APSCE

UPM
UNIVERSITI PUTRA MALAYSIA
BERILMU BERBAKTI

JUBLI EMAS
UPM 1971-2021

30TH INTERNATIONAL CONFERENCE ON COMPUTERS IN EDUCATION

ICCE 2022

KUALA LUMPUR, MALAYSIA
28 NOVEMBER - 2 DECEMBER 2022

CONFERENCE PROCEEDINGS

VOLUME I



HOSTED BY

UPM
UNIVERSITI PUTRA MALAYSIA
BERILMU BERBAKTI

JUBLI EMAS
UPM 1971-2021

FAKULTI
PENGAJIAN PENDIDIKAN
FACULTY OF EDUCATIONAL STUDIES
فاكولتي ڤڠاجين ڤندидikan

FACULTY OF EDUCATIONAL STUDIES,
UNIVERSITI PUTRA MALAYSIA, MALAYSIA

ORGANIZED BY

APSCE

THE ASIA-PACIFIC SOCIETY FOR
COMPUTERS IN EDUCATION (APSCE)

SUPPORTED BY

Malaysia Convention & Exhibition Bureau

Meet in Malaysia

Malaysia Truly Asia

SPONSORED BY

openlearning

TOURISM MALAYSIA®

30th International Conference on Computers in Education Conference Proceedings Volume I

Editors: Sridhar IYER, Ju-Ling SHIH, Weiqin CHEN, Mas Nida MD KHAMBARI

Publisher: Asia-Pacific Society for Computers in Education (APSCE)
Address: Center for Science and Technology for Learning,
National Central University, No. 300, Jhongda Road, Zongli District,
Taoyuan City 32001, Taiwan
Telephone: +886-3-4227151 ext. 35407
FAX: +886-3-4227151 ext. 35407
Email: service@apsce.net
Website: http://www.apsce.net

Date of publication: November 28, 2022

ISBN: 978-986-972-149-3

Copyright 2022 Asia-Pacific Society for Computers in Education.

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, transmitted, in any forms or any means, without the prior permission of the Asia-Pacific Society for Computers in Education. Individual papers may be uploaded on to institutional repositories or other academic sites for self-archival purposes.



EDITORS

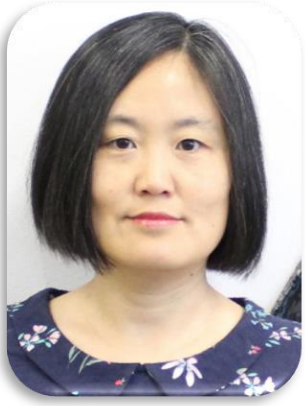
Sridhar IYER, Indian Institute of Technology, India
Ju-Ling SHIH, National Central University, Taiwan
Weiqin CHEN, Oslo Metropolitan University and University of Bergen, Norway
Mas Nida MD KHAMBARI, Universiti Putra Malaysia, Malaysia

ASSOCIATE EDITORS
(in alphabetical order)

Mouna DENDEN, University of Tunis, Tunisia
Rwitajit MAJUMDAR, Kyoto University, Japan
Liliana Cuesta MEDINA, Universidad de La Sabana, Colombia
Shitanshu MISHRA, MGIEP, UNESCO, India
Sahana MURTHY, Indian Institute of Technology Bombay, India
Patcharin PANJABUREE, Mahidol University, Thailand
Daner SUN, The Education University of Hong Kong, Hong Kong



MESSAGE FROM THE CONFERENCE CHAIR



Weiqin CHEN
Conference Chair
Oslo Metropolitan University and University of Bergen, Norway

On behalf of the organizing committee, I would like to welcome all participants of the 30th International Conference on Computers in Education (ICCE) 2022, the flagship conference series of the Asia-Pacific Society for Computers in Education (APSCE). After two years of virtual conferences in 2020 and 2021, this year we have our first hybrid conference with both in-person and online participants. ICCE is coming back to beautiful Malaysia that has successfully hosted ICCE 2010. Malaysia, with its pristine beaches, breath-taking islands and its diverse culture and rich history, will undoubtedly give participants a unique experience. The conference theme of ICCE 2022, "Optimising technology for sustainable quality education in the new norm" signifies the importance of sustainable learning augmented by technological innovations in the post-COVID era.

Four outstanding keynote speakers will share their insights across varying areas in the field of computers in education. Gwo-Dong Chen from the National Central University, Taiwan, will present how Digital Theater, an alternative approach to existing digital realities such as AR and VR, can support situational learning in the classroom. Rebecca Ferguson from the Open University, UK, will share with us innovative pedagogical models relevant for remote teaching and learning. Ryan Baker from the University of Pennsylvania, United States, will explicate machine-learned detectors in AIED systems and evidence and situations for successful model generalization. Su Luan Wong from Universiti Putra Malaysia (UPM) will talk about how students' interest in educational technology can be developed through learning activities designed based on Interest Driven Creator (IDC) theory. There will also be three equally inspiring theme-based invited speeches. Hyo-Jeong So from Ewha Womans University, Korea, will discuss possibilities and challenges for mobile learning to make use of learner data in learning design and analysis. Jan van Aalst from the University of Hong Kong will share with us his insights and reflections on knowledge building as an educational model in the post-truth era. Yu-Ju Lan from the National

Taiwan Normal University will talk about new paradigm of language learning supported by emerging digital technologies. These speeches connect with the essence of the conference theme in different ways and will stimulate reflections and inspire us to rethink sustainable quality education in the new normal post-COVID era.

Indeed, organizing such a large-scale conference requires the orchestrated efforts and unwavering support from the conference organizing committee members and conference paper reviewers. The hybrid mode of ICCE 2022 also poses new challenges that we have not experienced previously. I would like to express my sincere appreciations to all the individuals who have rendered their help in every possible way to make this conference a reality. The names of the hard-working Local Organizing Committees (LOC) chair and team members, International Program Coordination (IPC) chairs, Sub-conference chairs, Program Committee (PC) members and reviewers, organizers of Workshops, Tutorials, Work-In-Progress Posters (WIPP), Doctoral Student Consortium (DSC), Posters, Early Career Workshops (ECW), and Executive Summaries (ES) are enlisted in the proceedings. I am also grateful to all the paper authors and registered participants for their exciting academic contributions to the fruitful intellectual exchange in this conference.

Last but not least, I would like to express my heartfelt appreciation to the Managing Secretary of APSCE Pham-Duc Tho for all he has done for the conference, the standing committee for being flexible and proactive, and the consultants for sharing their experiences and wisdom and advising us along the way.

I hope all participants will have opportunities to renew friendships, forge new friendships and professional collaborations. I hope that you will have a productive and fun-filled time at this very special conference and leave Kuala Lumpur - a vibrant city with rich and amazing heritage —with beautiful, affectionate memories.

Thank you!

MESSAGE FROM THE INTERNATIONAL PROGRAM COORDINATION CHAIRS



Sridhar IYER
International Program Coordination Chair
Indian Institute of Technology, India



Ju-Ling SHIH
International Program Coordination Co-Chair
National Central University, Taiwan

Welcome to the 30th International Conference on Computers in Education (ICCE)! Organized by the Asia-Pacific Society for Computers in Education. ICCE 2022 is being held physically in Kuala Lumpur, Malaysia, from November 28 to December 2, 2022. The conference is in hybrid mode, so those who are unable to attend physically can do so virtually.

ICCE 2022 continues the meta-conference tradition of the previous ICCEs. As such, the conference is organized into seven sub-conference programs specializing specific themes:

- **C1 : ICCE Sub-Conference on Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS)**
- **C2 : ICCE Sub-Conference on Computer-supported Collaborative Learning (CSCL) and Learning Sciences (LS)**
- **C3 : ICCE Sub-Conference on Advanced Learning Technologies (ALT), Learning Analytics, Platforms and Infrastructure**
- **C4 : ICCE Sub-Conference on Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)**

C5 : ICCE Sub-Conference on Educational Gamification and Game-based Learning (EGG)
C6 : ICCE Sub-Conference on Technology Enhanced Language Learning (TELL)
C7 : ICCE Sub-Conference on Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

In addition to the main program with seven sub-conferences, ICCE 2022 includes various program components, such as Keynote Speeches, Theme-based Invited Speeches, Workshops, Tutorials, Work-in-Progress Posters (WIPP), Extended Summaries (ES), Doctoral Student Consortiums (DSC), and Early Career Workshops (ECW). All the papers in these program components are compiled and published in a separate volume with its own ISBN. Pre-conference events are held on the first two days of the conference, including workshops, tutorials, panels, DSC, ECW, APSCE Student Wing Workshops, and SIG community building sessions.

The International Program Committee is led by a strong and dedicated team, which includes the Conference Chair, the Program Coordination Chair and Co-Chair, Sub-Conference Chairs and Co-Chairs and experts in the field of Computers in Education from many different countries or economies. Former ICCE local organizing and program coordination chairs have played important roles as consultants in overseeing the organization process of this conference.

ICCE 2022 received a total of 153 submissions (120 full, 27 short, and 6 posters) from 36 different countries or economies. Top five countries with the highest number of submissions are Japan, Malaysia, Taiwan, India and China. Submissions were also received from the Middle East, Europe, America and Africa, which signals the international interest toward ICCE 2022.

All papers were subjected to a rigorous review process by at least three reviewers from the respective Sub-Conference program committees. After the reviews were completed, a meta-review was provided for each paper. In total, 544 reviews and 153 meta-reviews were received.

After the discussion period within the individual program committees led by the Sub-Conference Executive Chairs and Co-Chairs, recommendations were made to the Program Coordination Committee Chair and Co-Chair, who oversaw the review process and quality for all Sub-Conferences. This resulted in 32 full papers, 60 short papers, and 27 posters accepted across seven Sub-Conferences. The overall acceptance rate for full papers is 26.6%, which reflects our efforts to continue the maintenance of the quality of presentations at ICCE 2022.

The complete statistics of paper acceptance is shown in Table 1.

Table 1: Paper Acceptance Statistics

	Total Submissions	Submitted as Full Only	Accepted as Full	Full Only (%)	Accepted as Short	Accepted as Poster	Total Accepted	Overall Accepted (%)
C1 - AIED/ITS	31	26	7	26.92	15	1	23	74.19
C2 - CSCL/LS	22	13	4	30.77	7	7	18	81.82
C3 - ALT/LA	27	24	4	16.67	7	10	21	77.78
C4 - CUMTEL	13	9	3	33.33	6	1	10	76.92
C5 - EGG	18	15	4	26.67	7	2	13	72.22
C6 - TELL	11	8	3	37.50	4	3	10	90.91
C7 - PTP	31	25	7	28.00	14	3	24	77.42
Totals	153	120	32	26.67	60	27	119	77.78

The submission statistics by country, across all submissions (Main Conference, Workshops, WIPP, ES) is shown in Table 2.

Table 2: Submissions by country

Country	Submissions	Country	Submissions	Country	Submissions
Japan	62	Sweden	4	UAE	1
Malaysia	31	Spain	3	Italy	1
Taiwan	25	South Korea	3	Oman	1
India	23	Estonia	2	New Caledonia	1
China	22	Indonesia	2	UK	1
Philippines	21	Switzerland	2	Austria	1
United States	20	Croatia	2	Brunei Darussalam	1
Hong Kong	16	Germany	2	Nepal	1
Thailand	14	Poland	2	New Zealand	1
Singapore	12	France	2	Norway	1
Canada	6	Tunisia	2	Peru	1
Australia	5	Vietnam	2	South Africa	1

We are grateful to all who contributed to ICCE 2022's success. We thank all the paper authors for choosing ICCE 2022 as the venue to present their research. We would also like to thank the IPC Executive Chairs/Co-Chairs and members, who undertook the responsibility of reviewing and selecting papers that represent research of high quality. Specially thanks to our Keynote and Invited Speakers for accepting our invitations and sharing inspiring research with the ICCE 2022 participants. We are grateful to the APSCE community for continuing to make the valuable contributions to education that will help shape the minds, hearts, and spirits of future generations.



MESSAGE FROM THE LOCAL ORGANIZING COMMITTEE CHAIR



Mas Nida MD KHAMBARI
Local Organizing Committee Chair
Universiti Putra Malaysia, Malaysia

Selamat Datang, Welcome to Malaysia!

On behalf of the organising committee, I would like to extend my warm welcome to all delegates of the 30th International Conference on Computers in Education (ICCE 2022), held for the third time in Malaysia.

It is a great pleasure and honour to host ICCE 2022 as it recommences physically this year — adopting the hybrid mode. I am sure many of us are excited to meet each other again. ICCE 2022 is an important and timely event for researchers, educators, and practitioners to reconnect after having to meet virtually for the past two years (ICCE 2020 & ICCE 2021). The theme of the conference, "Optimising technology for sustainable quality education in the new norm," aptly reflects what we had recently experienced and looking forward to, in the future. Digital technologies adoption has taken an exponential leap during the global COVID-19 pandemic – transforming education in the new norm. When digital education ensues, technologies are widely optimised for learning. Universiti Putra Malaysia, along with other Research Universities in Malaysia, has taken an ardent role in ensuring the sustainability of education during the pandemic. Extensive efforts were taken to equip academics and students with digital literacy and competency, and digital devices. This process was progressive as we extended the efforts to equip classrooms with hyflex facilities to support the transactional distance caused by COVID-19. Universiti Putra Malaysia especially, has built Putra Future Classroom, Putra Hybrid Classrooms, Putra®Smart Classroom, to name a few.

It is a great privilege to share our beautiful country with you. I hope you will be able to enjoy Kuala Lumpur, one of the most vibrant cities in Asia that possesses a distinct and charming character. Situated in the Golden Triangle, Impiana KLCC, our conference venue, is adjacent to the Petronas Twin Towers, the world-renowned icon of the country. Fondly known for its authentic Malaysian hospitality, this hotel is an oasis for business and leisure travellers. There are many possibilities for sightseeing and experiencing Kuala Lumpur and Malaysian culture before and after the conference; all of which are within a 10-minute walking distance via the sky-bridge. I encourage you to explore our website page for tourist information where you will find useful information on places to go.

I would like to thank the APSCE Executive Committee for giving us this wonderful opportunity. Our sincere thanks to all the sponsors, the standing committee, the International Program Committee, reviewers, authors, participants and student volunteers. We trust all of you will enjoy the conference, and take home great memories from Kuala Lumpur, Malaysia.

Sampai bertemu! Till later!

ORGANIZATION

Organized by: Asia-Pacific Society for Computers in Education

Standing Committee

Conference Chair

Weiqin CHEN, Oslo Metropolitan University and University of Bergen, Norway

International Program Coordination Chair

Sridhar IYER, Indian Institute of Technology, India

International Program Coordination Co-Chair

Ju-Ling SHIH, National Central University, Taiwan

Local Organizing Committee Chair

Mas Nida MD KHAMBARI, Universiti Putra Malaysia, Malaysia

Consultants

Maria Mercedes T. RODRIGO, Ateneo de Manila University, Philippines
Lung-Hsiang WONG, National Technology University, Singapore
Thepchai SUPNITHI, National Electronics and Computer Technology Center, Thailand

Local Host Committee

Patron

Mohd Roslan SULAIMAN, Universiti Putra Malaysia, Malaysia

Main Advisor

Soaib ASIMIRAN, Universiti Putra Malaysia, Malaysia

Advisory Board

Siti Raba'ah HAMZAH, Universiti Putra Malaysia, Malaysia
Habibah AB JALIL, Universiti Putra Malaysia, Malaysia
Abu Bakar MOHAMED RAZALI, Universiti Putra Malaysia, Malaysia

Consultants

Su Luan WONG, Universiti Putra Malaysia, Malaysia
Nor Aniza AHMAD, Universiti Putra Malaysia, Malaysia

Local Host Chair

Mas Nida MD. KHAMBARI, Universiti Putra Malaysia, Malaysia



Local Host Co-chair
Ahmad Fauzi MOHD AYUB, Universiti Putra Malaysia, Malaysia

Secretary
Nadiah KAMARUDDIN, Universiti Putra Malaysia, Malaysia

Finance Coordinators
Arnida ABDULLAH, Universiti Putra Malaysia, Malaysia
Husaini HUSSAIN, Universiti Putra Malaysia, Malaysia
Siti Nor Ain IBRAHIM, Universiti Putra Malaysia, Malaysia

Social Events and Cultural Experience Coordinators
Sharifah Intan Sharina SYED ABDULLAH, Universiti Putra Malaysia, Malaysia
Nadiah KAMARUDDIN, Universiti Putra Malaysia, Malaysia
Nor Azni ABDUL AZIZ, Universiti Putra Malaysia, Malaysia

Proceedings Coordinators
Marzni MOHAMED MOKHTAR, Universiti Putra Malaysia, Malaysia
Norzihani SAHARUDDIN, Universiti Putra Malaysia, Malaysia
Zeiti Zulhani ZAKARIA, Universiti Putra Malaysia, Malaysia
Nor Syazila ABDUL RAHIM, Universiti Putra Malaysia, Malaysia
Nurul Nadwa ZULKIFLI, Universiti Putra Malaysia, Malaysia

Protocol and Registration Coordinators
Norliza GHAZALI, Universiti Putra Malaysia, Malaysia
Natassah OTHMAN, Universiti Putra Malaysia, Malaysia
Najihah ZAHARI, Universiti Putra Malaysia, Malaysia

Publicity & Social Media Coordinators
Shwu Pyng HOW, Universiti Putra Malaysia, Malaysia
Mohd. Syafiq Farhan ROSHIDI, Universiti Putra Malaysia, Malaysia
Muhammad Hirzi MOHAMED HALMI, Universiti Putra Malaysia, Malaysia
Intan Zuliana ZULKIPLI, Universiti Putra Malaysia, Malaysia

Logistics & Volunteers Coordinators
Nur Aira ABD RAHIM, Universiti Putra Malaysia, Malaysia
Nor Ain Syaheera NORDIN, Universiti Putra Malaysia, Malaysia
Fuad MOHAMMAD, Universiti Putra Malaysia, Malaysia
Suhairi ISHAK, Universiti Putra Malaysia, Malaysia
Othman ISHAK, Universiti Putra Malaysia, Malaysia

Website, Technical, Virtual Platform and Certificates Coordinators

Mohd Hazwan MOHD PUAD, Universiti Putra Malaysia, Malaysia
Muhammad Faisfadly SABARUDDIN, Universiti Putra Malaysia, Malaysia
Muhamad Amirul Azreen MOHD ZULKEFLI, Universiti Putra Malaysia, Malaysia
Pham Duc THO, Asia-Pacific Society for Computers in Education HQ

Sub-Conference

C1: Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS) and Adaptive Learning

PC Executive Chair

Patcharin PANJABUREE, Mahidol University, Thailand

PC Co-Chairs

Ryan BAKER, University of Pennsylvania, United States
Sabine SEUFERT, University of St. Gallen, Switzerland

C2: Computer-supported Collaborative Learning (CSCL) and Learning Science

PC Executive Chair

Sahana MURTHY, Indian Institute of Technology Bombay, India

PC Co-Chairs

Daniel BODEMER, University of Duisburg-Essen, Germany
Camillia MATUK, New York University, United States
Chew Lee TEO, Nanyang Technological University, Singapore
Kate THOMPSON, Queensland University of Technology, Australia

C3: Advanced Learning Technologies (ALT), Learning Analytics and Digital Infrastructure

PC Executive Chair

Rwitajit MAJUMDAR, Kyoto University, Japan

PC Co-Chairs

Mohammed SAQR, University of Eastern Finland, Finland
Gökhan AKCAPINAR, Hacettepe University, Turkey
Khalid KHAN, Charles Darwin University, Australia
Lakshmi GANESH, Kotak Education Foundation, India



C4: Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)

PC Executive Chair

Daner SUN, The Education University of Hong Kong, Hong Kong

PC Co-Chairs

Martina HOLENKO DLAB, University of Rijeka, Croatia
Grace QI, Massey University, New Zealand
Yuqin YANG, Central China Normal University, China

C5: Educational Gamification and Game-based Learning (EGG)

PC Executive Chair

Mouna DENDEN, University of Tunis, Tunisia

PC Co-Chairs

Jewoong MOON, University of Alabama, United States
Hafed ZARZOUR, Souk Ahras University, Algeria
Junfeng YANG, Hangzhou Normal University, China

C6: Technology Enhanced Language Learning (TELL)

PC Executive Chair

Liliana CUESTA, Universidad de La Sabana , Colombia

PC Co-Chairs

Carl ANDERSON, Universidad de La Sabana, Colombia
Michelle MARSEE, Chandler-Gilbert Community College, United States
Rosa Dene DAVID, Universidad de La Sabana, Colombia

C7: Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)

PC Executive Chair

Shitanshu MISHRA, MGIEP, UNESCO, India

PC Co-Chairs

Ivica BOTICKI, University of Zagreb, Croatia
Dan KOHEN-VACS, Holon Institute of Technology, Israel
Jayakrishnan MADATHIL, Indian Institute of Technology Madras, India

Workshops/ Interactive Events

Chair

Chiu-Lin LAI, National Taipei University of Education, Taiwan

Co-Chairs

Atima KAEWSA-ARD, Mae Fah Luang University, Thailand
Shwu Pyng HOW, Universiti Putra Malaysia, Malaysia

Tutorials

Chair

Kaushal Kumar BHAGAT, Indian Institute of Technology Kharagpur, India

Co-Chair

Sagaya AMALATHAS, University of Southampton, Malaysia campus, Malaysia

Work-in-Progress Posters (WIPP)

Chair

Mi Song KIM, University of Western Ontario, Canada

Co-Chairs

Atima KAEWSA-ARD, Mae Fah Luang University, Thailand
Wen YUN, Nanyang Technological University, Singapore
Siti Khadijah ALI, Universiti Putra Malaysia, Malaysia
Mohd Hazwan MOHD PUAD, Universiti Putra Malaysia, Malaysia

Doctoral Student Consortium (DSC)

Chair

Bo JIANG, Zhejiang University of Technology, China

Co-Chairs

Hiroaki OGATA, Kyoto University, Japan
Yanjie SONG, The Education University of Hongkong, Hongkong
Vwen Yen Awyln LEE, Nanyang Technological University, Singapore
Nor Azni ABDUL AZIZ, Universiti Putra Malaysia, Malaysia

Early Career Workshops (ECW)

Chair

Maiga CHANG, Athabasca University, Canada

Co-Chairs

Anita DIWAKAR, VJTI, India
Muhd Khaizer OMAR, Universiti Putra Malaysia, Malaysia

Panels

Chair

Jayakrishnan MADATHIL, Indian Institute of Technology Madras, India

Co-Chair

Sharifah Intan Sharina SYED ABDULLAH, Universiti Putra Malaysia, Malaysia

Extended Summaries (ES)

Chair

Lydia Yan LIU, East China Normal University, China

Co-Chairs

Hang SHU, Jiangnan University, China
Arnida ABDULLAH, Universiti Putra Malaysia, Malaysia

Merit Scholarships

Chair

Gökhan AKCAPINAR, Hacettepe University, Turkey

Co-Chairs

Erkan ER, Middle East Technical University, Turkey
Nur Aira ABD RAHIM, Universiti Putra Malaysia, Malaysia



Special Interest Groups (SIG) 2022-2023

S1: Artificial Intelligence in Education/Intelligent Tutoring Systems/Adaptive Learning (AIED/ITS/AL)
May Marie TALANDRON-FELIPE, University of Science and Technology, Phillipines

S2: Computer-supported Collaborative Learning and Learning Sciences (CSCL)
Elizabeth KOH, Nanyang Technological University, Singapore

S3: Advanced Learning Technologies, Learning Analytics, Platforms and Infrastructure (ALT)
Eunice SARI, UX, Indonesia

S4: Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)
Daner SUN, The Education University of Hong Kong, Hong Kong

S5: Educational Gamification and Game-based Learning (EGG)
Ahmed TLILI, Beijing Normal University, China

S6: Technology Enhanced Language Learning (TELL)
Vivian WU, Asia University, Taiwan

S7: Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)
Mas Nida MD KHAMBARI, Universiti Putra Malaysia, Malaysia

S8: Development of Information and Communication Technology in the Asia-Pacific Neighborhood (DICTAP)
Patcharin PANJABUREE, Mahidol University, Thailand

S9: Educational Use of Problems/Questions in Technology-Enhanced Learning (EUPQ)
Takahito TOMOTO, Tokyo Polytechnic University, Japan

S10: Learning Analytics and Educational Data Mining (LAEDM)
Ramkumar RAJENDRAN, Indian Institute of Technology Bombay, India

S11: Computational Thinking Education & STEM Education (CTE&STEM)
Chee Kit LOOI, Nanyang Technological University, Singapore



C1 PC Members

Sagaya Amalathas, Taylors University, Malaysia
Ryan Baker, University of Pennsylvania, United States
Kritya Bunchongchit, Mahidol University, Thailand
Chih-Yueh Chou, Yuan Ze University, Taiwan
Philippe Fournier-Viger, Harbin Institute of Technology, China
Claude Frasson, University of Montreal, Canada
Yuki Hayashi, Osaka Prefecture University, Japan
Bastiaan Heeren, Open University, The Netherlands
Tomoya Horiuchi, Kobe University, Japan
Sharon Hsiao, Arizona State University, United States
Akihiro Kashihara, The University of Electro-Communications, Japan
Tomoko Kojiri, Kansai University, Japan
Tatsuhiro Konishi, Shizuoka University, Japan
Noboru Matsuda, North Carolina State University, United States
Tatsunori Matsui, Waseda University, Japan
Tanja Mitrovic, University of Canterbury, New Zealand
Riichiro Mizoguchi, Japan Advanced Institute of Science and Technology, Japan
Roger Nkambou, Université du Québec à Montréal, Canada
Ange Tato, Université du Québec à Montréal, Canada
Jose Luis Perez De La Cruz, Universidad de Malaga, Spain
Elvira Popescu, University of Craiova
Ma. Mercedes T. Rodrigo, Ateneo de Manila University, Philippines
Olga C. Santos, aDeNu Research Group, Spain
John Stamper, Carnegie Mellon University, United States
Thepchai Supnithi, NECTEC, Thailand
Benedict Du Boulay, University of Sussex, United Kingdom



C2 PC Members

Sahana Murthy, Indian Institute of Technology Bombay, India
Daniel Bodermer, University of Duisburg-Essen, Germany
Camillia Matuk, New York University, United States
Kate Thompson, Queensland University of Technology, Australia
Sakinah Alhadad, Griffith University, Australia
Natasha Arthars, The University of Sydney, Australia
Chris Blundell, Queensland University of Technology, Australia
Jürgen Buder, Leibniz-Institut für Wissensmedien, Germany
Sanjay Chandrasekharan, Tata Institute of Fundamental Research, India
Jennifer Clifton, Queensland University of Technology, Australia
Julia Eberle, Ruhr-University Bochum, Germany
Lakshmi T. G., Kotak Education Foundation, India
Steven Kickbusch, Queensland University of Technology, Australia
Simon Knight, University of Technology, Australia
Elizabeth Ruilin Koh, Nanyang Technological University, Singapore
Aditi Kothiyal, EPFL, Switzerland
Simon Leonard, University of South Australia, Australia
Rose Liang, National University of Singapore, Singapore
Pei-Yi Lin, National Tsing-Hua University, Taiwan
Lori Lockyer, Queensland University of Technology, Switzerland
Michelle Mukherjee, Queensland University of Technology, Australia
Johanna Pöysä-Tarhonen, University of Jyvaskyla, Finland
Shilpa Sahay, New York University, United States
Lenka Schnaubert, University of Duisburg-Essen, Germany
Antonette Shibani, University of Technology, Australia
Seng Chee Tan, Nanyang Technological University, Singapore
Hongzhi Yang, The University of Sydney, Australia

C3 PC Members

Aditi Kothiyal, IIT Gandhinagar, India
Atsushi Shimada, Kyushu University, Japan
Chewlee Teo, Nanyang Technological University, Singapore
Erkan Er, Middle East Technical University, Turkey
Gökhan Akcapinar, Hacettepe University, Turkey
Huiyong Li, Kyoto University, Japan
Jerry Chih-Yuan Sun, National Yang Ming Chiao Tung University, Taiwan
Jon Mason, Charles Darwin University, Australia
Judith Azcarraga, De La Salle University, Philippines
Khalid Khan, Charles Darwin University, Australia
Kyosuke Takami, Kyoto University, Japan
Lakshmi Ganesh, Kotak Education Foundation, India
Luis Anido Rifon, Universidade de Vigo, Spain
Manuel Caeiro Rodríguez, University of Vigo, Spain
Marc Jansen, University of Applied Sciences Ruhr West, Germany
Minhong Wang, The University of Hong Kong, Hong Kong
Mohammed Saqr, University of Eastern Finland, Finland
Oluwafemi Samson Balogun, University of Eastern Finland, Finland
Prajish Prasad, IIT Bombay, India
Ramkumar Rajendran, IIT Bombay, India
Regina Motz, Universidad de la República, Uruguay
Rekha Ramesh, Mumbai University, India
Riina Vuorikari, Institute for Prospective Technological Studies (IPTS), Spain
Rwitajit Majumdar, Kyoto University, Japan
Sonsoles López-Pernas, Universidad Politécnica de Madrid, Spain
Tore Hoel, Oslo Metropolitan University, Norway
Victoria Abou Khalil, ETH Zurich, Switzerland
Vladimir Costas, Universidad Mayor de San Simón, Bolivia
Yang-Hsueh Chen, National Chengchi University, Taiwan
Yiling Dai, Kyoto University, Japan



C4 PC Members

Daner Sun, The Education University of Hong Kong, Hong Kong
Martina Holenko Dlab, University of Rijeka, Croatia
Grace Qi, Massey University, New Zealand
Yuqin Yang, Central China Normal University, China
Chiu-Pin Lin, National Tsing Hua University, Taiwan
Longkai Wu, Central China Normal University, China
Kaushal Kumar Bhagat, Indian Institute of Technology, Kharagpur, India
Fuhua (Oscar) Lin, Athabasca University, Canada
Ivica Boticki, University of Zagreb, Croatia
Su Cai, Beijing Normal University, China
Yanjie Song, The Education University of Hong Kong, Hong Kong
Chih-Ming Chu, National Ilan University, Taiwan
Ting-Ting Wu, National Yunlin University of Science and Technology, Taiwan
Noriko Uosaki, Osaka University, Japan
Yih-Ruey Juang, Jinwen University of Science and Technology, Taiwan
Jun-Ming Su, National University of Tainan, Taiwan
Ivana Bosnić, University of Zagreb, Croatia
M. Carmen Juan Lizandra, Universitat Politècnica De València, Spain
Igor Mekterović, University of Zagreb, Croatia
Tai-Chien Kao, National Dong Hwa University, Taiwan
Ju-Ling Shih, National University of Tainan, Taiwan
Hsu-Cheng Chiang, National Taiwan Normal University, Taiwan
Gwo-Haur Hwang, National Yulin Technology University, Taiwan
Chen-Yu Lee, Ling Tung University, Taiwan
Kuo-Liang Ou, National Tsing Hua University, Taiwan
Tzu-Chi Yang, National Chiao Tung University, Taiwan
Ben Chang, National Central University, Taiwan
Chiu-Lin Lai, National Taipei University of Education
Haiguang Fang, Capital Normal University, China
Feng-Kuang Chiang, Shanghai Normal University, China
Jing Leng, East China Normal University, China
Guang Chen, Beijing Normal University, China
Ping He, Tianjin University, China
Huiying Cai, Jiangnan University, China
Bian Wu, East China Normal University, China
Zhihong Wan, The Education University of Hong Kong, Hong Kong
Luo Ma, East China Normal University, China
Ying Zhan, The Education University of Hong Kong, Hong Kong
Maja Gligora Markovic, University of Rijeka, Croatia

C5 PC Members

Zhi-Hong Chen, National Taiwan Normal University, Taiwan
Boussaha Karima, Universite Badji Mokhtar Annaba, Algeria
Hiroyuki Mitsuhara, Tokushima University, Japan
Ju-Ling Shih, National Central University, Taiwan
Amina Zedadra, University of Guelma, Algeria
Chih-Pu Dai, Florida State University, United States
Jorge Simões, Institute Superior Politécnico Gaya, Portugal
Susan Gwee, English Language Institute of Singapore, Singapore
Abdelmalek Bouguettaya, Research Centre in Industrial Technologies (CRTI), Algeria
JaeHwan Byun, Wichita State University, United States
Luke West, Florida State University, United States
Khaled Halimi, Universite 8 Mai 1945 Guelma, Algeria
Kaoru Sumi, Future University Hakodate, Japan
Liz Bacon, Abertay University, Scotland
Jina Kang, University of Illinois Urbana-Champaign, United States
Gi Woong Choi, University of Cincinnati, United States
Mohamed Koutheaïr Khribi, Mada-Qatar Assistive Technology Center, Qatar
Mahnane Lamia, University of Badji Mokhtar, Annaba, Algeria
Ting-Wen Chang, Beijing Normal University, China
Jakub Swacha, University of Szczecin, Poland
Yassine Safsouf, LIMIE Laboratory, ISGA Group, Morocco
Sungwoong Lee, University of West Georgia, United States
Hafidi Mohamed, University of Badji Mokhtar, Algeria
Yanjun Pan, Florida State University, United States
Samia Drissi, Univeristé de Souk Ahras, Algeria
Fezile Ozdamli, Near East University,Turkey
Gheorghita Ghinea, Brunel University, United Kingdom
Ana Manzano-León, University of Almería, Spain
Sabine Graf, Athabasca University, Canada

C6 PC Members

Oi Misato, Kyushu University, Japan
Brendan Flangan, Kyoto University, Japan
Ju-Ling Shih, National Central University, Taiwan
Sahana Murthy, Indian Institute of Technology Bombay, India
Michelle Marsee, Chandler-Gilbert Community College, United States
Yanjie Song, The Education University of Hong Kong, Hong Kong
Agnieszka Palalas, Athabasca University, Canada
Apostolos Koutropoulos, University of Massachusetts Boston, United States
Alex Boulton, University of Lorraine, France
Xin Chen, Indiana University, United States
Yasushige Ishikawa, Kyoto University of Foreign Studies, Japan
Jiahang Li, Michigan State University, United States
Yanhui Han, The Open University of China, China
Jie-Chi Yang, National Central University, Taiwan

C7 PC Members

Ahmed Mohammed, Linnaeus University, Sweden
Ajita Deshmukh, MIT-ADT University, Pune
Anabil Munshi, Vanderbilt University, United States
Arriel Benis, Holon Institute of Technology, Israel
Ashutosh Raina, Indian Institute of Technology Bombay, India
Bernard Yett, Vanderbilt University, United States
Brendan Flanagan, Kyoto University, Japan
Caitlin Snyder, Vanderbilt University, United States
Eran Gal, Holon Institute of Technology, Israel
Gayithri Jayathirtha, University of Pennsylvania, Philadelphia
H. Ulrich Hoppe, University Duisburg-Essen / RIAS Institute Duisburg, Germany
Hayley Weigelt-Marom, Holon Institute of Technology, Israel
Joke Voogt, University of Amsterdam, Netherlands
Kapil Kadam, Indian Institute of Technology Bombay, India
Lucian Vumilia Ngeze, Indian Institute of Technology Bombay, India
Marc Jansen, University of Applied Sciences Ruhr West, Germany
Marcelo Milrad, Linnaeus University, Sweden
Martina Holenko Dlab, University of Rijeka, Croatia
Narasimha Swamy, Indian Institute of Technology Bombay, India
Navneet Kaur, Indian Institute of Technology Bombay, India
Pankaj Chavan, Deogiri Institute of Engineering and Management Studies, India
Rotem Israel-Fishelson, Tel Aviv University, Israel
Shu-Shing Lee, Nanyang Technological University, Singapore
Veenita Shah, Indian Institute of Technology Bombay, India
Winnie Wai Man Lam, The Education University of Hong Kong, Hong Kong
Yogendra Pal, NIIT University, India

APSCE FELLOW PROGRAM

Inaugurated in 2019, the APSCE Fellowship recognizes outstanding members of the Asia-Pacific Society for Computers in Education (APSCE) in the field of computers in education. The title of APSCE fellow indicates, (1) Sustained and distinguished academic contributions to the advancement of research in the field of computers in education at the international level; (2) A strong track record in academic networking and services within the Asia-Pacific region.

The full guidelines for APSCE Fellows are available here:

APSCE-Fellows-Program-Guidelines
https://new.apsce.net/wp-content/uploads/2022/10/APSCE-Fellows-Program-Guidelines.pdf

APSCE-Fellow-Nomination-Form
https://new.apsce.net/apsce-fellows/

Following a decision made by the APSCE Fellows Committee, Prof. Antonija Mitrovic (New Zealand) is inducted as the new APSCE Fellow for 2022.



Antonija MITROVIC
APSCE Fellow 2022
University of Canterbury, New Zealand

Biography: Dr Antonija (Tanja) Mitrovic is a full professor at the Department of Computer Science and Software Engineering at the University of Canterbury, Christchurch, New Zealand. She is the leader of ICTG (Intelligent Computer Tutoring Group). Dr Mitrovic received her PhD in Computer Science from the University of Nis, Yugoslavia, in 1994. She is an associate editor of the following journals: Practice in Technology Enhanced Learning (RPTEL), International Journal on Artificial Intelligence in Education (IJAIED) and Journal of Universal Computer Science (JUCS). She is a Distinguished member of ACM, and senior member of AAAI and IEEE. She was awarded the Distinguished Researcher Award in 2011 by APSCE.

Dr Mitrovic's primary research interests are in student modeling. ICTG has developed a number of constraint-based intelligent tutoring systems in a variety of domains, which have been thoroughly evaluated in real classrooms, and proven to be highly effective. These systems provide adaptive support for acquiring both problem-solving skills and meta-cognitive skills. Although most of the ITSs developed by ICTG support students learning individually in areas such as database querying (SQL-Tutor), database design (EER-Tutor and ERM-Tutor), data normalization (NORMIT), there are also constraint-based tutors for object-oriented software design and collaborative skills, various engineering topics (thermodynamics, mechanics), training to interpret medical images and language-learning. ICTG has also developed ASPIRE, a full authoring and deployment environment for constraint-based tutors. Her recent research focuses on AI-based support for active learning from videos.

DISTINGUISHED RESEARCHER AWARD WINNER



Maiga Chang

Professor, School of Computing and Information Systems, Athabasca University, Canada

Dr. Maiga Chang is a Full Professor in the School of Computing and Information Systems at Athabasca University, Canada. Dr. Chang has given more than 135 talks and lectures in different events. He also has (co-)authored more than 240 book chapters, journal and international conference papers. He is an IEEE member since 1996.

Dr. Chang is currently the chair of the IEEE (Institute of Electrical and Electronics Engineers) Technical Community of Learning Technology (TCLT) as well as the editor-in-chief of Educational Technology and Society, which is indexed by Web of Science's Social Science Citation Index (SSCI). In addition, He is also the editor-in-chief of the International Journal of Distance Education Technologies (IJDET) and the Bulletin of Technical Committee on Learning Technology – both are open access publication indexed by Web of Science Emerging Sources Citation Index (ESCI).

Dr. Chang is now Vice President (2022~) of International Association of Smart Learning Environments (IASLE); Executive Committee member of Asia-Pacific Society for Computers in Education (APSCE, 2017~2024), IEEE Computer Society Technical & Conference Activities Board (2022), Global Chinese Society for Computing in Education (GCSCE, 2016~2025), IEEE Computer Society Special Technical Communities (2021~); and Chair (2021~) of Educational Activities Committee, IEEE Northern Canada Section. Dr. Chang is also a Steering Committee member (2020~) for International Conference on Intelligent Tutoring Systems (ITS).

In the last two decades, Dr. Chang's research directions are Data Analytics Service and Game-based Learning. His Data Analytics research focus on the design of algorithms and methodologies for item generation (for quizzes and exams), learning activity generation, behaviour pattern extraction from sequential data and preference prediction and recommendation service. The Personalized Study Guide is research aiming to enable an open-source learning management system (i.e., Moodle) to provide students with a personalized study guide for their online learning by via the use of graph structures to analyze the learning objects in an online course and compare the learning behaviours, strategies, and preferences of individual students studying online. Three open-source Moodle plugins have been developed, reviewed, approved, and included in Moodle plugins directory.

On the Game-based Learning direction, his research continuously designs games for teaching, learning, rewards, and assessment. For instances, different games (including 3D, mobile and role-playing games) have been designed and developed to help students learning various topics, including botany, culture and history, finance, programming, management information systems, and meta-cognitive skills. His Trading Card Game (TCG) and the use of In-game Card as Educational Reward (ICER) research also inspired Trinity Primary School in the United Kingdom to develop "Character Clash!" to encourage children to read more. The Multiplayer Educational Game for All (MEGA World) is a web-based massively multiplayer educational game platform which supports any languages and is capable of access any existing external resources (e.g., multimedia, materials, online meetings, etc.). Teachers can create their virtual worlds as well as create learning and assessment activities (i.e., quests in the game) for students. Students can learn specific knowledge and reach the learning goal by taking and solving those quests while playing.

EARLY CAREER RESEARCHER AWARD WINNER (2022)



Daner Sun

Assistant Professor, Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong

Dr Sun Daner is an assistant professor at the Department of Mathematics and Information Technology, the Education University of Hong Kong (EdUHK), Hong Kong SAR, China. Prior to joining EdUHK in 2015, she had been working as a postdoc fellow at the Learning Sciences Lab of the National Institute of Education, Nanyang Technological University, Singapore. Dr Sun follows a long–term research agenda spanning ICT-supported science education, AI in education, mobile learning, technology-oriented STEM education, and higher-order thinking in interdisciplinary education. In her research, she addresses the prevailing hot topics and challenges in these fields, thereby developing a solid understanding of the developments in these fields as well as making her own contributions. In brief, with great enthusiasm for research, Dr Sun has published 25 journal papers with 8 (top1%-5%) papers and 13 (top 5% -15%) papers. Besides, Dr Sun published eight book chapters and served as the editor/co-editor for four conference proceedings, and one special issue in RPTEL journal. Meanwhile, she actively leads or contributes to research projects including Co-I for five external projects and PI for 1 General Research Fund (GRF) project.

Dr Sun spares no efforts in taking various leading roles in local and international academic events and activities for contributing to the professional research community. She has taken leading roles as an organizer of ICCE and GCCCE international conferences. Three of her conference papers have been awarded nominations for the best student paper, the best overall paper, the best design paper, and in referred international conferences (i.e. GCCCE2018, ICCE2013, CSCL2013). She is the chair of the Special Interest Group (SIG) of CUMTEL of APSCE (2022-23); She is the Associate Editor of APSCE official journal: RPTEL. She is a reviewer of SSCI journals and serves as a member of relevant societies and associations. She is the awardee of the Outstanding Reviewer Award of Computers & Education 2018. Besides, APSCE Early Career Researcher Award (ECRA) 2022, Dr Sun is also the awardee of the President's Awards for Outstanding Performance in Knowledge Transfer (Team) 2019/2020 and Dean's Research Output Prize 21/22 in EdUHK.

LAST TEN YEARS' DISTINGUISHED RESEARCHER AWARD WINNERS

2021 - APSCE Distinguished Researcher Award
Maria Mercedes T. Rodrigo, Ateneo de Manila University, Philippines

2020 - APSCE Distinguished Researcher Award
Wenli CHEN, Nanyang Technological University, Singapore

2015 - APSCE Distinguished Researcher Award
Lung-Hsiang WONG, Nanyang Technological University, Singapore

2014 - APSCE Distinguished Researcher Award
Hiroaki OGATA, Kyushu University, Japan

2011 - APSCE Distinguished Researcher Award
Antonija MITROVIC, University of Canterbury, New Zealand
Chen-Chung LIU, National Central University, Taiwan

LAST TEN YEARS' EARLY CAREER RESEARCHER AWARD WINNERS

2021 - APSCE Early Career Researcher Award
Bo Jiang, East China Normal University, China

2020 - APSCE Early Career Researcher Award
Kaushal Kumar BHAGAT, Indian Institute of Technology, Kharagpur, India

2019 - APSCE Early Career Researcher Award
Cheng-Jiu YIN, Kobe University, Japan

2018 - APSCE Early Career Researcher Award
Ting-Chia HSU, National Taiwan Normal University, Taiwan

2017 - APSCE Early Career Researcher Award
Jon MASON, Charles Darwin University, Australia

2015 - APSCE Early Career Researcher Award
Morris Siu-Yung JONG, The Chinese University of Hong Kong, Hong Kong

SPEAKERS OF APSCE WEBINAR SERIES (December 2021 – November 2022)

APSCE Webinar #20: Sustainable Learning Analytics in the Digital Age
Date: 04 April 2022 (Monday)
Speakers: Prof. Stephen Yang (National Central University, Taiwan)
Moderator: Prof. Ramkumar Rajendran (Indian Institute of Technology Bombay, India)
Curated by: APSCE Learning Analytics & Educational Data Mining SIG

APSCE Webinar #21: Intelligent Educational Gamification System Based on the Learner's Personality and Needs
Date: 10 May 2022 (Tuesday)
Speaker: Dr. Mouna Denden (Polytechnic University of Hauts-de-France (UPHF), France)
Moderator: Prof. Ahmed Tlili (Beijing Normal University, China)
Curated by: APSCE Educational Gamification and Game-based Learning (EGG) SIG

APSCE Webinar #22: Open-Ended Learning Environments Supporting STEM+C Learning
Date: 02 June 2022 (Thursday)
Speaker: Prof. Gautam BISWAS (Vanderbilt University, USA)
Moderator: Prof. Chee Kit LOOI (Nanyang Technological University, Singapore)
Curated by: APSCE Computational Thinking in Education & STEM (CTE-STEM) SIG

APSCE Webinar #23: Learning by Re-composition for Shared Understanding Between Peers and Between Peers and Teacher
Date: 23 June 2021 (Thursday)
Speaker: Prof. Tsukasa Hirashima (Hiroshima University, Japan)
Moderator: Prof. Takahito TOMOTO (Tokyo Polytechnic University, Japan)
Curated by: APSCE Educational Use of Problems/Questions in Technology (EUPQ) SIG

APSCE Webinar #24: Shaping Mobile Learning Futures for The Common Good
Date: 04 July 2022 (Monday)
Speaker: Prof. Agnes KUKULSKA-HULME (The Open University, UK)
Moderator: Dr. Daner SUN (The Education University of Hong Kong, Hong Kong)
Curated by: APSCE Classroom, Ubiquitous and Mobile Technology Enhanced Learning (CUMTEL) SIG

APSCE Webinar #25: Breaking the Barriers in the Learning Aids Development for the Visually Impaired and the Blind

Date: 30 August 2022 (Tuesday)
Speaker: Assoc. Prof. Aaron Raymond SEE (Southern Taiwan University of Science and Technology, Taiwan)
Moderator: Dr. May Marie P. TALADRON-FELIPE (University of Science and Technology of Southern Philippines, Philippines)
Curated by: APSCE Artificial Intelligence in Education, Intelligent Tutoring Systems & Adaptive Systems (AI-ED) SIG

APSCE Webinar #26: Multimodal Data in CSCL

Date: 08 September 2022 (Thursday)
Panelist:
Mutlu CUKUROVA (University College London, UK)
Muhterem DINDAR (Tampere University, Finland)
Roberto Martinez MALDONADO (Monash University, Australia)
Chair: Elizabeth KOH (Nanyang Technological University, Singapore)
Curated by: APSCE Computer-Support Collaborative Learning / Learning Sciences (CSCL/LS) SIG

APSCE Webinar #27: Design Thinking for Educators

Date: 06 October 2022 (Thursday)
Speaker: Dr. Eunice SARI (UX Indonesia and Customer Experience Insight Australia)
Moderator: Mr. Zachary Roland Anthony (Universiti Tunku Abdul Rahman, Malaysia)
Curated by: APSCE Practice-Driven Research, Teacher Professional Development and Policy of ICT in Education (PTP) SIG

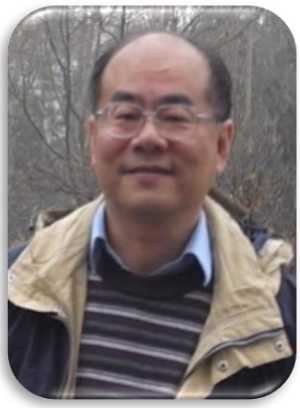
APSCE Webinar #28: Virtual Reality and Language Education: Research Trend and Teaching Practice

Date: 15 October 2022 (Saturday)
Speaker: Prof. Chunping ZHENG (Beijing University of Posts and Telecommunications)
Moderator: Prof. Vivian Wen-Chi WU (Asia University)
Curated by: APSCE Technology-Enhanced Language Learning (TELL) SIG

APSCE Webinar #29: Emulating Real-World Contexts Using Social Robots and IoT-Based Tangible Objects to Provide Embodied and Interactive Learning Experiences

Date: 09 November 2022 (Wednesday)
Speaker: Prof. Nian-Shing CHEN – National Taiwan Normal University, Taiwan
Moderator: Assoc. Prof. Patcharin Panjaburee (Khon Kaen University, Thailand)
Curated by: APSCE Development of ICT in the Asia-Pacific Neighbourhood (DICTAP) SIG

KEYNOTE SPEAKERS



Gwo-Dong CHEN

National Central University, Taiwan

Digital Theater for Situational Learning in the Classroom

To support situational learning in the classroom, the teachers need to arrange scenarios according to the teaching content and context so that the students can immerse in the scenarios to experience situational learning and teaching. Teachers commonly use Augmented reality AR, Virtual Reality VR, and 2D/3D computer screen displays to apply situational learning in the classroom. However, the existing digital reality mechanisms do not let the students see how they perform in the reality to do reflection learning and show it to their classmates so that the students feel responsible for learning and perform better to get recognition from their classmates. We designed and developed an alternative approach called Digital Theater to be used by teachers to apply situational learning in the classroom. Using video and skeleton capturing techniques, the video and sound of the students as actors can be captured, mixed, and put into the digital reality generated by the computer according to the context scenario of the learning content. The students formed groups to perform on the stage in the classroom. The Digital Theater system displays the performance as a stage drama that shows to the students who act as the actors or the audience in the classroom. The digital theater mechanism can be easily applied in an existing classroom. Moreover, the AI cognitive recognition mechanisms such as facial recognition, gesture recognition, speech recognition, and dialog language processing can be integrated into the digital theater. Therefore, the student's physical performance can be captured, recognized, and evaluated in the digital space. The digital space, including virtual humans and robots, can interact and respond based on the designed learning script. After that, Digital Theater was formed as a digital game room for students to explore. Novel learning design can be devised on the digital theater mechanism. The virtual human and physical robots can be integrated into the digital theater to improve the student's learning performance. The physical robot can be used as a learning and presentation tool for designing how to learn and presenting learning results.

Biography: Dr. Gwo-Dong Chen is a Chair Professor Department of Computer Science and the director of the research center of learning science and technology at National Central University, Taiwan. He obtained a BS, MS, and Ph.D. from the Department of Electrical Engineering, National Taiwan University, Taiwan, in 1979, 1982, and 2000 respectively. From 2004 to 2010, he was the CEO of the Taiwan National e-learning program office, which involved 13 ministries and organizations of the government. Dr. Chen was the director general of the Department of Science Education, National Science Council, Taiwan, from 2010 to 2013. His current research focuses on constructing a Digital Theater system to support the teachers can quickly build a digital learning space in the physical space in the classroom according to the learning context. The learning design is to let the learners immerse bodily in the digital space to participate in the learning activities based on situated learning and embodied cognition. Dr. Chen and his students got the best student paper from ICALT 2022 and ICALT 2020, the best technique paper from ICCE2021, the best student paper from GCCCE 2022 (in Chinese), and the best paper from GCCCE 2020 (in Chinese), respectively.

KEYNOTE SPEAKERS



Rebecca FERGUSON
Institute of Educational Technology (IET),
The Open University United Kingdom

Pedagogies for Today's World

The COVID-19 pandemic forced educators around the world to make a sudden move to remote learning, often without the benefit of appropriate training and resources. As we begin to adapt to the 'new normal', there are opportunities to incorporate the findings of more than 20 years research on online education into our daily practice. For the last decade, The Open University in the UK has been producing 'Innovating Pedagogy' reports. These introduce new approaches together with sound advice based on evidence, common sense, and clarity. The pedagogies range from small-scale innovations, which individual educators can try out in their classes, to sweeping trends that may shape education futures around the world. In her keynote, Professor Rebecca Ferguson, lead author on the report for several years, introduces some of the pedagogies most relevant for a world in which teaching increasingly takes place at a distance. These include hybrid models that combine face-to-face and online approaches, enriched realities that extend the possibilities for learning, microcredentials, online laboratories, virtual studios, wellbeing education, and student-led analytics.

Biography: Rebecca Ferguson is Professor of Learning Futures in the Institute of Educational Technology (IET) at The Open University in the UK, and a senior fellow of the Higher Education Academy. Her primary research interests are educational futures and how people learn together online. She has worked on the Innovating Pedagogy Reports since the series was founded in 2012, including two years as lead author. This highly cited series of high-profile annual reports explores new forms of teaching, learning and assessment in order to guide educators and policy makers around the world. Rebecca was an executive member of the Society for Learning Analytics Research (SoLAR) for four years and is currently one of the Editors in Chief of the Journal of Learning Analytics. Her work has been influential in shaping the field, supporting implementation across Europe, and promoting a focus on social learning analytics and on ethics. Her other current focus is microcredentials, on which she has worked as Academic Lead at The Open University, due to her extensive experience in related areas. This experience includes a leading role in the creation of Future Learn and work as Pedagogic Adviser for the company in its first years, OU PI on the European-Commission-funded European MOOC Consortium: Labour Market (EMC-LM) project, and current Academic Lead of the international Future Learn Academic Network.

KEYNOTE SPEAKERS



Ryan BAKER

University of Pennsylvania, United States

When Might a Detector Generalize?

Abstract: Machine-learned based detectors have become an increasingly important part of contemporary AIED systems, measuring and/or predicting constructs ranging from knowledge, to disengagement, to affect, to stopout. However, often when models are developed, they are only tested to a very limited degree (usually just on held-out students from the original data set) and are then used in different situations without further evaluation. In this talk, I will discuss evidence around when detectors generalize -- and when they don't -- in terms of student identity and changes in the learning system itself, using examples from multiple studies in our research group spanning from stopout, to gaming the system, to wheel-spinning, to affect. I will offer some simple guidelines about the situations that seem to be linked to successful model generalization and propose some steps forward for better understanding this challenge.

Biography: Ryan Baker is Professor at the University of Pennsylvania, and Director of the Penn Center for Learning Analytics. His lab conducts research on engagement and robust learning within online and blended learning, seeking to find actionable indicators that can be used today but which predict future student outcomes. Baker has developed models that can automatically detect student engagement in over a dozen online learning environments, and has led the development of an observational protocol and app for field observation of student engagement that has been used by over 150 researchers in 7 countries. Predictive analytics models he helped develop have been used to benefit over a million students, over a hundred thousand people have taken MOOCs he ran, and he has coordinated longitudinal studies that spanned over a decade. He was the founding president of the International Educational Data Mining Society, is currently serving as Editor of the journal Computer-Based Learning in Context, is Associate Editor of the Journal of Educational Data Mining, was the first technical director of the Pittsburgh Science of Learning Center Data Shop, and currently serves as Co-Director of the MOOC Replication Framework (MORF). Baker has co-authored published papers with over 400 colleagues.

KEYNOTE SPEAKERS



Su Luan Wong

Universiti Putra Malaysia, Malaysia

From theory to practice: Igniting Students' Interest in Educational Technology through Interest Driven Creator (IDC) Theory

IDC theory owes its origins to a group of prominent Asian researchers concerned with the worrying trends of students' declining interest in learning. Indeed, their concerns about examination-driven education have struck a chord among scholars, especially in the Asian context. IDC theory posits that once interest in learning is piqued through interest-driven learning activities, students will be engaged in the knowledge creation process. Learning habits then ensue through repetition of the aforesaid process in the students' daily routine. This keynote focuses on interest-driven learning with the main aim of bridging theory and practice. In this talk, I will illustrate the trajectory of IDC theory in Asian classrooms since its inception in 2014. I will then share the findings of a recent study focusing on the first loop — interest. I will showcase empirical data illustrating how interest in educational technology can be developed through learning activities designed based on IDC tenets. I will end my talk by highlighting the key characteristics that emerged from the interest loop and how teacher education can benefit from IDC theory.

Biography: Dr. Su Luan Wong is a Professor at the Faculty of Educational Studies, Universiti Putra Malaysia (UPM). She has extensive teaching experience in educational technology, teacher education and teaching methods; and regularly conducts teaching and learning related training courses for educators. She is an active and passionate researcher and has published more than 200 scholarly papers with a special focus on teaching and learning in ICT. She has served as the Principal Associate Editor of the Asia-Pacific Educational Researcher Journal (SSCI), and Associate Editor of the Research and Practice in Technology Enhanced Learning (Scopus). She was also the Editor-in-Chief of the Pertanika Journal of Social Sciences and Humanities (Scopus). In recognition of her active role as a scholar, she was elected as an Executive Council member for the Asia-Pacific Society for Computers in Education (APSCE) in 2006, a post she still holds today. In 2011, she established a Special Interest Group — Development of Information and Communication Technology in the Asia-Pacific Neighbourhood (DICTAP) under the auspices of APSCE to bridge the research gap between scholars from developing and developed countries. Her hard work and dedication to serving the research community have led to a highly coveted association with APSCE as the President for 2016-2017.

THEME-BASED INVITED SPEAKERS

Hyo-Jeong SO
Ewha Womans University, Korea



When Mobile Learning and Learner Data Meet – Possibilities and Challenges

The field of mobile learning has expanded significantly with the wide adoption of mobile devices and seamless network connections. It is time to reflect on whether the promises of mobile seamless learning have been realized and whether mobile learning has transformed the culture of teaching and learning beyond easy access. In this talk, I will discuss some inherent tensions and challenges that researchers in the field of mobile learning are likely to experience when they adopt learner data-driven approaches in their design and analysis. Using examples from my projects, I will emphasize the necessity of building models of Asian learners and tackling the challenges of interdisciplinary knowledge. I will discuss these points based on my recent research on designing affective feedback mechanisms for Korean learners and diversity-based recommendation systems for mobile learning.

Biography: Dr. Hyo-Jeong So is Professor in the Department of Educational Technology, Ewha Womans University in Korea. She received her Ph.D. degree from Instructional Systems Technology, Indiana University. Her main research interests include mobile learning, computer-supported collaborative learning (CSCL), and informal learning. She is particularly interested in examining how to integrate emerging technologies for teaching and learning from collaborative knowledge building perspectives. Currently, she is serving as Editor-in-Chief of Research in Practice in Technology-Enhanced Learning (RPTEL), and associate editors of Learning: Research and Practice, and IEEE Transactions on Learning Technologies. She has conducted several research projects on emerging technologies in education funded by Google, Microsoft, and the Korean National Research Foundation and international development projects in Indonesia and Mongolia. She has published several research papers in the international journals and working paper series with UNESCO.

THEME-BASED INVITED SPEAKERS



Jan VAN AALST

The University of Hong Kong, Hong Kong

What Have We Made? Examining Knowledge Building in the Post-Truth Era

When I was a graduate student in the 1990s, online discussion environments were called "CSILE-like," after the original environment that Marlene Scardamalia and colleagues developed to support knowledge building. Knowledge building was—and is—an effort to introduce students to a culture and the processes of knowledge creation. Nowadays, discussion forums are ubiquitous: Twitter, WeChat, Facebook, Whatsapp, TikTok, Instagram, and more. There also is an abundance of personal websites, vlogs, and podcasts. And where I live, newscast report on their own opinion polls and at talk shows scientists sit next to comedians and restaurant operators to debate the science of climate change or the Corona pandemic. People often do some Googling before they speak or consult their social networks. At first sight at least, all this it does look somewhat like knowledge building—e.g., democratization of knowledge and idea diversity. In this presentation I consider knowledge building as an educational model considering these developments. Is it an idea that has been passed by, or can it be a way forward? What kinds of educational development are then needed?

Biography: Jan van Aalst is a recently retired associate professor at the University of Hong Kong, where he also has served as associate dean for research. He continues to be a Visiting Scholar at the University of Twente. He has served as Co-editor in Chief of Journal of the Learning Sciences, and as member of the Board of Directors of the International Society of the Learning Sciences (ISLS). He was elected Fellow of ISLS in 2020. Van Aalst's research has focused on knowledge building, an educational approach that emphasizes students' agency and metacognition, collaborative learning, and inquiry-based learning within a community. His team has studied pedagogical designs for knowledge building in schools in Hong Kong and China and has developed web-based tools, the Knowledge Connections Analyzer, and practices to support students' self-directed assessment of their knowledge building. His work involves students across a wide range of achievement levels. His work on knowledge building is published in International Journal of Computer-Supported Collaborative Learning, Journal of the Learning Sciences, and American Educational Research Journal. In 2013 he created an. M. Ed. program focusing on putting the key findings from the field of the learning science into educational practice, which has been popular with teachers from China, Hong Kong and farther abroad. This program emphasized educational design and partly led to a co-authored reference book on research methods, Learning Sciences Research for Teaching (Routledge). In the last few years Van Aalst's interests have begun to emphasize equity and social justice more, and he is currently collaborating on a long-range project to address these issues in an urban center in the Netherlands.

THEME-BASED INVITED SPEAKERS

Yu-Ju LAN
National Taiwan Normal University, Taiwan



New Research Trends in Emerging Technologies for Language Learning

Learning a new language other than one's first language (L1) is always challenging. It takes time, effort, focus, motivation, and sustained involvement. The use of language for communication and social interaction has always been a key competency in the 21st century. Technology plays a significant role in helping today's learners to acquire a language. Research on language learning has become highly interdisciplinary, drawing upon a range of fields such as psychology, education, neuroscience, and recently machine learning. TELL study follows the theoretical and methodological advances in education, cognitive science, and neuroscience. Moreover, TELL research depends heavily on the latest technologies. In this speech, I will illustrate the new paradigm of language learning in the modern digital era, examine TELL based on various methods and platforms supported by new digital technologies such as AI, mobile computing, VR, and digital games. It will also include how emerging technologies can further extend the influence of TELL research.

Biography: Dr. Yu-Ju Lan is a Research Chair Professor in the Department of Chinese as a Second Language at National Taiwan Normal University. She is currently the Editor-in-Chief of Educational Technology & Society, Associate Editor of Language Learning & Technology, and on the editorial board of Ampersand. She is the president of the Taiwan Pedagogy and Practice in TELL Association. Her research interests include technology-enhanced foreign language learning, language learning in virtual worlds, mobile learning, and online synchronous teacher training. She has published nearly 50 SSCI journal papers. Dr. Lan has proposed the principles of designing tasks and VR contexts based on empirical evidence. As the need for using VR in learning grows, her pioneering works could provide essential implications for academia, education, and industries. For her outstanding research performance, she was awarded the Outstanding Research Award by the Ministry of Science and Technology (MOST), Taiwan, in 2022.

Table of Content

No	Title of papers	Track	Page
C1 : Artificial Intelligence in Education/Intelligent Tutoring System (AIED/ITS) and Adaptive Learning.			
1	Metacognitive Accuracy in Homework Assignments, Time-Limited Quizzes, and Learning Objectives *Huanyi CHEN & Paul A.S. WARD*	C1	1
2	Investigation on Practical Effects of the Explanation in a K-12 Math Recommender System *Yiling DAI, Kyosuke TAKAMI, Brendan FLANAGAN & Hiroaki OGATA*	C1	7
3	Transferable Student Performance Modeling for Intelligent Tutoring Systems *Robin SCHMUCKER & Tom M. MITCHELL*	C1	13
4	Automated Matching of Exercises with Knowledge Components *Zejie TIAN, Brendan Flanagan , Yiling DAI & Hiroaki Ogata*	C1	23
5	Using Unsupervised Machine Learning to Model Taiwanese High-School Students' Digital Distraction Profiles Concerning Internet Gaming Disorder *Yu-Lin HO, Chien CHOU, Chen-Hsuan LIAO & Jiun-Yu WU*	C1	32
6	Use of Professor Comments in Predicting Student Success *Timothy H. BELL, Christel DARTIGUES, Florent JAILLET & Christophe GENOLINI*	C1	38
7	Learning Algorithm Implementation Structures for Multilabel Classification via CodeBERT *Karl Frederick ROLDAN, Gerd Lowell JANA, John Kenneth LESABA & Joshua MARTINEZ*	C1	44
8	Micro- and Macro-Level Features of NLP-Based Writing Tools in Higher Education *Michael BURKHARD, Sabine SEUFERT, Patcharin PANJABUREE, Chailerd PICHITPORNCHAI & Christina NIKLAUS*	C1	50
9	Investigating Bloom's Cognitive Skills in Foundation and Advanced Programming Courses from Students' Discussions *Joel J. W. LIM, Swapna GOTTIPATI & Kyong Jin SHIM*	C1	56
10	I will help you, but will you help me? How The Perception of a Teachable Agent May Influence Performance *Eva-Maria TERNBLAD, Magnus HAAKE & Betty TÄRNING*	C1	63

No	Title of papers	Track	Page
11	'I know that I clicked but not if I read': An Exploratory Study Comparing Data Traces and Self-Reports on Feedback Engagement *Eva-Maria TERNBLAD, Agneta GULZ & Betty TÄRNING*	C1	73
12	Exploring Relationships Between Temporal Patterns of Affect and Student Learning *Anthony F. BOTELHO, Seth A. ADJEI, Vedant BAHEL, & Ryan S. BAKER*	C1	83
13	Engagement Estimation using Time-series Facial and Body Features in an Unstable Dataset *Xianwen ZHENG1, Minh-Tuan TRAN, Koichi OTA, Teruhiko UNOKI & Shinobu HASEGAWA1*	C1	89
14	Investigating How Achievement Goals Influence Student Behavior in Computer Based Learning *J. M. Alexandra L. ANDRES, Stephen HUTT, Jaclyn OCUMPAUGH, & Ryan S. BAKER*	C1	95
15	How Much Support is Necessary for Self-Regulated Learning? *Antonija MITROVIC & Jay HOLLAND*	C1	101
16	Development of a Learning Companion Robot with Adaptive Engagement Enhancement *Bowei YAO, Koichi OTA, Akihiro KASHIHARA, Teruhiko UNOKI & Shinobu HASEGAWA*	C1	111
17	Practical Use of an Error-based Problem Presentation System in Mechanics *Nonoka AIKAWA, Shintaro MAEDA, Tomohiro MOGI, Kento KOIKE, Takahito TOMOTO, Isao IMAI, Tomoya HORIGUCHI & Tsukasa HIRASHIMA*	C1	118
18	Prerequisite-driven Q-matrix Refinement for Learner Knowledge Assessment: A Case Study in the Online Learning Context *Wenbin Gana & Yuan Sun*	C1	124
19	Transforming Brainwave Signals into Symbolic Strings Towards Academic Emotion Recognition *Judith AZCARRAGA & Juan Francesco SALCEDA*	C1	134
20	Considerate, Unfair, or Just Fatigued? Examining Factors that Impact Teacher *Ashish GURUNG, Anthony F. BOTELHO, Russell THOMPSON, Adam C. SALES, Sami BARAL & Neil T. HEFFERNAN*	C1	140

No	Title of papers	Track	Page
21	Adaptive Scaffolding to Support Strategic Learning in an Open-Ended Learning Environment *Anabil Munshi, Gautam BISWAS, Eduardo DAVALOS, Olivia LOGAN, Gayathri NARASIMHAM & Marian RUSHDY*	C1	150
C2 : Computer-supported Collaborative Learning (CSCL) and Learning Sciences (LS).			
22	Unveiling the Process of Collaborative Learning Through the Use of Digital Whiteboard Historical Action Logs *Hua Leong WA*	C2	157
23	The Development of Online Learning Readiness Scale for Junior Students (OLRS-J) in Taiwan *Chien CHOU, Ruo-Yu LI, Li-Fang KO & Min-Ling HUNG*	C2	164
24	Analysis of CSCL for Mathematical Proof Based on the Log Data of Learners' Verbal and Nonverbal Communications *Masataka KANEKO, Hironori EGI & Takeo NODA*	C2	167
25	Estimating Divergence-or-Convergence in Discussions based on Abstractness of Words in Utterance *Ryunosuke NISHIMURA, Taisei MURAOKA, Risa IHARADA & Hironori EGI*	C2	170
26	The Influence of Awareness of a Difference between Concept Maps on Transfer: Experimental Investigation on the Efficacy in Collaborative Learning *Shigen SHIMOJO, Yoshimasa OHMOTO, Junya MORITA & Yugo HAYASHI*	C2	176
27	The Effect of Individual Ideation before Discussion on Computer Supported Collaborative Argumentation in a Primary Classroom *Wenli CHEN, Qianru LYU, & Junzhu SU*	C2	182
28	Interaction and Monitoring Matter: Comparison of High and Low-performing Groups in CSCL *Wenli CHEN, Junzhu SU, Qianru LYU, Aileen CHAI, & Wei Liang TOH*	C2	192
29	Effects of Group Awareness Tools on Student Engagement and Enjoyment in Online Collaborative Writing *Xiaohui SONG, You SU, Jing REN & Rushi GONG*	C2	200
30	Understanding Areas of Parallelograms Through Virtual Geometrical representations: A Pilot Study *Eva-Maria TERNBLAD*	C2	206

No	Title of papers	Track	Page
31	Understanding Sociomaterial Encounters in Collaborative Creative Making: A Case Study from a Makerspace in India *Alekh V & Chandan DASGUPTA*	C2	210
32	Unpacking Emergent Creativity in Online Collaborative Making *Alekh V & Chandan DASGUPTA*	C2	220
33	Support System for Understanding How Teammates Grasp Game Situations *Kazuteru Sakamoto, Kota Kunori & Tomoko Kojiri*	C2	226
34	GWpulse: Supporting Learner Modelling and Group Awareness in Online Forum with Sentiment Analysis *Yuta NAKAMIZO, Changhao LIANG, Izumi HORIKOSHI, Rwitajit MAJUMDAR, Brendan FLANAGAN & Hiroaki OGATA*	C2	230
35	Improving Spatial Perspective Taking Ability of Middle School Students Using Augmented Reality *Anveshna SRIVASTAVA & Chandan DASGUPTA*	C2	233
36	Noticing Critical Information for Productive Uncertainty Management During Engineering Design Activity *Navneet KAUR & Chandan DASGUPTA*	C2	240
37	Investigating the Role of Students' Feedback-Seeking Actions in Harnessing and Adapting to the Distributed Competence *Narasimha SWAMY, Anveshna SRIVASTAVA & Chandan DASGUPTA*	C2	243
38	Understanding Learners' Negotiation Processes During Ill-Structured Engineering Estimation Problem-Solving *Ulfa KHWAJA & Sahana MURTHY*	C2	253
C3 : Advanced Learning Technologies (ALT), Learning Analytics and Digital Infrastructure			
39	Monitoring of Learners' Activities in Software Structure Design Exercises *Yasuhiro NOGUCHI, Kanta INOUE, Satoru KOGURE, Koichi YAMSHITA & Tatsuhiro KONISHI*	C3	259
40	Innovative and Effective Spreadsheet Tool for Learning Sentiment Analysis and Prediction *Michelle CHEONG & Jean CHEN*	C3	262

No	Title of papers	Track	Page
41	Topic-Based Representation of Learning Activities for New Learning Pattern Analytics *Jinghao WANG, Tsubasa MINEMATSU, Yuta TANIGUCHI , Fumiya OKUBO & Atsushi SHIMADA*	C3	268
42	Assessment of At-Risk Students' Predictions From E-Book Activities Representations In Practical Applications *Erwin D. LOPEZ Z., Tsubasa MINEMATSU, Yuta TANIGUCHI, Fumiya OKUBO & Atsushi SHIMADA*	C3	279
43	Topic-Level Social Network and Language Correlation in Course Discussion Forums *Ezekiel Adriel LAGMAY & Maria Mercedes RODRIGO*	C3	289
44	Assessment of Students' Feedback Behavior in a Game-Based Automated Feedback System- A Cross-Cultural Replication Study *Annika SILVERVARG, Kristen BLAIR, Maria CUTUMISU and Agneta GULZ*	C3	292
45	Construction of a Role-Play Style Japanese Learning Support System that Allows Teachers to Edit Dialogue Situations *Satoru KOGURE, Yudai KAMATA, Yasuhiro NOGUCHI, Koichi YAMASHITA, Tatsuhiro KONISHI, & Makoto KONDO*	C3	302
46	Efficacy of an Online Course to Build Mindfulness in Adolescents *Renuka RAUTELA, Anurag DEEP, Mayank SHARMA & Nandini CHATTERJEE*	C3	305
47	Detection of At-Risk Students in Programming Courses *Ikkei IGAWA, Yuta TANIGUCHI , Tsubasa MINEMATSU , Fumiya OKUBOa & Atsushi SHIMADA*	C3	308
48	Learning Support System Visualizing Relationships Among Classes and Objects Based on Teacher's Intent of Instruction *Koichi YAMASHITA, Yusuke SUZUKI, Satoru KOGURE, Yasuhiro NOGUCHI, Raiya YAMAMOTO, Tatsuhiro KONISHI & Yukihiro ITOH*	C3	314
49	Combining Data and Human Intelligence through Predictive Visual Analytics to Improve Educational Assessments *Yancy Vance PAREDES & I-Han HSIAO*	C3	317
50	The Influence of a Virtual Physics Experiment Learning Environment on Grade 9 Students' Motivation towards Physics Learning *Yushi HUANG, Alex Wing Cheung TSE*	C3	320

No	Title of papers	Track	Page
51	Extracting Students' Self-Regulation Strategies in an Online Extensive Reading Environment using the Experience API (xAPI) *Chia-Yu HSU, Izumi HORIKOSHI, Huiyong LI, Rwitajit MAJUMDAR & Hiroaki OGATA*	C3	326
52	Process Models Enhancement with Trace Clustering *Wiem HACHICHA, Ronan CHAMPAGNAT, Leila GHORBEL & Corinne Amel ZAYANI*	C3	332
53	Applicability and Reproducibility of Peer Evaluation Behavior Analysis Across Systems and Activity Contexts *Izumi HORIKOSHI, Changhao LIANG, Rwitajit MAJUMDAR & Hiroaki OGATA*	C3	335
54	Association Analysis on Open-Ended Concept Maps using Data Mining *Didik Dwi PRASETYA, Setiadi Cahyono PUTRO, Muhammad ASHAR, Saida ULFA, & Tsukasa HIRASHIMA*	C3	346
55	Comparing Short-Term and Long-Term Online Courses Using the Kano Model and Neural Network Language Models *Daniel Moritz MARUTSCHKE & Yugo HAYASHI*	C3	352
56	A Curriculum Package of Social and Ethical Concerns in Metaverse Ecosystem in School Education Based on Bloom's Taxonomy *Percy Lai-Yin KWOK, Joe Kwong-Ngai CHEN & Jordan Tsz-Chun FUNG*	C3	358
57	PyGuru: A Programming Environment to Facilitate Measurement of Cognitive Engagement *Daevesh SINGH, Hema SUBRAMANIAM, & Ramkumar RAJENDRAN*	C3	363
C4 : Classroom, Ubiquitous, and Mobile Technologies Enhanced Learning (CUMTEL)			
58	Modelling Physical Activity Behaviour Changes for Personalised Feedback using a Health Education Application *Claudio DIAZ, Olivier GALY & Corinne CAILLAUD & Kalina YACEF*	C4	369
59	Research on Classroom Interaction Behavior in a Mobile Tablet-based Unit Teaching Setting *Jian KANG & Ruixue LIU*	C4	375
60	A Technique for Tracking the Reading Rate to Provide Students' Learning Feedback *Fuzheng ZHAO & Chengjiu YIN*	C4	381

No	Title of papers	Track	Page
61	An Evaluation of the Online Course on Climate Change and Emotion *Anurag DEEP*, Aditi PATHAK & Nandini CHATTERJEE*	C4	387
62	Augmented Reality Applications in the Classroom: Teachers' Experience *Hajar AL OMAIR, Antonette MENDOZA, Marie BODÈN & Nilufar BAGHAEI*	C4	392
63	Design and Effect of AR Sign Language Teaching Software Based on Natural Interaction *Jiangxu LI, Peng ZHANG, Junting CHANG & Su CAI*	C4	402
64	Designing a Recommender System for Mobile Applications Focusing on Relative Importance Weights of Learner-related Variables *Woorin HWANG, Hyo-Jeong SO, Chiyoung SONG, & Hyeji JANG*	C4	410
65	Learning Through Patient Stories: Fostering Sensemaking in Technology Enhanced Learning Environment for Clinical Diagnosis *Sunita RASTE & Sahana MURTHY*	C4	413
66	Unpacking Contextual Parameters Influencing the Quality of Personalized Adaptive Learning Edtech Applications *Gomathy SOUNDARARAJ, Vishwas BADHE, Ishika ISHIKA, Meera PAWAR, Chandan DASGUPTA, Sahana MURTHY*	C4	419
67	Instructional Design on Data Visualization Model of Using AR Sandbox Apps in Learning and Teaching Geography *Aaron Hok-him LIU & Percy Lai-yin KWOK*	C4	429
68	Pupils' Learning Characteristics, Self-directed Learning Ability, and Mathematical Ability by Applying A Gamified Math Platform for Distance Learning *Euphony F. Y. YANG & Tak-Wai CHAN*	C4	433
C5 : Educational Gamification and Game-based Learning (EGG)			
69	Learning Computational Thinking Efficiently with Block-based Parsons Puzzles *Jeff BENDER, Alex DZIENA & Gail KAISER*	C5	439

No	Title of papers	Track	Page
70	*For People and Planet*: A Pilot Study of an Educational Mobile Game on the Sustainable Development Goals *Maria Mercedes T. RODRIGO, Louise Marie C. TULAYBA, Marianne U. LIM, Marlon C. MORALEJO, Ananea ARCEGA, Maria Celeste A. MARASIGAN-LASCANO & Allyah Elizabeth ORIO*	C5	449
71	An Analysis of Filipino Primary School Learners' Game Experience and STEM Interest within Minecraft *Jonathan CASANO & Maria Mercedes T. RODRIGO*	C5	455
72	Ludonotation: A Visualization System of Recording and Analyzing Interactive Strategic Gaming Behaviors *Ju-Ling SHIH*	C5	461
73	Design a Robot as a Double with Micro-expression to Participate in a Virtual Situational Learning Environment and its Effect on Students' Learning Performance *Vando Gusti Al HAKIM, Su-Hang YANG, Jen-Hang WANG, Yi-Jing LIa, Yi-Hsin CHEN & Gwo-Dong CHEN*	C5	467
74	Achievement, Behaviors, and STEM Interest of Frustrated and Bored Learners Using Minecraft *Maricel A. ESCLAMADO, Maria Mercedes T. RODRIGO & Jenilyn A. CASANO*	C5	477
75	Two-way Knowledge Transfer Among University Academics, Young Entrepreneurs, NGOs and Students in STEM and IoT Metaverses: Conceptual Model, Research Agendas and Contextual Challenges *Percy Lai-Yin KWOK, Joe Kwong-Ngai CHEN & Jordan Tsz-Chun FUNG*	C5	487
76	What Do Teachers Want to Know about Game-Based Learning Analytics: Cross-Case Study *Yoon Jeon KIM & Jennifer SCIANNA*	C5	493
77	Changing Students' Perceptions of a History Exploration Game Using Different Scripts *Stefan SLATER, Ryan S. BAKER, David GAGNON, Erik HARPSTEAD, Juliana Ma. Alexandra L. ANDRES, Luke SWANSON*	C5	499
78	Talk & Play: A VR Application To Improve Volume And Silences Use In Verbal Communication *Maria Esther RUIZ-CAPILLAS, Alejandro ROMERO-HERNANDEZ & Borja MANERO*	C5	505

No	Title of papers	Track	Page
79	Impact of Gaming on the Mental Well-Being and Academic Performance among High School Students *Aaron James CAPINPIN, Roi Victor ROBERTO, Ramon Diego TAN & Judith AZCARRAGA*	C5	511
80	Shelter GO: Multiplayer Location-based Game for Learning Evacuation *Hiroyuki MITSUHARA & Masami SHISHIBORI*	C5	514
C6 : Technology Enhanced Language Learning (TELL)			
81	Evaluating the Performance of Chinese Multi-Label Grammatical Error Detection Using Deep Neural Networks *Tzu-Mi LIN, Chao-Yi CHEN, Lung-Hao LEE & Yuen-Hsien TSENG*	C6	524
82	An Exploration of a Novice Kindergarten Teacher's Enactment of Multiliteracies Pedagogy during the Pandemic: A Case Study of a Virtual Kindergarten Classroom *Fengchao YU, Lian CHANG & Mi Song KIM*	C6	527
83	Using Group Awareness Tools to Enhance Students' Behavioral and Cognitive Engagement with Peer Feedback in Online Collaborative Essay Writing *Jing RENa, You SU, Xiaohui SONG & Haizhen GUO*	C6	537
84	Research on Activity Design and Application of AR Gamification Resources to Promote Children's Second Language Learning *Min-Sheng FAN, Zhen-Hua LIU & Xian-Ping JIN*	C6	543
85	Ask4Summary: A Summary Generation Moodle Plugin Using Natural Language Processing Techniques *Mohammed SALEH, Maria F. IRIARTE & Maiga CHANG*	C6	549
86	What Criteria are Important for Evaluating The Quality of English Language Learning Edtech Products? Evidence from Literature *Minu NANDAKUMAR, Leena BHATTACHARYA & Sahana MURTHY*	C6	555
87	Explainable English Material Recommendation Using an Information Retrieval Technique for EFL Learning *Kensuke TAKII, Brendan FLANAGAN, Huiyong LI, Yuanyuan YANG & Hiroaki OGATA*	C6	561
88	The Effects of Learning with Digital Storytelling on Classroom Engagement in a Grade 6 English Class *Yingxue LIU & Alex Wing Cheung TSE*	C6	571

No	Title of papers	Track	Page
89	A Learning Path Recommendation System for English Grammar Quiz Using Knowledge Map *Naomichi TANIMURA, Kensuke TAKII, Brendan FLANAGAN & Hiroaki OGATA*	C6	581
C7 : Practice-driven Research, Teacher Professional Development and Policy of ICT in Education (PTP)			
90	Role of Peer Assessment in Facilitating Computational Thinking Among Pre-Service Teachers *Xin Pei VOON, Su Luan WONG, Lung Hsiang WONG, Mas Nida MD KHAMBARI & Sharifah Intan Sharina SYED ABDULLAH*	C7	584
91	Authentic Assessments for Digital Education: Learning Technologies Shaping Assessment Practices *Tristan LIM, Swapna GOTTIPATI & Michelle CHEONG*	C7	587
92	The Comparative Effects of 'What if/what if not' and 'Main Ideas' Scaffolding for Online Student Question-generation on Science Learning *Fu-Yun YU & Hsiu-Chen CHEN*	C7	593
93	When the School Door Closes, Do Teachers Open a Window? Using Mixed Methods to Investigate Teachers' Online Teaching Practices and Momentary Experiences in Crisis *Jo-Chi HSIAO, Chao-Yang CHENG & Sunny S. J. LIN*	C7	598
94	Agile Learning as a Solution for Class Absence *Maria Francesca ROIG-MAIMÓ, Ramon MAS-SANSÓ*	C7	607
95	Evaluation of the Quality of The Educational Software From the Perspective of Experiences of Italian and Polish Pre-Service Teachers *Łukasz TOMCZYK, Laura FEDELI, Anna WŁOCH, Pierpaolo LIMONE, Monika FRANIA , Piergiorgio GUARINI, Michał SZYSZKA, Maria Lidia MASCIA, & Joanna FALKOWSKA*	C7	613
96	The Perspectives of Distance Learning in Higher Education: Lessons Learned from the COVID-19 Pandemic *Dariusz DYMEK, Mariusz GRABOWSKI & Grażyna PALIWODA-PĘKOSZ*	C7	619
97	Media Exposure and Students' Attitude as Mediators Between Subjective Norms and Choice Intention for Science, Technology, Engineering, and Mathematics Careers *Tiny Chiu Yuen TEY, Priscilla MOSES & Phaik Kin CHEAH*	C7	629

No	Title of papers	Track	Page
98	Rural Teachers' Professional Development under the Model of "Internet plus" Volunteer Teaching in Normal College *Jian ZHAO & Wenjing YANG*	C7	639
99	Exploring Predictive Indicators of Reading-Based Online Group Work for Group Formation Teaching Assistance *Changhao LIANG, Izumi HORIKOSHI, Rwitajit MAJUMDAR, Brendan FLANAGAN & Hiroaki OGATA*	C7	642
100	The Budding Botanist Paradox: Automating Human Inquiry with Immersive Technology *Eric HAWKINSON*	C7	648
101	Coding Initiative Provides Different Approaches to Inspire Girls for Programming *Sarina GURSCH*	C7	654
102	An Integrative Evaluation of STEM Policy in Secondary Education: Empirical Reflections, Conceptual Discussions and Policy Implications *Percy Lai-yin KWOK & Heidy Jia-hui WANG*	C7	660
103	A Pedagogy in STEM Classrooms for Primary Students to Develop Knowledge of Electric Circuits and Problem-Solving Skills *Siu Cheung KONG & Wai Ying KWOK*	C7	666
104	Computational Thinking Development: Validating an Instrument for Self-regulated Learning Using Animation *Siu-Cheung KONG, Ye HU & Ming LAI*	C7	675
105	Rural Teacher ICT Literacy Professional Learning and Teaching Practice in China: A Multiple Case Study Investigation *Yanfen HUANG , Sharifah Intan Sharina SYED-ABDULLAH, Nurul Nadwa ZULKIFLI & Norliza GHAZALI*	C7	681
106	How Teachers Used ICT in Education in 2015 and 2022 in Croatia: A Cross-Sectional Study *Ivica BOTICKI, Marcela MATAS, Lana JURAK & Bruno SEGVIC*	C7	684
107	The Influence of an Online Community of Practice on Professional Development for Online Teaching: Case Study on Kindergartens *Yuk Mui HEUNG & Alex Wing Cheung TSE*	C7	690

No	Title of papers	Track	Page
108	Development and Validation of a brief Digital Pedagogy Competency Scale (DiPeCoS) *Shraddha RAWAT, Shreya TIWARI, Mayank SHARMA & Nandini CHATTERJEE*	C7	700
109	How to Cluster Students Based on Their Digital Competence for Learning? *Külli KALLAS & Margus PEDASTE*	C7	709
110	Copyright in the Scope of Education *Branko KIRIN & Rwitajit MAJUMDAR*	C7	715
111	A First-order Action Research Study to Uncover Students' Conceptual Gaps in an Online Statistics Course using Extended Matching Questions *Ishaan TANEJA, Prajish PRASAD, & Jayakrishnan WARRIEM*	C7	720
112	Teachers' Classroom Digital Technology Integration Practices *Doris Kristina RAAVE, Eric ROLDAN ROA, Margus PEDASTE & Katrin SAKS*	C7	730

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Metacognitive Accuracy in Homework Assignments, Time-Limited Quizzes, and Learning Objectives

Huanyi CHEN[a*] & Paul A.S. WARD[a]
Department of Electrical and Computer Engineering, University of Waterloo, Canada
*huanyi.chen@uwaterloo.ca

Abstract: How confident students are about their solutions is essential to instructors. It reflects how students understand the course materials and how they learn in the course. Studies usually use "Metacognitive Accuracy" to refer to the relationship between a student's confidence or estimated performance and actual performance.

In our study, 170 students estimated their performances for individual tasks in homework assignments and their total performances in time-limited quizzes. The study was conducted in an upper-year database course. In this paper, we propose research questions related to students' estimations in homework assignments, time-limited quizzes, and different learning objectives. We found that students can estimate their homework assignments and time-limited quizzes differently. They also expose different over- and underestimation patterns under different learning objectives.

Keywords: Metacognitive Accuracy, Metacognition, Learning Objective, Self-assessment, Student Performance

1. Introduction

Understanding one's ability to solve a problem is one's metacognitive skill. A good metacognitive skill can guide the self-regulated learner to learn well. Relevant studies are gaining more and more interest in computing education in recent years (Prather et al., 2020). A commonly accepted hypothesis is that high-performing students have higher metacognitive accuracy than low-performing students (Aghababyan, Lewkow, & Baker, 2018; Hacker, Bol, & Bahbahani, 2008; Hacker, Bol, Horgan, & Rakow, 2000; Harrington, Peng, Jin, & Khan, 2018; Murphy & Tenenberg, 2005). A possible reason why low-performing students tend to have poor metacognitive accuracy is that they do not know what they do not know, and they tend to overestimate their marks compared to those who perform well (Harrington et al., 2018; Kruger & Dunning, 2000). Instructors may conduct early intervention if they have such observations and thus improve students learning.

Although there are several valuable observations, we consider some aspects that can be further explored. First, we wondered if the evaluation format will affect students' performance estimation. For example, homework assignments usually give students days to finish, while the time is usually limited for quizzes. Limit time may lead to stress. Under stress, students may make an inaccurate estimation more often than usual (Reyes, Silva, Jaramillo, Rehbein, & Sackur, 2015). Second, prior studies rarely look into different learning objectives in a course. For example, in a database course, the learning objectives may include aspects from conceptual and theoretical knowledge to practical development and implementation skills (Pahl, Barrett, & Kenny, 2004). Given that the difficulties of different learning objectives may vary from each other, students may have different estimations for different types of learning objectives (John L. Nietfeld, Li Cao, & Jason W. Osborne, 2005). It may suggest learners to take different self-regulated learning strategies on different learning objectives.

Our study was conducted within a database course, and we focused on two research questions:

RQ1: *Is there any difference in the metacognitive accuracy between homework assignments and time-limited quizzes?*

RQ2: *What is the metacognitive accuracy for different types of learning objectives?*

2. Related Work

A commonly accepted hypothesis is that high-performing students have higher metacognitive accuracy than low-performing students (Hacker et al., 2008, 2000; Harrington et al., 2018; Murphy & Tenenberg, 2005). For example, Harrington et al. asked students to predict their grades on each question of their final examination in a CS1 course (Harrington et al., 2018). They analyzed the summed predicted grade and found a relatively strong correlation that students who performed poorly in the course were more likely to overestimate their marks than those who performed well. Stephens-Martinez asked students to predict before and after exams, and they found that low performers were worse at predicting than other performers (Stephens-Martinez, 2021). Lee & Liao also recognized similar findings that weak students tend to be more confident than strong students when given identical actual scores (Lee & Liao, 2021).

Those studies primarily collect the metacognitive data with the same time-restriction condition. For example, Lee & Liao used Google Forms without time restrictions (Lee & Liao, 2021); while Hacker et al. collected the data from exams (Harrington et al., 2018), which had a time restriction for students to make their score estimations. However, we consider that students may estimate differently under time stress. Therefore, in our study, we have two different types of time restrictions. Homework assignments allowed students to finish within days, while time-limited quizzes were typically 50 minutes to 1 hour each. Students may show different behavior in their performance estimation under these two time-restriction conditions. We will discuss it in this paper.

John L. Nietfeld et al. found that the difficulty of the question is related to students' metacognitive accuracy (John L. Nietfeld et al., 2005). However, a learning objective may contain test items at multiple difficulty levels, and it is not easy to categorize them accordingly. To our knowledge, we did not see any paper mentioning how students estimate their performance under different types of learning objectives. Our paper will present our findings on how students estimate their performance over coding practice, theory, theory & practice, and design questions.

3. Course Background

This study was conducted in a database course in an R1 institution (name omitted for ethics reason) in Canada. The learning objectives of this course cover a wide range of topics that include but are not limited to entity-relationship models, relational algebra, functional dependencies, basics of SQL, and performance.

There were three homework assignments, and three quizzes were used in the study from 170 students who made performance estimations for all assignments and quizzes. The sample solutions to these homework assignments were provided soon after the due date, with a detailed explanation of the thinking process regarding how the answers were written up, hoping to help students gain helpful feedback before the marking is done. The quizzes were used similarly to evaluate students' knowledge using exams and help the instructional team better understand how students learned.

The assignments and quizzes were released alternatively, meaning assignment 1, quiz 1, assignment 2, quiz 2, assignment 3, then quiz 3. There was a fourth assignment in the course. However, since we cannot access the individual score of its tasks while we can for other assignments, so it was excluded from the study. The questions in a quiz were primarily related to the same learning objectives covered in its previous assignment. The only exception is that quiz 3 contains one question to cover some learning objectives in the fourth assignment. Since we focused on metacognitive accuracy within assignments and quizzes separately, we did not exclude that quiz. The solutions for quizzes were not provided to students.

Depending on the difficulties, homework assignments gave students a different number of days ranging from 10 days to 14 days to finish. There were multiple tasks in each assignment. A task in an assignment will have a clear section name that connects to a learning objective. The last task of each assignment is a metacognition task, which asks students to provide their estimated performances out of 100 percent for every other task.

All quizzes were open-book. The first two quizzes were 50 minutes each, and the last quiz was 60 minutes. A short extension period was provided for every quiz, but with a penalty (1% per minute). The last question of each quiz is a metacognition question, which asks students to provide their

estimated performance out of 100 percent for the entire quiz. Students could see how many points there were for each question in a quiz, but they cannot see that in assignments, even the total points.

The metacognition task was a blip in the grade, and its effect was lost in the final grade as the instructor rounded up the grade to the least large letter grade equivalent. However, this process is hidden from students.

4. Results

For all the actual performance discussed in this section, penalties were not included, as we consider students were not likely to include the penalty in their estimation.

4.1 RQ1: Is there any difference in the metacognitive accuracy between homework assignments and time-limited quizzes?

This research question aims to see if students estimate their performances differently under different assignment formats, with or without time constraints. We used two sample two tailed Welch's t-test in our experiments.

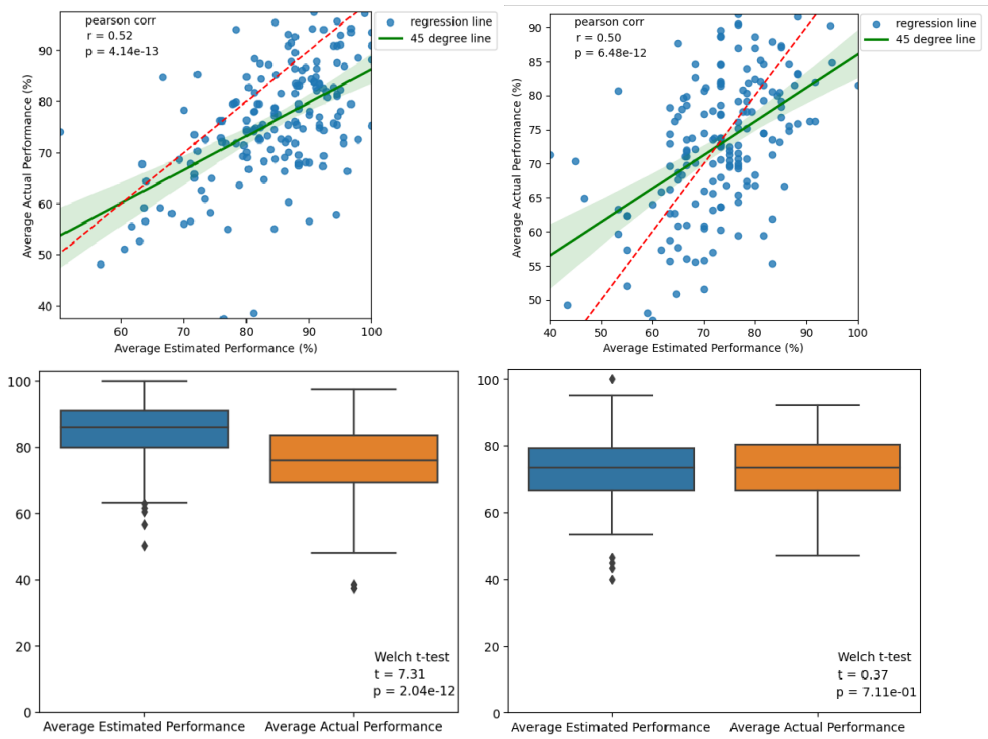


Figure 1. Average Estimated Performances and Average Actual Performances of Each Student on Assignments (left) and Quizzes (right) separately

Figure 1 (top-left and top-right) tell us that the correlations between students' average estimated performances and their average actual performances is at a similar level for homework assignments ($r = 0.52$) and time-limited quizzes ($r = 0.50$). A predictive model may be built based on this correlation.

The interesting finding comes from Figure 1 (bottom-left and bottom-right), which are box plots of the absolute average values of students' estimated performances and their actual performances. From Figure 1 (bottom-left), it is showing that when estimating their assignment performances, students have a significant trend to overestimate their performance ($t = 7.31$, $p < 0.01$); however, students tend to make accurate estimation when it comes to time-limited quizzes ($t = 0.37$, $p > 0.05$), as shown in Figure 1 (bottom-right). A significant difference in the decrease between students' actual performance and their estimation was also observed ($p = 0.005 < 0.05$).

For homework assignments, since students were given more time to finish, they could refer to course materials or ask for help from the instructional team whenever they faced any difficulties in

completing the assignment. The online discussion forum shows that the average response time in the course was half an hour. Students may gain further understanding as well as confidence as they get help. Therefore, they may have more confidence in their assignment answers, although the answers may not be entirely correct.

However, in time-limited quizzes, students must quickly pick up their knowledge and apply it to the questions. There will not be much extra time to seek help or review course materials. Under such conditions, students may have less confidence in their answers, thus resulting in low estimation. The more significant decrease in their estimation from assignments to quizzes also provides evidence that students were not so confident about their answers. However, the high metacognitive accuracy in the quizzes is inconsistent with prior work (Harrington et al., 2018). We will discuss it in the discussion section later.

4.2 RQ2: What is the metacognitive accuracy for different types of learning objectives?

Table 1 shows us the types of different learning objectives for assignment tasks. We grouped tasks based on their type, and we ended up with three Coding Practice tasks, four Theory & Coding Practice tasks, one Design task, and one Theory task.

Table 1. *The Types of Learning Objectives for Assignment Tasks*

Assignment	Task	Short Description	Type
1	1	CLI and SQL Commands	Coding Practice
	2	Embedding SQL in a client program	Coding Practice
	3	Relational Algebra Queries	Theory
2	1	Relational Algebra to SQL	Theory & Coding Practice
	2	Attribute Decomposition	Coding Practice
	3	Primary and foreign keys	Theory & Coding Practice
3	1	Entity-Relationship Model	Design
	2	Translation from ER Model to Relational Schema	Theory & Coding Practice
	3	Dependencies	Theory & Coding Practice

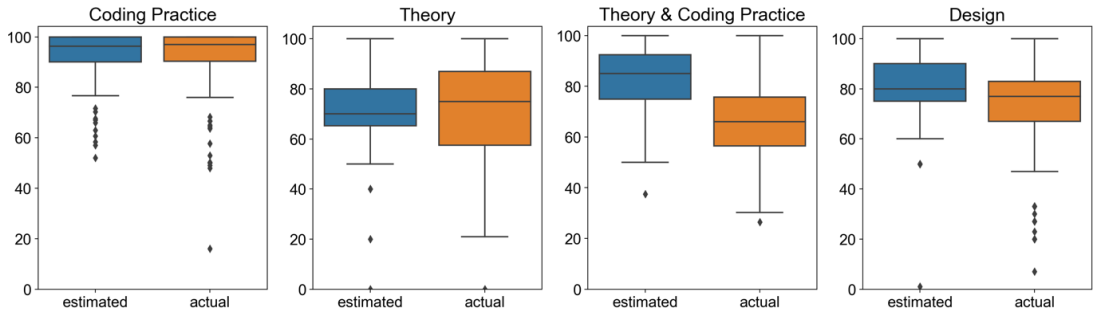


Figure 2. Average Estimated Performance and Actual Performance for different types of Learning Objectives

The coding practice task does not need students to understand complex concept. All students needed to do was coding. The theory task is a math-like question. There is no coding component. Students need to understand the concept to get the answer correct. The theory & coding practice task needs students to write code. However, they can only write correct code if they understand the underlying theory correctly. The design task asks students to create an entity-relationship diagram. There can be multiple correct solutions.

Figure 2 shows us the average estimated performance and actual performance of each student for different types of learning objectives. We can see that in both coding practice and the theory tasks, students made accurate estimations; however, for theory & coding practice tasks and the design task, overestimation can be observed.

Given that effort and practice are related to mindset for programming, programming-efficacy, and performances (Scott & Ghinea, 2014), it is reasonable that students achieved a reasonable metacognitive accuracy for coding practice tasks ($p = 0.68 > 0.05$, Welch's t-test on students' estimation and their actual performance). Students could simply try out their solutions using the provided database server to see whether a solution was correct. Therefore, it is expected that they achieved both high metacognitive accuracy and high performance for those coding practice tasks.

Surprisingly, students did not make significant over- or underestimation in the theory-only task ($p = 0.25 > 0.05$). It may suggest that students have good metacognition skills in estimating theory questions in general. However, we can still see that the lowest estimation was still higher than the lowest actual performance. It suggests that low-performing students tend to overestimate their performance, which is consistent with prior work (Harrington et al., 2018; Lee & Liao, 2021).

Students were generally overestimating for theory & coding practice and design tasks. For the design task ($p < 0.01$), it is reasonable since students rarely had such type of question. The correlation between their estimated performance and their actual performance was also low ($r = 0.03$, Figure not shown in the paper). Thus, we consider they were overestimating it under the Dunning–Kruger effect (Kruger & Dunning, 2000). In fact, most students tend to believe their solutions were (at least half) correct, thus overestimating the task blindly. For the theory & coding practice tasks ($p < 0.01$), we consider it is the great combined difficulty of such tasks that leads students to overestimate. It is found in other fields that people tend to overestimate their actual performance when tasks are difficult (Krawczyk & Wilamowski, 2019; Moore, Healy, & Hall, 2008).

5. Discussion

Our result (Section 4.1) implies that there is a difference in metacognitive accuracy between students' assignment performances and students' quiz performances. The data in our study shows that students could make reasonably accurate estimations of their actual performance in quizzes without receiving markings or feedback for the previous assignment. It has been found that students' estimation can be calibrated on easy items (John L. Nietfeld et al., 2005). As we provided the sample solution of assignments to students, we consider the sample solution played an important role in calibrating students' metacognitive accuracy. However, we do not consider all the quizzes easy. A preliminary experiment (not shown due to page limit) shows that students only show an inaccurate estimation in quiz 3, which mainly contains design questions. However, they had accurate estimations for quiz 1 and quiz 2. It is accurate for quiz 1 because quiz 1 mainly contains coding practice and theory questions. Students tend to make accurate estimations of these types of learning objectives (see Figure 2). However, quiz 2 mainly contains the theory & coding practice questions, which are considered difficult; thus, calibration may not exist. Further analysis is needed to reveal the real reason.

6. Threats to Validity

The first threat to validity would be that students cooperate sometimes. It was not likely that students over-cooperated in quizzes since those quizzes had a very intense time setup. However, it was possible that some students over-cooperated for some assignments in the course. Therefore, some of the answers for assignments may not reflect their real metacognition. However, those cases should be considered trivial. Otherwise, they would be recognized by the teaching assistants who marked their assignments.

Second, which could potentially be the biggest threat, it is possible that students did not submit their thoughtful estimation, but instead, they filled up a random number. Those numbers should be considered outliers. However, since their estimation was worth a few points (although its effect would be lost when calculating the final grades, students do not know that), we consider that most of them would try their best to make accurate estimations.

7. Conclusion

This paper investigated students' metacognitive accuracy in homework assignments, time-limited quizzes, and different learning objectives in a database course. We found that the correlations between students' average estimated performances and their average actual performances are similar; however, when estimating their assignment performances, students have a significant tendency to overestimate their performance, but they tend to make accurate estimations when it comes to time-limited quizzes, although the reason behind remains hidden. We found that students made accurate estimations for the coding practice and the theory tasks; however, overestimation can be observed for the theory & coding practice and the design tasks.

References

Aghababyan, A., Lewkow, N., & Baker, R. S. (2018). Enhancing the Clustering of Student Performance Using the Variation in Confidence. In R. Nkambou, R. Azevedo, & J. Vassileva (Eds.), *Intelligent Tutoring Systems* (pp. 274–279). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-91464-0_27

Hacker, D. J., Bol, L., & Bahbahani, K. (2008). Explaining calibration accuracy in classroom contexts: The effects of incentives, reflection, and explanatory style. *Metacognition and Learning*, *3*(2), 101–121. https://doi.org/10.1007/s11409-008-9021-5

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, *92*(1), 160–170. https://doi.org/10.1037/0022-0663.92.1.160

Harrington, B., Peng, S., Jin, X., & Khan, M. (2018). Gender, confidence, and mark prediction in CS examinations. *Proceedings of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, 230–235. Larnaca Cyprus: ACM. https://doi.org/10.1145/3197091.3197116

John L. Nietfeld, Li Cao, & Jason W. Osborne. (2005). *Metacognitive Monitoring Accuracy and Student Performance in the Postsecondary Classroom*. 7–28.

Krawczyk, M., & Wilamowski, M. (2019). Task difficulty and overconfidence. Evidence from distance running. *Journal of Economic Psychology*, *75*, 102128. https://doi.org/10.1016/j.joep.2018.12.002

Kruger, J., & Dunning, D. (2000). *Unskilled and Unaware of It: How Difficulties in Recognizing One's Own Incompetence Lead to Inflated Self-Assessments*. 14.

Lee, P., & Liao, S. N. (2021). Targeting Metacognition by Incorporating Student-Reported Confidence Estimates on Self-Assessment Quizzes. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 431–437. Virtual Event USA: ACM. https://doi.org/10.1145/3408877.3432377

Moore, D., Healy, P. J., & Hall, A. (2008). *The Trouble with Overconfidence*. 61.

Murphy, L., & Tenenberg, J. (2005). *Do Computer Science Students Know What They Know?: A Calibration Study of Data Structure Knowledge*. 5.

Pahl, C., Barrett, R., & Kenny, C. (2004). *Supporting Active Database Learning and Training through Interactive Multimedia*. *Vol.36*(3), 27–31.

Prather, J., Becker, B. A., Craig, M., Denny, P., Loksa, D., & Margulieux, L. (2020). What Do We Think We Think We Are Doing?: Metacognition and Self-Regulation in Programming. *Proceedings of the 2020 ACM Conference on International Computing Education Research*, 2–13. Virtual Event New Zealand: ACM. https://doi.org/10.1145/3372782.3406263

Reyes, G., Silva, J. R., Jaramillo, K., Rehbein, L., & Sackur, J. (2015). Self-Knowledge Dim-Out: Stress Impairs Metacognitive Accuracy. *PLOS ONE*, *10*(8), e0132320. https://doi.org/10.1371/journal.pone.0132320

Scott, M. J., & Ghinea, G. (2014). On the Domain-Specificity of Mindsets: The Relationship Between Aptitude Beliefs and Programming Practice. *IEEE Transactions on Education*, *57*(3), 169–174. https://doi.org/10.1109/TE.2013.2288700

Stephens-Martinez, K. (2021). A Study of the Relationship Between a CS1 Student's Gender and Performance Versus Gauging Understanding and Study Tactics. *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, 679–685. Virtual Event USA: ACM. https://doi.org/10.1145/3408877.3432365

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Investigation on Practical Effects of the Explanation in a K-12 Math Recommender System

Yiling DAI[a*], Kyosuke TAKAMI[a], Brendan FLANAGAN[a] & Hiroaki OGATA[a]
[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*dai.yiling.4t@kyoto-u.ac.jp

Abstract: Providing explanations in educational recommender systems is expected to improve students' motivation and learning outcomes. In this study, we aim at exploring the practical effects of the explanation in a K-12 math recommender system. We implemented a concept-explicit quiz recommender system and generated explanations of the estimated quiz difficulty and the expected learning gains on the concepts. We then conducted an A/B experiment (N=281) in a Japanese high school to verify the explanation's effects on students' acceptance of the recommendations and related learning behaviors. We found that the explanations led to a higher click rate of recommendations. Besides, among the group displayed with explanations, the students who used the system more frequently and consistently gave more positive feedback on the explanations.

Keywords: Recommender system, math quiz, explanation, recommendation acceptance, A/B experiment

1. Introduction

Recommending learning materials based on individual student's knowledge state has become an essential task in learning management systems. Unlike consuming entertaining products such as movies or music, reading learning materials or solving quizzes requires higher level of motivation and determination of users. Therefore, it is important to provide explanations to the students when recommending learning materials. Some pioneering works have explored the effects of explanations in educational recommender systems and found the explanations may: increase the students' attention towards recommended practices, the willingness to open them (Barria-Pineda et al., 2021); increase the students' trust in hints, perceived usefulness of them, and the intention to use them again (Conati, Barral, Putnam, & Rieger, 2021); increase the students' perceived unexpectedness, novelty of recommended courses, and the interests in them (Yu, Pardos, Chau, & Brusilovsky, 2021); increase the students' clicks of the recommended quizzes (Takami, Dai, Flanagan, & Ogata, 2022). However, it is still under-researched how the explanations affect students' learning behaviors and performance in real classroom settings especially at the level of K-12 education. In this study, we aim at enriching this research topic by exploring the real-world effects of explanations in a K-12 math quiz recommender system.

In the context of learning math, Birenbaum et.al. (1993) suggested that identifying specific misconcepts and difficult areas is more instructive than a test score for remediation. Therefore, it is important to recommend math quizzes that address the students' weak points which are readable math concepts. Previously, we proposed a simple and explainable math quiz recommender system in which estimations are made based on human-readable concepts (Dai, Flanagan, Takami, & Ogata, 2022). We recommended quizzes that are appropriately difficult and help address the student's weak areas, companioned with the corresponding explanations. The explanation is expected to increase the students' belief of their ability to solve the quiz and the perceived values of solving the quiz, which are suggested as strong predictors of learning motivation (Wigfield & Eccles, 2000). On the other hand, the explanations on the math concepts are expected to help the students construct relationships between the quizzes they are going to solve and the math knowledge they already possess, which is consistent with the subsumption theory proposed by Ausubel (1962).

To verify the learning effects of the explanations, we conducted an A/B experiment in a Japanese high school. We focused on the following research questions:

RQ1: Do the explanations motivate the students to accept the recommendations?

RQ2: Do the students who favor the explanations have any characteristics in learning behavior?

The remaining of this paper is structured as: In Section 2, we introduce previous works about explainable recommender systems in educational context. Section 3 overviews the design and implementation of the concept-explicit math quiz recommender system. In Section 4, we report the real classroom experiment and discuss the findings of the explanation's effects. We conclude this study in Section 5.

2. Related Work

In general recommender systems, providing explanations on why the item is recommended is supposed to raise the user's trust (X. Wang et al., 2018), promote the user to accept the recommendation (Huang et al., 2019; X. Wang et al., 2018; Xie, Hu, Cai, Zhang, & Chen, 2021), and support the user in making a good decision (N. Wang, Wang, Jia, & Yin, 2018; X. Wang et al., 2018). In educational recommender systems, many researchers have attempted to generate explanations for a recommended learning material (Barria-Pineda et al., 2021; Conati et al., 2021; Rahdari et al., 2020; Takami et al., 2022; Yu et al., 2021). However, many works did not include an experiment evaluating the effects of the explanations, such as (Rahdari et al., 2020) in the educational context. Some of the works only conducted simulation-based experiments (Bilgic & Mooney, 2005; Conati et al., 2021; N. Wang et al., 2018; X. Wang et al., 2018). For example, Conati et.al. (2021) conducted a control experiment to investigate participants' feedback toward the explanations in a laboratory environment which lasted for two or three hours. We consider that it is insufficient to conclude something about the long-term effect in learning. Yu et.al. (2021) compared the effects of three types of explanations in a course recommender system. We consider that course selection is different from the scenario of doing math quizzes as the latter one requires users' continuous engagement with the system. To the best of our knowledge, only two works evaluated the explanation's practical effect in real settings: Barria-Pineda et.al. (2021) conducted a long-term experiment to test the learning effects of explanations on recommending programming exercises to university students. Different from their experiment setting where both groups were exposed to the explanation, our study verifies the effect of explanation by comparing groups with and without explanations. Takami et.al. (2022) explored the explanation's effect on the usage of recommended math quizzes. Our study focuses on a different recommender system and succeeded to collect more log data and analyze in more depth.

3. Overview of Our Concept-Explicit Quiz Recommender System

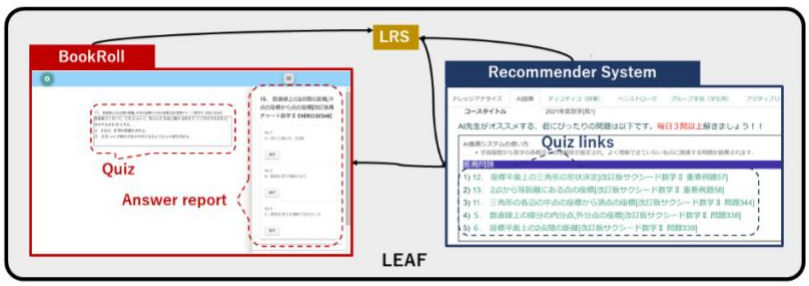


Figure 1. The overview of the system.

In our previous work (Dai et al., 2022), we proposed a concept-explicit quiz recommender system where the estimations are made based on the math concepts. We recommend math quizzes and generate explanations based on two strategies: 1) quizzes of appropriate difficulty. 2) quizzes that improve students' weak points. We implemented the recommender system under the Learning Evidence Analytics Framework (LEAF) (Flanagan & Ogata, 2018) where an ebook reading module BookRoll is used to display the math quizzes and collect the students' answers, and a learning record storage module

LRS is used to record the learning activities in the system. As illustrated in Figure 1, our recommender system guides the students to the BookRoll pape of the quiz based on the students' log data generated in the system.

4. Practical Effects in Real Classrooms

4.1 Experiment Design

To answer the research questions, we conducted a quasi-experiment (Cohen, Manion, & Morrison, 2011) where the experimental group was exposed to the recommender system with explanations and the control group was exposed to the recommender system without explanations. As Figure 2 shows, the treatment of the control group only displays five links of the recommended quizzes. In addition to the links, the treatment of the experimental group contains i) a concept panel that demonstrates the student's current mastery level, ii) the estimated difficulty level of the quiz, and iii) the expected mastery level update. Note that this is one of various possible visualizations and exploring the effects of different visualizations is out of the scope of this study.



Figure 2. The treatment of the control group and experimental group.

A Japanese high school adopts the LEAF system as a supplement of its brick and mortar education. 7 classes of first-year students (N=281) were invited to participate in the experiment and we obtained their consent on providing the learning data for research use. The 7 classes were separated into control (4 classes) and experimental (3 classes) groups randomly. During the experiment period (a short school holiday from February 9, 2022 to February 20, 2022), the students were required to complete at least 20 assigned quizzes and 10 recommended quizzes for reviewing a previously learnt topic. The target topic "geometry and equations" contains 262 descriptive problems from two quiz books. Normally, the students are supposed to complete the quizzes at their own paces and the teachers give supervision or intervention when they think necessary. Thus, it is always the case that an average student cannot finish all the quizzes and there is a need to personalize the quizzes of a doable size. Note that the students have both the paper-version and digital-version of the quizzes at hand, and it is observed they are more comfortable with the traditional paper version.

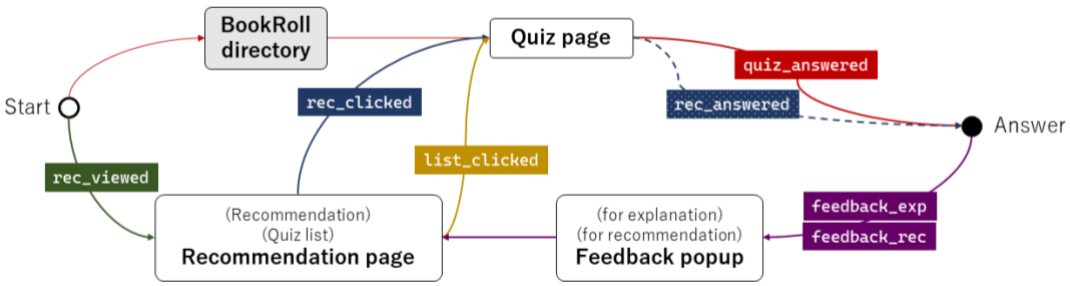


Figure 3. The student workflow and the collected log data.

Figure 3 illustrates the student workflow and the collected log data in our system. To access a quiz page, the student can directly open one topic from the directory and move around the pages in the topic. If the student chooses to answer the recommended quizzes, s/he can open the recommendation

page, where a recommendation block displays the recommendations and associated explanations, and a quiz list block displays all the links to the quizzes in the current topic. Every time a student opens the recommendation page, some recommendations are generated and recorded as *rec_viewed*. The clicks of the links in the recommendation block and quiz list block are recorded as *rec_clicked* and *list_clicked*, respectively. After a quiz page is opened, the student is supposed to solve the quiz either using the digital pen or their own notes, check the answer, and then report whether the answer is correct or not. Only 0 (incorrect) or 1 (correct) is recorded as *quiz_answered*. If the *quiz_answered* action happens after a *rec_clicked* in a predefined time window (one hour in this study), the *quiz_answered* is also counted as an answered recommendation *rec_answered*. After the student answers a quiz from recommendation, s/he is required to rate the explanation (if displayed) and the recommendation with a score from 1 to 5, which is recorded as *feedback_exp* and *feedback_rec*, respectively.

4.2 Data Collection and Preprocessing

We used the following process to preprocess the log data:

1. For each student, we sorted her/his log data in temporal sequence.
2. We treated an action as a start of a new session if it happened more than one hour later than its previous action.
3. We computed the average time spent on one quiz in each session. We then deemed a session as valid if the time spent on one quiz is longer than one minute.
4. We excluded all the actions in invalid sessions.

After the filtering process, we had 30,673 records of 184 unique students and 341 unique quizzes remained in our dataset. We processed the data mainly using Python and jamovi.

4.3 Results and Discussion

4.3.1 Explanation's Effects Over Students Who Used the Recommendations

Table 1. *Welch's T-test of Students' Recommendation Usage in Experimental and Control Groups*

Variable (per student)	Group	N	Mean	SD	Welch's T-test	
					Statistics	df
Recommendation click rate	**Control**	**71**	**0.154**	**0.178**	**2.511****	**90.7**
	Experimental	**58**	**0.264**	**0.290**		
Recommendation answer rate	Control	70	0.817	0.156	1.382	121.8
	Experimental	55	0.853	0.135		
Perception of recommendation	Control	69	3.592	0.896	0.662	123.7
	Experimental	57	3.690	0.779		

Hypothesis: Control<Experimental
* p<.05, ** p<.01, *** p<.001

To answer **RQ1**, we are interested in knowing whether the explanation motivates the students to accept the recommended quizzes and perceive the recommendations positively. To do so, we first computed recommendation click rate as $\frac{\#\text{ rec_clicked}}{\#\text{ rec_viewed}}$, and recommendation answer rate as $\frac{\#\text{ rec_answered}}{\#\text{ rec_clicked}}$, respectively. We also took the average of the *feedback_rec* as a student's overall perception of the recommendations. We then conducted a Welch's T-test on the values between two groups. As we can see from Table 1, there is a significant difference between the experimental (0.264) and control (0.154) groups in terms of the recommendation click rate. This indicates that the student is more likely to click a recommendation if the explanation is provided, which provides a positive answer to **RQ1**. Our result is a reinforcement of the findings in Takami et.al.'s work (2022) as we analyzed the click rate at the level of every individual student, which excludes the biases brought by specific "heavy" users.

The answer rate was higher than 0.8 in both groups. In other words, if the student clicks the quiz, s/he always answers the quiz given this learning context. To this end, raising the students' attention and interest toward a quiz at an earlier stage is more important for increasing the recommendation

acceptance rate. In terms of the perception of the recommendation, we did not find a significance between two groups. The reason could be complicated as the goodness of the recommendation is difficult to define at the first place. Unlike watching a movie, doing a math quiz is not always enjoyable. The students' feedback on the recommendation may be influenced by their values and attitudes towards math learning. In the future work, we need to collect more specified students' feedback on the recommendation.

4.3.2 Explanation's Relationship with Learning Characteristics

Table 2. *Descriptive Statistics and Correlations for Perception of the Explanation and Learning Behavior Characteristics*

		Descriptive statistics			Correlations
	Variable	N	Mean	SD	**1**
Perception of explanation	1. Mean of *feedback_exp*	57	3.72	0.797	**-**
Characteristic of learning behavior	2. *mean_quiz_time* (minute)	184	9.59	6.52	**0.286* (N=57)**
	3. *num_days*	184	2.52	1.66	**0.263* (N=57)**
	4. *max_cons_days*	184	1.41	0.733	**0.431*** (N=57)**

* p<.05, ** p<.01, *** p<.001
Correlations among Variables 2-4 are omitted since they are not our focus.

Among the students who were displayed with the explanations, who favors the explanations from the perspective of learning behavior? Utilizing the session information, we obtained three metrics which indicate different characteristics of the learning behavior: a) average time spent on answering a quiz, b) number of days using the system, and c) maximum number of days using the system consecutively. We took the average of the *feedback_exp* as the student's overall perception of the explanation. Table 2 shows the descriptive statistics and the correlations of the four variables. We found that all of the three metrics are positively related to the perception of the explanation. Especially, the students who used the system more consistently (i.e., larger number of consecutive days) tended to have a better perception of the explanation. To this end, we answer **RQ2** as the students who favored the explanations used the system more consistently. Although the causality remains unclear, we think this finding shows evidence for Bilgic and Mooney's argument (2005)--- explanations could improve user's satisfaction, foster user's trust and thus motivate user's continuous use of the system. In the future, we need to further investigate how the students' personalities and other learning strategies are related to their perceptions of the explanations and their usage of the recommendation.

5. Conclusions and Future Work

In this study, we focused on investigating the practical effects of a K-12 math quiz recommender system where explanations are generated to help students better understand why the quiz is recommended to them. We implemented the recommender system in an integrated learning framework and generated explanations of the estimated difficulty of the quiz and the expected master level update on the student's weak concepts. We then conducted a 12-day A/B experiment (N=281) in a Japanese high school to verify the explanation's effects on students' acceptance of recommendations and related learning behaviors. We found promising results such as: a significant difference between the click rates in groups with and without explanations; students who highly valued the explanations spent much time on each quiz, used the system frequently and consistently.

Some limitations need to be addressed in future works. In this study, we mainly investigated the explanation's effect on the acceptance of the recommendation. We need to further explore the effects on learning performance such as whether the weak points are addressed or not. Besides, we need to

analyze in what sequences the quizzes were recommended to the students and how did their mastery levels evolve during the usage of the recommender system. In addition, to personalize the recommendation mechanism and the corresponding explanations, we need to collect and analyze the data that indicate the students' personalities, emotions, attitudes towards learning, learning strategies, and so on.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, NEDO JPNP20006 and JPNP18013.

References

Ausubel, D. P. (1962). A Subsumption Theory of Meaningful Verbal Learning and Retention. *The Journal of General Psychology*, *66*(2), 213–224.

Barria-Pineda, J., Akhuseyinoglu, K., Želem-Ćelap, S., Brusilovsky, P., Milicevic, A. K., & Ivanovic, M. (2021). Explainable Recommendations in a Personalized Programming Practice System. *Proceedings of International Conference on Artificial Intelligence in Education*, 64–76.

Bilgic, M., & Mooney, R. J. (2005). Explaining Recommendations: Satisfaction vs. Promotion. *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces*.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing Knowledge States in Algebra Using the Rule-Space Model. *Journal for Research in Mathematics Education*, *24*(5), 442–459.

Cohen, L., Manion, L., & Morrison, K. (2011). Experiments, Quasi-Experiments, Single-Case Research and Internet-based Experiments. In *Research Methods in Education (7th ed.)* (pp. 312–334).

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*, 103503.

Dai, Y., Flanagan, B., Takami, K., & Ogata, H. (2022). Design of a User-Interpretable Math Quiz Recommender System for Japanese High School Students. *Companion Proceedings of the 11th International Conference on Learning Analytics and Knowledge.*

Flanagan, B., & Ogata, H. (2018). Learning Analytics Platform in Higher Education in Japan. *Knowledge Management & E-Learning*, *10*(4), 469–484.

Huang, X., Fang, Q., Qian, S., Sang, J., Li, Y., & Xu, C. (2019). Explainable Interaction-Driven User Modeling over Knowledge Graph for Sequential Recommendation. *Proceedings of the 27th ACM International Conference on Multimedia (MM'19)*, 548–556.

Rahdari, B., Brusilovsky, P., Thaker, K., & Barria-Pineda, J. (2020). Using Knowledge Graph for Explainable Recommendation of External Content in Electronic Textbooks. *Proceedings of the Second International Workshop on Intelligent Textbooks 2020 Co-Located with 21st International Conference on Artificial Intelligence in Education (AIED 2020)*, *2674*, 50–61.

Takami, K., Dai, Y., Flanagan, B., & Ogata, H. (2022). Educational Explainable Recommender Usage and Its Effectiveness in High School Summer Vacation Assignment. *Proceedings of the 12th International Learning Analytics and Knowledge Conference (LAK22)*, 458–464.

Wang, N., Wang, H., Jia, Y., & Yin, Y. (2018). Explainable Recommendation via Multi-Task Learning in Opinionated Text Data. *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18)*, 165–174.

Wang, X., Chen, Y., Yang, J., Wu, L., Wu, Z., & Xie, X. (2018). A Reinforcement Learning Framework for Explainable Recommendation. *Proceedings of 2018 IEEE International Conference on Data Mining (ICDM)*, 587–596.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Xie, L., Hu, Z., Cai, X., Zhang, W., & Chen, J. (2021). Explainable recommendation based on knowledge graph and multi-objective optimization. *Complex & Intelligent Systems*, *7*(3), 1241–1252.

Yu, R., Pardos, Z., Chau, H., & Brusilovsky, P. (2021). Orienting Students to Course Recommendations Using Three Types of Explanation. *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, 238–245.

Zhang, Y., & Chen, X. (2020). Explainable Recommendation: A Survey and New Perspectives. *Foundations and Trends® in Information Retrieval*, *14*(1), 1–101.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Transferable Student Performance Modeling for Intelligent Tutoring Systems

Robin SCHMUCKER[a*] & Tom M. MITCHELL[a]
[a]*Carnegie Mellon University, USA*
*rschmuck@cs.cmu.edu

Abstract: Millions of students worldwide are now using intelligent tutoring systems (ITSs). At their core, ITSs rely on student performance models (SPMs) to trace each student's changing ability level over time, in order to provide personalized feedback and instruction. Crucially, SPMs are trained using interaction sequence data of *previous* students to analyze data generated by *future* students. This induces a *cold-start* problem when a new course is introduced, because no students have yet taken the course and hence there is no data to train the SPM. Here, we consider transfer learning techniques to train accurate SPMs for new courses by leveraging log data from existing courses. We study two settings: (i) In the *naive transfer* setting, we first train SPMs on existing course data and then apply these SPMs to new courses without modification. (ii) In the *inductive transfer* setting, we fine tune these SPMs using a small amount of training data from the new course (e.g., collected during a pilot study). We evaluate the proposed techniques using student interaction sequence data from five different mathematics courses taken by over 47,000 students. The naive transfer models that use features provided by human domain experts (e.g., difficulty ratings for questions in the new course) but no student interaction training data for the new course, achieve prediction accuracy on par with standard BKT and PFA models that use training data from thousands of students in the new course. In the inductive setting our transfer approach yields more accurate predictions than conventional SPMs when only limited student interaction training data (<100 students) is available to both.

Keywords: performance modeling, knowledge tracing, transfer learning

1. Introduction

Intelligent tutoring systems (ITSs) are an educational technology that provides millions of students worldwide with access to learning materials and personalized instruction. Even though ITS offerings come at a much lower cost, they can in certain cases be as effective as a personal human tutor (VanLehn, 2011). ITSs can mitigate the academic achievement gap and help disadvantaged students (Huang et al., 2016). At their core, ITSs rely on student performance models (SPMs), to trace each student's changing ability level over time (Corbett & Anderson, 1994), to enable personalized curricula and feedback.

The increasing popularity of ITSs induces a need for SPM techniques that are flexible enough to support frequent releases of new courses, as well as changes to existing courses. The *cold-start* problem, which arises when a new course is released for which no student log data is available for SPM training, prevents us from applying conventional modeling approaches. In practice this means that the first batch of students does not enjoy the full benefits offered by the ITS. Future students then have the advantage that the log data generated by the early students can be used to train an accurate SPM.

In this paper we consider transfer learning (TL) techniques to improve the learning experience of early adopter students by mitigating the SPM cold-start problem for new courses. We show that TL can be used to train accurate SPMs for a new course by leveraging student log data collected from existing courses. We study two settings: (i) In the *naive transfer* setting where no data is available for the new course, we learn *course-agnostic* SPMs – i.e., models whose parameters can be trained using student interaction sequence data from existing courses and that can be applied to any new course. (ii) In the *inductive transfer* setting where small-scale new course data is available, we tune pre-trained

course-agnostic SPMs to the new course by learning new course-specific question and knowledge component (KC) (i.e., skill) difficulty parameters. This inductive transfer setting mimics the case where the course designer can run a pilot with a small number of students before large-scale deployment.

We evaluate the proposed TL techniques using learning trajectory data from over 47,000 students collected from five different mathematics courses offered by a single ITS organization. In both settings, we find that the proposed techniques mitigate the cold-start problem for all courses. We hope that TL methods will become a standard tool for ITS designers and improve the learning experience of early students. The key contributions of this paper include:

- **Course-agnostic student performance models**. We present the first course-agnostic modeling techniques for predicting student performance on future questions in newly introduced courses where no previous students have yet taken this course. Even though our course-agnostic models have no access to training data logs of students taking the new course, they exhibit predictive performance comparable to conventional BKT and PFA models – found in many real-world ITSs – which were trained on data from thousands of students taking the new course.
- **Inductive transfer learning for efficient tuning**. We use transfer learning techniques to efficiently tune our pre-trained course-agnostic performance models to individual new courses by learning question- and KC-specific parameters. Our experiments show how our approach leads to more accurate performance predictions than conventional modeling techniques in settings in which only limited student log data from the new course is available (<100 students).
- **Guidance for practice**. By analyzing data from five different courses offered by a large-scale ITS this work provides insights which can inform the design of future ITSs. Among others, our experiments show how manually assigned difficulty ratings and information about distinct learning contexts provided by human domain experts during content creation can be used to boost the prediction accuracy of course-agnostic SPMs. Further, going against common guidance, our study of various existing SPM approaches reveals that large logistic regression models can outperform classical lower dimensional SPMs even in data starved settings (when training on <10 students).

2. Related Work

2.1 Transfer Learning

Transfer learning (TL) techniques are a class of machine learning (ML) algorithms which aim to improve model performance in a *target* domain (e.g., a new course) by leveraging data from a different but related *source* domain (e.g., existing courses) (Zhuang et al., 2020). TL is particularly attractive when only limited target domain data is available, but source domain data is abundant. Via pre-training on source domain data, TL can acquire a model for the target domain even when *no target domain data* is available. TL techniques enjoy great popularity in domains such as image classification and machine translation but have also been applied to various educational data mining (EDM) problems.

In the context of learning management systems (LMS), TL methods that combine data from multiple different courses or from multiple offerings of the same course have been explored for predicting academic performance (e.g., Tsiakmaki et al., 2020). Data collected from multiple courses has been used to predict the student's likelihood of completing future courses (Huynh et al., 2020) and their degree program (Hunt et al., 2017). In the setting of massive open online courses (MOOCs) TL can improve dropout predictions (e.g., Boyer & Veeramachaneni, 2015). Unlike all above-mentioned transfer approaches, in this work we do not predict a single attribute related to a current course (e.g., pass/fail, student dropout), but rather trace the changing likelihood with which students answer individual questions inside an ITS correctly over time based on their interaction history.

More related to the ITS setting considered in this paper, Paquette et al. (2015) studied the transfer of student gaming detection models between different courses and ITSs. Using simulated students, Spaulding et al. (2021) investigated an approach for transferring cognitive models of language learning between educational games. Multi-task learning has been proposed to learn useful representations via pre-training on response correctness and interaction time prediction tasks (Kim et al., 2021). Baker (2019) framed the problem of transferring student models (e.g., gaming detection models, SPMs, …)

between different learning systems as an open challenge at the EDM2019 conference. Recently, Baker et al. (2021) surveyed related work and discussed directions for future research on sharing models across different learning systems. While our work does not consider the transfer of SPMs across different ITSs, it focuses on the question of transferring SPMs between different courses inside the *same* ITS.

2.2 Student Performance Modeling

Tutoring systems rely on SPMs to estimate a student's ability level based on sequential log data that describes their prior interactions with the system. There are three major categories of SPMs: (i) Markov process-based inference, (ii) logistic regression and (iii) deep learning-based approaches. Markov process-based techniques, such as Bayesian Knowledge Tracing (BKT) (Corbett & Anderson, 1994) and BKT+ (Khajah et al., 2016), are well established and can for example be found in the Cognitive Tutor (Koedinger & Corbett, 2006) and the ASSISTments system (Feng et al., 2009). Most probabilistic approaches estimate a student's ability level by performing inference in a two state Hidden Markov Model – one state to represent mastery and one for non-mastery. Logistic regression models rely on a set of manually specified features which summarizes the student's interaction sequence. Given an input vector with feature values, the regression-based SPM estimates the probability that the student is proficient in a certain question or KC. Some approaches in this class are PFA (Pavlik et al., 2009), DAS3H (Choffin et al., 2019), Best-LR (Gervet et al., 2020) and Best-LR+ (Schmucker et al., 2022). Deep learning-based approaches take as input the same sequence data, but unlike logistic regression techniques can learn suitable features on their own without requiring human feature engineering. Deep learning models benefit from large-scale training data, but as of today, BKT- and logistic regression-based SPMs are still competitive with deep learning in multiple domains (e.g., Khajah et al., 2016; Schmucker et al., 2022). A survey on recent deep learning-based SPMs is provided by Liu et al. (2021).

Importantly, all above-mentioned SPM approaches rely on course-specific parameters (e.g., parameters that represent the difficulty of individual questions and KCs in the target course) that need to be learned from target course data. This makes these models inapplicable in our cold start setting where a new course is first introduced and there is no data for training these parameters available yet.

Lastly, we want to mention recent works (Gervet et al., 2020; Zhang et al., 2021) which investigated another SPM related cold-start problem. There, the question is how accurate are SPM predictions for new students for which we have only observed a few interactions. This is different from the cold-start problem studied in this paper – it addresses the question of how to handle a new cold-start student in an existing course, whereas we address the question of how to handle a new cold-start course. Related to the inductive transfer setting studied in this work, is a short-paper by Zhao et al. (2020) which proposed an Attentive Neural Turing Machine architecture that requires less training data than an LSTM based approach. Unlike our study, they only experiment with small-scale student log data (<30 students, <1000 responses) and do not leverage data collected from existing courses for knowledge transfer.

3. Problem Setting

3.1 The Student Performance Modeling Problem

Formally, we denote the sequence of student s's past interaction with the system as $\boldsymbol{x_{s,1:t}} = (x_{s,1}, \dots, x_{s,t})$. The tuple $x_{s,t} = (y_{s,t}, q_{s,t}, c_{s,t})$ represents the data collected for student s at time-step t. Variable $q_{s,t}$ indicates the answered question, $y_{s,t} \in \{0,1\}$ is binary response correctness and $c_{s,t}$ is an aggregation of additional information about question difficulty, learning context, read materials, watched videos and time. Provided student s's history $\boldsymbol{x_{s,1:t}}$ and a question $q_{s,t+1}$, a SPM f_w estimates $p(y_{s,t+1} = 1 \mid q_{s,t+1}, \boldsymbol{x_{s,1:t}})$ as the probability of s responding correctly to $q_{s,t+1}$ if it were asked next.

All SPMs considered in this paper are parametric and defined by a vector $w \in R^d$. Using training data $D = \{\boldsymbol{x_{s_1,1:t_1}}, \dots, \boldsymbol{x_{s_n,1:t_n}}\}$ capturing interaction logs from *previous* students one can determine a vector w_D for predicting the performance of *future* students by solving the minimization problem

$$w_D = arg\ \min_{w \in R^d} \sum_{s \in D} \sum_{t=1}^{t_s} L\left(f_w\left(q_{s,t}, \boldsymbol{x_{s,1:t-1}}\right), y_{s,t}\right). \quad (1)$$

Here, $L(\hat{y}_{s,t}, y_{s,t}) = -(y_{s,t} log(\hat{y}_{s,t}) + (1 - y_{s,t}) \log(1 - \hat{y}_{s,t}))$ is the negative conditional log-likelihood of observed student response correctness $y_{s,t}$ given model prediction $\hat{y}_{s,t} = f_w(q_{s,t}, \boldsymbol{x_{s,1:t-1}})$ and student history $\boldsymbol{x_{s,1:t-1}}$. This function penalizes predictions $\hat{y}_{s,t}$ that deviate from observation $y_{s,t}$.

3.2 Dataset

For our analysis we rely on the *Squirrel Ai ElemMath2021* dataset (Schmucker et al., 2022) which provides log data from multiple mathematics courses for elementary school students collected over a 3-month period. Overall, the dataset describes about 62,500,000 interactions from over 125,000 students. Going beyond pure question-solving activities, *ElemMath2021* provides insights into how students interact with learning materials. During content creation human domain experts assign each question a difficulty rating between 10 and 90 and specify a prerequisite graph to describe dependencies between individual KCs. *ElemMath2021* further records information about the learning context by assigning each learning activity to one of six categories of study modules (e.g., pre-test, post-test, review, …).

Our study of the transferability of SPMs partitions *ElemMath2021* into multiple course-specific datasets. We selected the five courses with the most students, which we refer to as *C6*, *C7*, *C8*, *C9* and *C40*. Together, these five courses capture approximately 26,300,000 interactions from over 47,000 students (Table 1). Each student only participates in a single course which implies *disjoint* student populations across courses. In terms of covered KCs and used questions the courses are also *disjoint* except for *C9* and *C40* which have an overlap of less than 5%. These properties allow us to measure the transferability of SPMs to different courses involving disjoint students and disjoint questions and KCs.

Table 1. *Five largest ElemMath2021 courses by student number. Avg. responses is the average number of submitted responses per student. Avg. correctness is the proportion of correct student responses.*

course	*C6*	*C7*	*C8*	*C9*	*C40*
# of students	11,864	9,423	10,296	8,531	7,487
# of questions	2,483	2,226	2,438	2,407	1,307
# of KCs	164	145	159	157	87
# of responses	3,262k	1,934k	2,142k	1,407k	1,228k
avg. responses	275	227	187	165	164
avg. correctness	71.30%	69.62%	69.47%	68.68%	62.39%

4. Approach

4.1 Naive Transfer Approach

The naive transfer setting is concerned with using student log data D_S from existing *source* courses $S = \{S_1, \dots, S_k\}$ to learn an SPM that can be applied to any future *target* course T. Crucially, such a *course-agnostic* SPM approach *cannot* rely on any parameters that describe *course-specific* features. Because existing SPMs rely on parameters that capture properties of individual questions and KCs, they require access to target course data D_T for training and are thus not applicable when such data is not available.

As a first step in the design of course-agnostic SPMs we identify a set of features which induces model parameters that do not require target course data for training. For this we study existing logistic regression-based SPMs. Each regression model relies on a distinct feature function $\Phi = (\phi_1, \dots, \phi_d)$ which outputs a real-valued feature vector that describes student s's prior interaction history $\boldsymbol{x_{s,1:t}}$ and information about the next question $q_{s,t+1}$. The trained model then uses this feature vector as input to estimate the probability that s will respond correctly to question $q_{s,t+1}$ if it were asked next as

$$p\left(y_{s,t+1}=1 \mid q_{s,t+1}, \boldsymbol{x_{s,1:t}}\right)=\sigma\left(w^{\intercal}\Phi\left(q_{s,t+1}, \boldsymbol{x_{s,1:t}}\right)\right). \quad (2)$$

Here $w \in R^d$ is the learned weight vector that defines the model and $\sigma(x) = 1/(1+e^{-x}) \in [0,1]$ is the sigmoid function whose output can be interpreted as the probability of correct response.

Because conventional SPMs use feature functions that target course-specific features they do not generalize to new courses. As an example, consider the Best-LR model by Gervet et al. (2020). It features an ability parameter α_s for each individual student and difficulty parameters δ_q and β_k for each individual question q and KC k. Further, Best-LR uses count features for the number of prior correct (c_s) and incorrect (f_s) responses of student s overall and for each individual KC k (i.e., $c_{s,k}$ and $f_{s,k}$). Defining scaling function $\phi(x) = log(1+x)$, the Best-LR prediction is

$$p_{Best\text{-}LR}\left(y_{s,t+1}=1 \mid q_{s,t+1}, \boldsymbol{x_{s,1:t}}\right)=\sigma(\alpha_s-\delta_{q_{s,t+1}}+\tau_c\phi(c_s)+\tau_f\phi(f_s) + \sum_{k\in KC(q_{s,t+1})} \beta_k+\gamma_k\phi(c_{s,k})+\rho_k\phi(f_{s,k})). \quad (3)$$

One can interpret the Best-LR feature function as a tuple $\Phi = (\Phi_A, \Phi_T)$ where Φ_A is course-agnostic (i.e., total counts) and Φ_T is target course-specific (i.e., student ability, question and KC difficulty and counts). Because – to the best of our knowledge – this is the first work that investigates course-agnostic SPMs we introduce simple but reasonable baselines by taking conventional SPM approaches and reducing them to their course-agnostic feature sets.

From Best-LR we derive a course-agnostic SPM called A-Best-LR. A-Best-LR uses overall count features c_s and f_s to indicate the number of student *s*'s prior correct and incorrect responses. The two parameters γ and ρ consider the number of prior correct ($c_{s,k}$) and incorrect responses ($f_{s,k}$) for the *current* KC k – the *same* γ and ρ parameters are used for *all* KCs. Best-LR's ability parameters are reduced to a *single* bias term α that is constant over time for *all* students. The A-Best-LR prediction is

$$p_{A\text{-}Best\text{-}LR}\left(y_{s,t+1}=1 \mid q_{s,t+1}, \boldsymbol{x_{s,1:t}}\right)=\sigma\left(\alpha+\tau_c\phi(c_s)+\tau_f\phi(f_s)+\gamma\phi\left(c_{s,k}\right)+\rho\phi\left(f_{s,k}\right)\right). \quad (4)$$

By avoiding course-specific features A-Best-LR can be trained on source data D_S from existing courses and then be used for any new course T. Giving a similar treatment to other common SPMs we define:

- A-BKT: We train a single BKT parameter set shared for all KCs. We then estimate student performance by using this parameter set to initialize a separate BKT model for each individual KC.
- A-IRT: We train an IRT (Rasch) model that uses the same difficulty parameter (δ) for all questions. We then use this single difficulty parameter to trace each student's ability over time for each KC and derive performance predictions. The student ability parameters are updated after each response.
- A-PFA: We train a reduced 3-parameter PFA model that uses the same difficulty (δ), correctness (γ), and incorrectness count parameters (ρ) for all KCs.
- A-DAS3H: We train a reduced DAS3H model that uses a shared difficulty parameter (δ) for all questions and KCs, a shared constant ability bias term (α) for all students and a single a set of time-window based correctness and incorrectness count parameters for all KCs.
- A-Best-LR+: We train a reduced Best-LR+ model that augments the A-Best-LR feature set (Equation 3) with response pattern and smoothed average correctness features (Schmucker et al., 2022). In addition, the model learns a single set of DAS3H time-window (Choffin et al., 2019), R-PFA (Galyardt & Golding, 2015) and PPE (Walsh et al., 2018) count parameters used for all KCs.

Related to A-BKT, Corbett & Anderson (1994) evaluated a version of BKT which uses a single set of BKT parameters for all KCs that is trained and tested on data from the same course. Related to A-PFA, Maier et al. (2020) proposed to learn PFA parameters for KCs with enough training data and to use the average of the parameters to model KCs with insufficient data in the same course. A-BKT and A-PFA are different in that they train on data from existing courses and then make predictions for a new course.

Conventional SPMs – including all the above – base their estimates solely on log data that describes the student's question-answering behavior. Recently, it has been shown how alternative types of log data collected by modern ITSs can improve logistic regression-based SPMs (Schmucker et al., 2022). The use of such alternative types of features is particularly interesting in the naive transfer setting

because most conventional SPM features are course-specific and are thus not transferable. The *ElemMath2021* dataset captures various types of student interaction data. In our experiments we consider information related to student video and reading material usage, learning context, question difficulty ratings assigned by human domain experts during content creation, KC prerequisite structure and the response- and lag-time features introduced by SAINT+ (Shin et al., 2021).

4.2 Inductive Transfer Approach

Most conventional SPM approaches rely on parameters that capture question- and KC-specific attributes. By training and testing on *target* course data D_T using a 5-fold cross validation, Table 2 compares the performance of course-agnostic SPMs with models that use the same *course-agnostic* feature set, but which are allowed to learn additional *course-specific* parameters to capture question- and KC-difficulty. We observe that the inclusion of question- and KC-specific parameters leads to large improvements in prediction accuracy and closes the gap to conventional SPM techniques (Table 3).

Table 2. *When training and testing on data from the same course, adding course-specific question- and KC-difficulty parameters to the course-agnostic A-AugLR model increases accuracy (ACC) and AUC.*

	ACC (%)	AUC (%)
A-AugLR	72.02	69.48
A-AugLR+KC	74.00	74.99
A-AugLR+quest.	76.34	79.39
A-AugLR+KC+quest.	**76.37**	**79.39**

Motivated by this observation, we propose an inductive TL approach that uses small-scale target course data D_T to tune a pre-trained course-agnostic SPM to a new course T by learning additional question- and KC-specific parameters. Formally, the pre-trained agnostic and target model are defined by weight vectors $w_S \in R^{|\Phi_S|}$ and $w_T \in R^{|\Phi_S|+|\Phi_T|}$ respectively. We use L_2 regularization to subject the target weights w_T to a Gaussian prior $\mathcal{N}((w_S, 0)^\intercal, 1)$ and control the degree of regularization using a penalty parameter $\lambda \in R_{\geq 0}$. The regularized maximum likelihood objective is

$$w_T = arg\ \min_{w \in R^d} \frac{\lambda}{2} ||w - \binom{w_S}{0}||_2^2 + \sum_{s \in D_T} \sum_{t=1}^{t_s} L\big(f_w\big(q_{s,t}, \boldsymbol{x_{s,1:t-1}}\big), y_{s,t}\big). \quad (5)$$

By using a prior for w_T that is based on the previously learned w_S, we can mitigate overfitting and can learn a suitable target model using only very limited training data D_T. With increasing amounts of recorded learning histories in D_T the objective focuses increasingly on model fit. For our experiments we determine the penalty parameter value by evaluating $\lambda \in \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 100\}$ using the first split of a 5-fold cross validation on the *C6* training data. We found $\lambda = 5$ to be most effective for different amounts of tuning data and use it for all our experiments.

5. Experiments

5.1 Evaluation Methodology

As is common in prior work (e.g., Choffin et al., 2019; Gervet et al., 2020) we filter out students with less than ten answered questions. In the naive transfer setting, we use each course once to simulate a new target course $T \in \{C6, C7, C8, C9, C40\}$. For each target course T we train one course-agnostic SPM using source data D_S from the other four courses and then evaluate predictions on the unseen target dataset D_T. For the inductive transfer experiments, we perform a 5-fold cross-validation on the student level where in each fold 80% of students are used as training set $D_{T,train}$ and the remaining 20% are used as test set $D_{T,test}$. To simulate small-scale training data, we sample a limited number of students $(5, 10, \ldots)$ from training set $D_{T,train}$. Because the *ElemMath2021* courses tend to introduce topics in the same sequential order, we only sample students that reached the last topic – sampled

students might have skipped or revisited individual topics. This approach mimics the case where the course designer can collect interaction log data from a small number of students during a pilot study before large-scale deployment. We report model performance using accuracy (ACC) and area under curve (AUC) metrics. AUC is a common evaluation metric for SPMs which can be interpreted as the probability that the model ranks a random correct student response higher than a random incorrect response.

Table 3. *Reference model performance. Performance metrics achieved by conventional course-specific student performance models that were trained and tested on data from the same course.*

	C6		*C7*		*C8*		*C9*		*C40*		Averaged	
model \ in %	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Always correct	71.30	50.00	69.62	50.00	69.47	50.00	68.68	50.00	62.38	50.00	68.29	50.00
BKT	74.89	73.39	71.66	69.35	72.24	70.43	72.01	70.09	68.09	71.00	71.78	70.85
PFA	74.66	73.02	71.52	69.19	72.13	70.21	71.87	69.94	67.85	70.87	71.61	70.65
IRT	75.52	75.66	73.05	73.22	73.28	73.21	72.40	72.36	68.66	72.05	72.58	73.30
DAS3H	77.31	78.15	74.59	76.06	75.05	76.18	74.09	75.38	70.87	75.20	74.38	76.19
Best-LR	78.42	80.30	75.95	78.44	76.58	78.97	76.33	79.08	73.10	78.07	76.08	78.97
Best-LR+	**78.75**	**80.85**	**76.18**	**78.83**	**76.90**	**79.39**	**76.69**	**79.58**	**73.62**	**78.81**	**76.43**	**79.49**

Our code builds on the public GitHub repository by Schmucker et al. (2022) which implements various SPMs. We build on their regression models and leave their hyperparameter choices unchanged. We use pyBKT (Badrinath et al., 2021) to implement the BKT experiments. For our naive and inductive transfer experiments we use PyTorch and train each model for 200 epochs using the Adam optimizer with learning-rate $\alpha = 0.001$. As a reference, Table 3 shows average performance metrics of common SPM approaches that were trained and tested on the *same* course using a 5-fold cross-validation. To increase reproducibility, we provide detailed descriptions of the evaluated features and SPMs in an external appendix hosted on GitHub (https://github.com/rschmucker/TransferableSPM-Appendix).

5.2 Naive Transfer Experiments

Feature Evaluation. We evaluate the benefits of different features for course-agnostic SPMs. For each feature, we train an augmented A-Best-LR+ model using the A-Best-LR+ feature set plus one of several possible additional features, described below. We use A-Best-LR+ because it combines features that were found most useful in earlier SPMs and it yields the most accurate predictions among all considered course-agnostic baseline models in our experiments (Table 5).

Table 4 shows the ACC and AUC scores when adding each of several additional features to A-Best-LR+. The most useful additions are the one-hot features that encode question difficulty ratings assigned by human domain experts during content creation – these improve performance on average over all five courses by 0.24% ACC and 1.07% AUC. The one-hot learning context features improve the average AUC score by 0.14%. The count features that track the number of prior correct and incorrect responses to questions of a certain difficulty or learning context, lead to smaller improvements compared to their one-hot counterparts. The lag time and response time features improve AUC scores on average by 0.15% and 0.11%. The post- and pre-requisite features derived from the KC dependency graph did not benefit the course-agnostic SPMs. Similarly, the count features that summarize the students' video and reading material usage did not improve the performance predictions. One limitation of these two count features is that they do not capture the relationship between the content covered by individual learning materials and questions.

Table 4. *Naive transfer feature evaluation. Using the A-BestLR+ feature set augmented with one additional feature we trained course-agnostic models on four source courses and evaluated on a new target course. The marker **X** indicates which additional features yielded the largest improvements.*

	C6		*C7*		*C8*		*C9*		*C40*		Averaged		
model \ in %	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	
A-BestLR+ (base)	73.86	67.33	71.62	65.41	71.92	65.79	72.36	68.92	67.59	68.31	71.47	67.15	
current lag time	73.91	67.48	71.55	65.42	72.00	66.00	72.40	69.02	67.65	68.38	71.50	67.26	**X**
prior resp. time	73.94	67.61	71.57	65.39	72.02	65.98	72.39	69.12	67.46	68.41	71.48	67.30	**X**
context one-hot	73.83	67.24	71.65	65.56	71.94	65.95	72.38	69.02	67.65	68.70	71.49	67.29	**X**

context count	73.87	67.38	71.53	65.33	72.04	65.79	72.41	69.09	67.71	68.54	71.51	67.23	**X**
difficulty one-hot	74.09	68.63	71.88	66.80	72.22	67.21	72.54	69.71	67.82	68.84	71.71	68.24	**X**
difficulty count	73.84	67.34	71.60	65.56	71.93	66.00	72.33	69.00	67.59	68.52	71.46	67.28	**X**
prereq count	73.88	67.27	71.58	65.44	71.91	65.83	72.31	68.92	67.55	68.28	71.45	67.15	
postreq count	73.61	66.48	71.68	65.03	71.98	65.96	72.38	69.22	67.39	68.15	71.41	66.97	
videos count	73.84	67.32	71.59	65.41	71.95	65.75	72.30	68.91	67.52	68.20	71.44	67.12	
readings count	73.83	67.41	71.60	65.48	71.96	65.82	72.37	68.93	67.53	68.19	71.46	67.17	

Agnostic AugmentedLR. Given the experimental results in Table 4, we created a final model (which we call A-AugLR) that incorporates all the features from A-Best-LR+ (base), plus all features marked by X in Table 4, namely lag and response time, learning context and question difficulty. To complete our experiments on the naive transfer setting, we then compared the performance of A-AugLR to the naive transfer baselines defined in Subsection 4.1. The results of this comparison are shown in Table 5. We observe that the course-agnostic SPMs derived from BKT, PFA, IRT and DAS3H struggle in the naive transfer setting and yield low AUC scores. The A-Best-LR+ model uses additional features to capture aspects of long- and short-term student performance over time. On average, its predictions yield 0.37% higher ACC and 1.09% higher AUC scores compared to the A-Best-LR model it builds on.

The A-AugLR model yields the best predictions in the naive transfer setting. Compared to A-Best-LR+, the A-AugLR models are on average 0.38% more accurate and their AUC scores are 1.76% higher. This shows how additional information provided by domain experts during content creation can enable accurate performance predictions on cold start courses in the naive transfer setting. Importantly, even though our course-agnostic A-AugLR models were fitted using data from different source courses their prediction accuracy is on par with course-specific BKT and PFA models which were trained on target course data (compare the A-AugLR row in Table 5 to the BKT and PFA rows from Table 3).

Table 5. *Naive transfer performance. We used each of the five courses to simulate a new target course and trained course-agnostic performance models using student data from the other four source courses.*

	C6		*C7*		*C8*		*C9*		*C40*		Averaged	
model \ in %	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
A-BKT	73.25	63.29	70.58	60.87	71.04	60.85	70.93	62.30	65.56	63.87	70.27	62.57
A-PFA	73.27	63.54	70.75	60.59	71.13	61.23	70.92	62.29	65.55	63.13	70.32	62.16
A-IRT	59.50	61.82	58.36	59.55	57.50	59.45	58.04	60.96	58.33	62.72	58.35	60.90
A-DAS3H	73.29	63.70	70.81	60.84	71.15	61.31	70.98	62.41	65.59	63.54	70.36	62.36
A-Best-LR	73.55	66.38	71.35	64.17	71.74	65.01	71.99	67.87	67.13	66.86	71.15	66.06
A-Best-LR+	73.86	67.33	71.62	65.41	71.92	65.79	72.36	68.92	67.59	68.31	71.47	67.15
A-AugLR	**74.28**	**69.11**	**71.80**	**67.21**	**72.35**	**68.19**	**72.76**	**70.52**	**68.05**	**69.52**	**71.85**	**68.91**

5.3 Inductive Transfer Experiments

Here, we evaluate our inductive TL approach (I-AugLR) that uses small-scale target course data D_T to tune a course-agnostic A-AugLR model – pre-trained on log data from the other courses – to the target course by learning course-specific difficulty parameters. We also evaluate a course-specific model (S-AugLR) which use the same feature set as I-AugLR but does not use a pre-trained model. We measure the amount of target course data for tuning in number of students and experiment with values in $\{0, 5, 10, 25, 50, 100, 250, 500, 1000\}$. The 0-student case is equivalent to the naive transfer setting.

Figure 1 compares the performance of our inductive TL method (I-AugLR) with conventional SPM approaches and S-AugLR trained using only target course data D_T. Due to the page limit, we only plot model performance for *C40*. By tuning a pre-trained A-AugLR model, I-AugLR can mitigate the cold-start problem for all five courses and benefits from small-scale log data. Given as little as data from 10 students, the I-AugLR models consistently outperform standard BKT and PFA models that were trained on logs from thousands of target course students (Table 3). Among all considered SPMs, I-AugLR yields the most accurate performance prediction up to 25 students for C7, up to 100 students for *C6* and *C8* and up to 250 students for *C9* and *C40*. Among the non-TL approaches, Best-LR is most data efficient and yields the best performance predictions when training on up to 500 students.

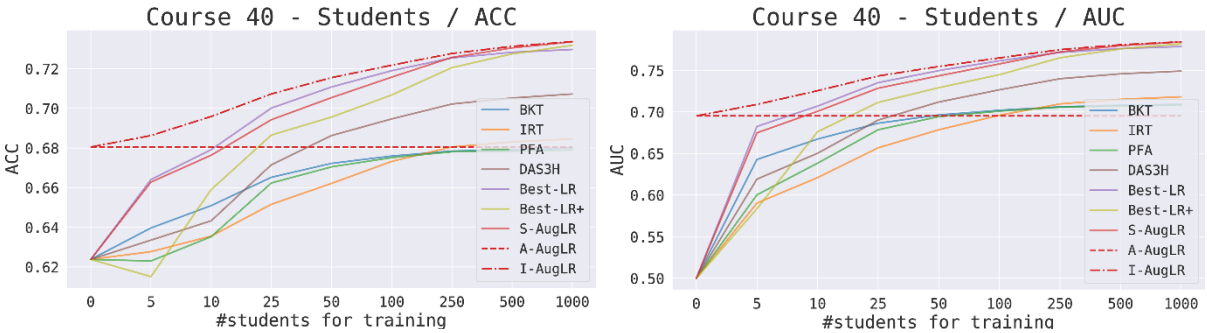


Figure 1. Training student performance models with different amounts of student log data.

6. Discussion

Our experiments show that our proposed TL techniques can mitigate the SPM cold-start problem for new courses by leveraging student interaction data from other existing courses. In the naive transfer setting where no target course data is available, the course-agnostic A-AugLR models trained on data from existing source courses yielded prediction accuracy on par with standard BKT and PFA models that use training data from thousands of students in the target course. One key ingredient of our course-agnostic SPMs is additional information about question difficulty and learning context provided by domain experts during content creation. While these features improve SPM predictions, the need for manual annotations puts an additional load on the content creators. Further, the success of our transfer approach depends to a degree on the domain expert's ability to assign accurate question difficulty labels.

In the inductive transfer setting we use small-scale target course data (e.g., collected during a pilot study) to tune pre-trained course-agnostic SPMs. This allows us to overcome the limitations of the naive transfer setting by learning target course specific question- and KC- difficulty parameters. Our parameter regularized approach yields better predictions than conventional SPM approaches when only limited target course data (<100 students) is available. Surprisingly, we found that among the non-TL approaches, Best-LR (Gervet et al., 2020) yielded the most accurate predictions when training on less than 500 students for all five courses. This is interesting because low-parameter models such as BKT (Corbett & Anderson, 1994), PFA (Pavlik et al., 2009) and IRT (Rasch, 1960) are commonly believed to be more data efficient than more complex logistic regression models that contain many more parameters. What sets Best-LR apart from these three models, is that its parameters describe student performance using multiple levels of abstraction (question-, KC- and overall-level). Future work might investigate this phenomenon further using log data from multiple different ITSs.

One limitation of our study is that it focuses on a set of five courses offered by the same ITS. This has advantages because the course log data is of consistent format and content creators follow similar protocols. Still, it prevents us from answering the question of whether SPMs are transferable between different tutoring systems (Baker, 2019). Another related limitation is that all considered courses cover mathematics topics for elementary school students. Our study did not investigate the transferability of SPMs across different subjects or grade levels (e.g., middle school, high school, …).

7. Conclusion

The increasing popularity of intelligent tutoring systems (ITSs) induces a need for student performance modeling (SPM) techniques that are flexible enough to support frequent new course releases as well as changes to existing courses. This paper proposes two transfer learning approaches for mitigating the cold-start problem that arises when a new course is introduced for which no training data is available. In the naive transfer setting where no new course data is available, we rely on student interaction sequences from existing courses to learn course-agnostic SPMs that can be applied to any future course. In the inductive transfer setting where small-scale new course data is available (e.g., collected during a pilot study), we show how one can tune pre-trained course-agnostic models to a specific course by learning question- and KC- (i.e., skill) difficulty parameters. Our experimental evaluation on student

log data from five different mathematics courses showed how both transfer approaches mitigate the cold-start problem successfully. This work represents a first step in the design of SPMs that are transferable between different ITS courses. The success of our approach depends significantly on (i) automatically learned course-independent parameters that characterize how quickly students learn a skill as a function of the number of prior practice attempts, and (ii) information provided by human domain experts in the form of difficulty values for questions in the new, target course. We hope that transfer learning techniques such as the ones discussed in this paper will enable ITS designers to provide effective adaptive instruction for early adopter students.

Acknowledgements

We thank Richard Tong and Dan Bindman from Squirrel Ai Learning for providing the *ElemMath2021* dataset. This work was supported in part through the CMU - Squirrel Ai Research Lab on Personalized Education at Scale, and in part by the AFOSR under award FA95501710218.

References

Badrinath, A., Wang, F. & Pardos, Z. (2021). pyBKT: An Accessible Python Library of Bayesian Knowledge Tracing Models. In *Proc. of the 14th Int. Conference on EDM* (pp. 468-474). Paris, France: EDM.

Baker, R. S. (2019). Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *Journal of Educational Data Mining*, *11*(1), 1-17.

Baker, R. S., McLaren, B. M., Hutt, S., Richey, J. E., Rowe, E., Almeda, M., Mogessie, M. & Andres, J. M. (2021). Towards sharing student models across learning systems. In *Proc. of the 22nd Int. Conference on AIED* (pp. 60-65). Utrecht, Netherlands: Springer

Boyer, S., & Veeramachaneni, K. (2015). Transfer Learning for Predictive Models in Massive Open Online Courses. In *Proc. of the 17th Int. Conference on AIED* (pp. 54-63). Madrid, Spain: Springer.

Choffin, B., Popineau, F., Bourda, Y., & Vie, J. J. (2019). DAS3H: Modeling Student Learning and Forgetting for Optimally Scheduling Distributed Practice of Skills. In *Proc. of the 12th Int Conference on EDM* (29-38). Montreal, Canada: EDM

Corbett, A. T., & Anderson, J. R. (1994). Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge. *User modeling and user-adapted interaction*, 4(4), 253-278.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User modeling and user-adapted interaction*, *19*(3), 243-266.

Galyardt, A., & Goldin, I. (2015). Move Your Lamp Post: Recent Data Reflects Learner Knowledge Better than Older Data. *Journal of Educational Data Mining*, *7*(2), 83-108.

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is Deep Learning the Best Approach to Knowledge Tracing?. *Journal of Educational Data Mining*, *12*(3), 31-54.

Huang, X., Craig, S. D., Xie, J., Graesser, A., & Hu, X. (2016). Intelligent tutoring systems work as a math gap reducer in 6th grade after-school program. *Learning and Individual Differences*, 47, 258-265.

Hunt, X. J., Kabul, I. K., & Silva, J. (2017). Transfer Learning for Education Data. In *KDD Workshop on Advancing Education with Data*. Nova Scotia, Canada: ACM

Huynh-Ly, T. N., Le, H. T., & Nguyen, T. N. (2020). Integrating courses' relationship into predicting student performance. *Int. Journal of Advanced Trends in Computer Science and Engineering*, 9(4). 6375-6383.

Khajah, M., Lindsey, R. V., & Mozer, M. C. (2016). How Deep is Knowledge Tracing?. In *Proc. of the 9th Int. Conference on EDM* (pp. 94-101). Raleigh, NC, USA: EDM

Kim, B., Yu, H., Shin, D., & Choi, Y. (2021). Knowledge Transfer by Discriminative Pre-training for Academic Performance Prediction. In *Proc. of the 14th Int. Conference on EDM* (pp. 287-294). Virtual: EDM

Koedinger, K. R., & Corbett, A. (2006). Cognitive Tutors: Technology Bringing Learning Sciences to the Classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of: The learning sciences* (pp. 61–77). Cambridge University Press.

Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). A Survey of Knowledge Tracing. *arXiv preprint arXiv:2105.15106*.

Maier, C., Baker, R. S. & Stalzer, S. (2021). Challenges to Applying Performance Factor Analysis to Existing Learning Systems. In *Proc. of the 29th ICCE* (pp. 57-62). Virtual: APSCE

Paquette, L., Baker, R. S., Carvalho, A. D., & Ocumpaugh, J. (2015). Cross-system Transfer of Machine Learned and Knowledge Engineered Models of Gaming the System. In *Proc. of the 23rd Int. Conference on UMAP* (pp. 183-194). Dublin, Ireland: Springer.

Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance Factors Analysis – A New Alternative to Knowledge Tracing. In *Proc. of the 14th Int. Conference on AIED* (pp. 531-538). Brighton, UK: IOS Press

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Schmucker, R., Wang, J., Hu, S., & Mitchell, T. M. (2022). Predicting the Performance of Online Students - New Data, New Approaches, Improved Accuracy. *Journal of Educational Data Mining*, *14*(1), 1–45.

Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., & Choi, Y. (2021). SAINT+: Integrating temporal features for EdNet correctness prediction. In *Proc. of the 11th Conference on LAK* (pp. 490-496). New York, NY, USA: ACM

Spaulding, S., Shen, J., Park, H., & Breazeal, C. (2021). Towards Transferrable Personalized Student Models in Educational Games. In *Proc. of the 20th Int. Conference on AAMAS* (pp. 1245-1253). Virtual: ACM

Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., & Ragos, O. (2020). Transfer Learning from Deep Neural Networks for Predicting Student Performance. *Applied Sciences*, *10*(6), 2145.

VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational psychologist*, *46*(4), 197-221.

Walsh, M. M., Gluck, K. A., Gunzelmann, G., Jastrzembski, T., Krusmark, M., Myung, J. I., Pitt, M. A., & Zhou, R. (2018). Mechanisms Underlying the Spacing Effect in Learning: A Comparison of Three Computational Models. *Journal of Experimental Psychology: General, 147*(9), 1325–1348.

Zhang, J., Das, R., Baker, R. S., & Scruggs, R. (2021). Knowledge Tracing Models' Predictive Performance when a Student Starts a Skill. In *Proc. of the 14th Int. Conference on EDM* (pp. 625-629). Virtual: EDM

Zhao, J., Bhatt, S., Thille, C., Gattani, N., & Zimmaro, D. (2020). Cold Start Knowledge Tracing with Attentive Neural Turing Machine. In *Proc. of the 7th Conference on L@S* (pp. 333-336). New York, NY, USA: ACM

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. & He, Q. (2020). A Comprehensive Survey on Transfer Learning. In *Proc. of the IEEE*, *109*(1), 43-76.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Automated Matching of Exercises with Knowledge components

Zejie TIAN[a*], Brendan Flanagan[b] , Yiling DAI[b] & Hiroaki Ogata[b]
[a]Graduate School of Informatics, Kyoto University, Japan
[b]Academic Center for Computing and Media studies, Kyoto University, Japan
*tianzejie123@gmail.com

Abstract: As intelligent tutoring systems (ITS) fast develop, they have been implemented in many classrooms combing with various tutorial tools. Especially, matching exercises with knowledge components is the fundamental task of many applications, such as automatic recommendation, student knowledge tracing etc. However, manually labelling educational data is labor intensive and time consuming. Therefore, a range of machine learning methods has been proposed to address this problem, while few of them focusing on Japanese educational dataset in the real high school. In this paper, we leverage natural language processing techniques with Keyphrase extraction methods based on Japanese math exercises. We evaluate the model performance with several state-of-the-art methods and how it works in a real educational task. The results show that our methods outperform several state-of-art methods and can effectively save time for managing math exercise.

Keywords: Keyphrase extraction, nature language processing, knowledge components

1. Introduction

Towards the last decade, with the fast development of AI and educational techniques, Intelligent Tutoring System (ITS) and Learning Management Systems (LMS) are widely used for helping students to promote their learning performance. They focused on finding an "optimal path" towards a set of skills, also known as Knowledge components (KCs). Therefore, properly matching exercises with knowledge components is important task for exercise-based applications such as personalized recommendation (Hou, Y., Zhou, P., Xu, J., et al, 2018) (Hou, Y., Zhou, P., Wang, T., et al, 2016), knowledge tracing (Piech, C., Bassen, J., Huang, J., et al, 2015) (Vie, J. J., & Kashima, H. ,2019) etc. Currently, in many schools, exercises sets are still manually labeled by experts. However, manual labelling costs quantity of time and requires expert labors. As the number of exercises fast grows, it requires automatically understand the meaning of exercises and labelling it with suitable knowledge components. Therefore, many research have been proposed to solve this problem. Usually the solution of this problem can be framed as the multinomial classification (Shen, J. T., Yamashita, M., Prihar, E., et al, 2021) which assigns the most relevant KCs label to the exercises. Prior research solutions included supervised learning methods (Pardos, Z. A., & Dadu, A., 2017) (Hage, H., & Aimeru, E., 2006), unsupervised learning methods (Desmarais, M. C., 2012) and deep learning methods (Shen, J. T., Yamashita, M., Prihar, E., et al, 2021) (Huang, T., & Li, X., 2021). However, existing methods most focus on English exercises and seldom of them based on Japanese exercises. Therefore, there are few of linguistic based analyzing rules can be referenced. Besides, most of them require a large number of datasets. Usually in the real school teaching environment, it is hard to collect as much data as Online learning platform and few of them focusing on reducing expert checking time by developing helpful tools. But it needs to blend the bridge to help expertise saves time for managing ITS.

Toward these challenges, in this paper, we propose a new model combing the word embedding based method with the Keyphrase extraction method. Both models may have some strength, word embedding based methods has great ability for representing and understanding the semantic meaning of text. Keyphrase extraction-based methods can utilize the domain knowledge to category the data

which has great interpretability. Especially, for the Japanese, the fine-grained information can be extracted according to the structure of grammar. Combing both of them can be complementary to each method. Word embedding based method can help the model to understand the semantic meaning of exercises from words and with the help of Keyphrase extraction method, it can focus on the "key concepts" of each exercise and improve the model interpretability. With the experience of word embedding based method, it can also improve the model performance of Keyphrase extraction method. We conduct the experiment to test how well it can assist experts for classifying the exercises and comparing the model performance with several state-of-art models. The experiments show that our methods can effectively save the expert oversights time and outperform several state-of-models. By showing the effectiveness of the system, my research paved the way towards a complete automatic knowledge component annotation system.

2. Related work

2.1 Matching exercises

There are wide range of researches of been posted in the last decades. Some researches organized exercises by finding similarity between them which also called as finding similar exercises. Generally, the similar exercises serve the same educational purpose and indicate the same KCs (Del Solato, T., & Du Boulay, B., 1995). Some prior works leveraged texts and concepts in the exercises to compare the similarity. VSM (Tsinakos, A., & Kazanidis, I., 2012) is one of the models by representing the exercises as TF-IDF and calculated the similarity by text distance methods. But they were lack of understanding of semantic meaning of exercises and accuracy was low. Recently, some deep learning-based methods have been proposed to solve this problem. (Liu, Q., Huang, Z., Huang, Z., et al, 2018) extended the resources by utilizing texts, images and KCs for systematically exploit the semantic information of data. (Huang, T., & Li, X., 2021) utilized BERT methods to release the risk of lacking of labeled exercises.

Some researches proposed to match exercises with the KCs of textbook. (Matayoshi, J., & Lechuga, C., 2020) used natural language processing and machine learning methods to match the ITS exercises with certain textbook content and the textbook content belongs to specific topics. By introducing the text book data and explanation of each exercise can effectively help model understand and classify the ITS exercises. In this research, we will utilize the high-performance ability of deep learning model and extension the resource according to the Keyphrase.

2.2 Keyphrase extraction

Keyphrases are the set of words which can effectively highlight the meaning of documents. Extraction of Keyphrases can help us quickly analyzing and organizing the huge amount of data, which can provide concepts or the theme of documents. (Chau, H., Labutov, I., Thaker, K., et al, 2021) proposed to use the machine learning methods by using the POS tagger to annotate the Keyphrases in the educational textbooks. (Alzaidy, R., Caragea, C., & Giles, C. L., 2019) used the deep learning approach to show the outperformance of the Bi-LSTM-CRF model. There are also some related researches (Siddiqi, S., & Sharan, A., 2015) used linguistic based approaches or specific domain knowledge to extract the related Keyphrases in the documents. By knowing or extracting the relationship of Keyphrases can also help us to build up the knowledge graph (Wong, W., Liu, W., & Bennamoun, M., 2012). Construction of educational knowledge graph can also effectively help some downstream task such as automatic recommendation and knowledge tracing etc. (Nakagawa, H., Iwasawa, Y., & Matsuo, Y., 2019) (Tong, S., Liu, Q., Huang, W., et al, 2020) To the best of our knowledge, those methods are still not fully exploit in the educational area and also there are few rule-based methods to reference for the Japanese math data.

3. The proposed model

In this paper, we propose an ensemble method combing both word embedding based method and Keyphrase extraction. We utilize Word2Vec (Church, K. W., 2017) to transform the exercises into the

vector space for understanding the semantic meaning of exercises. Moreover, the Keyphrase extraction can help to indicate the key knowledge components of each exercises. The overall structure of our model showed in figure 1.

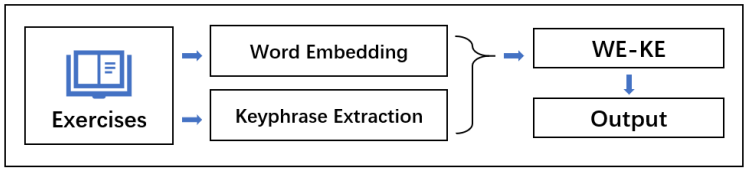


Figure 1. The Overview of Our Proposed Model.

The whole process can be seen as input the exercise text and output the most one of relevant KCs. We use the R_{WE} to represent the result of the word embedding based method and R_{KE} to represent the result of the Keyphrase extraction method. For each single model, the output is the vector of relevant probability for each exercise. We are combining both model results as the input feature for the final model. We will illustrate details in the following sections.

3.1 Word embedding based method

In our system (Flanagan, B., & Ogata, H., 2018), the exercises are in the digital format stored in the separate PDF file. Comparing with HTML file (Matayoshi, J., & Lechuga, C., 2020), it is hard to completely parse and represent the function information within math exercises. Therefore, we only utilize the text information extracting by the Pdf2text (pdftotext, 2021) Python library. Next, we preprocess the text data by utilizing the Nagisa (nagisa, 2020) Python library to segment text into words. During the process, we also filtered the noisy and stop words such as number, function etc. Then, we transformed the words into vector space which length is 300 by utilizing the FastText (Joulin, A., Grave, E., Bojanowski, P., et al, 2016) method. For representing the whole exercises sentences, we calculate the average of word vectors. It can be a simple and effectively way to represent the exercises, because the math questions usually contain few synonyms and have clearly topic. Moreover, we treated the task as the multi-classification problems and trained by machine learning methods: Input the vector representation of exercises and output the probability for each KCs' labels.

3.2 Keyphrase extraction

According to our observation, math concepts can be obviously found in the most of Japanese junior high school math exercises text. For example, Japanese words "因数分解(factorization)" in the exercises showed in figure 2 reveals that it is a question about factorization. Therefore, we collected those key math concepts from text books and expertise suggestions. Then we detecting each exercise about whether it has specific math concepts and save it in the Keyphrase list. As for the example question, the Keyphrase list should be ['2元2次式(binary quadratic)', '因数分解'(factorization)].

37. 2元2次式の因数分解 Factorization of binary quadratic form

次の式を因数分解せよ。Factorization of the following formulas

(1) $x^2-xy-2y^2-x-7y-6$ (2) $3x^2+7xy+2y^2-5x-5y+2$

Figure 2. The Math Exercises about Factorization in the Junior High School Math.

Inspired by the work (Wu, W., Li, H., Wang, H., & Zhu, K. Q,2012), for better extracted the relationship between Keyphrase and flourish the description of exercises, we extracted the relationship of KCs by Japanese linguistic rules. Japanese words 'の' often plays the role of possessor and modifier, usually it can deliver the detail information of KCs such as "整数の加法(Addition of integer)" which shows that the exercise is about addition of integer. 'を' usually uses to mark the object of word, which

often shows the purpose of exercises such as “式を計算” (function calculation). ‘と’ is like ‘and’ in the English used to connect similar word together. Conclusion of linguistic pattern and example shows in Table 1.

Table 1. The Example and Pattern of Linguistic Rules

Pattern	Example
_の_	多項式の次数 (Degree of polynomial)
_と_の_	乗法と除法の混じった計算 (Calculation with a mixture of product and division)
_の_と_	多項式の加法と減法 (Adding and subtracting polynomials)
_を_	同類項をまとめ (Summarize similar terms)

In addition, for better construction of relationship, we filtered the results by setting up the stop words and assigned nearest concepts to the pronoun. For distinguishing the key concepts and modifier, we assign the key concepts according to the key math concepts list extracting from the textbooks. As for the example of figure 2, we store the results as a tuple list: [(2元2次式(binary quadratic), 因数分解(factorization)), (式(formula), 因数分解(factorization))].It should be noticed that the order in the tuple didn’t influence the result.

We assume that *if exercises belong to the same KCs then it must have similar key concepts and key concepts relationship.* According to this assumption, we calculate the similarity score between exercise as follows:

$$sim(E^1, E^2) = \frac{E^1_{keywords} \cap E^2_{keywords}}{E^1_{keywords} \cup E^2_{keywords}} + \frac{E^1_{KR} \cap E^2_{KR}}{E^1_{KR} \cup E^2_{KR}}$$

Keywords is the key concepts set and KR is the key concepts relationship set. We calculate the average of similarity scores between exercises in the test set and a batch of exercises belong to certain knowledge components in the training se and store the final results into a matrix.

3.3 Ensemble model

The ensemble model was showed in the figure 1. Results from word embedding based method can help model better understand the semantic meaning of exercises from text and results from the Keyphrase extraction can help better focus on the key topic of exercises. The similar exercises should contain similar semantics meaning and key concepts. For better unite the two models results, we simply average weighted the both model results (Zhou, Z. H., 2012):

$$O = \mathrm{ML}((R_{WE} + R_{KE})/2, \lambda)$$

O is final output of model predictions which can be represented as $O(o_1, o_2, \dots, o_n)$ where n equals to the number of KCs. For the model input, R_{WE} is the result of word embedding based method, R_{KE} is the result of Keyphrase extraction method and λ is the hyperparameter for normalization to avoid overfitting. ML is the machine learning model in our study is based on XGBoost model. We show the detail analysis in the experiment session.

4. Experiment

4.1 Datasets and experiment settings

To better evaluate our model, we conduct our experiment on the Japanese math exercises collected in the real Junior high school. The datasets contained 830 exercises included four main topics: geometry, function, statistic and probability. Currently, our task is to classify the exercises into 13 detailed KCs. Follow the prior works, we measure the model performance with the macro F1-score and Accuracy

which calculated as (TP+TN)/ (TP + TN + FP + FN). We utilize the XGBoost model for the word embedding based methods and final state. To measure the generalizability of our proposed model, we apply 5-fold cross-validation at the KCs level, which means we split exercises of each KCs into 5 distinct groups and both training and testing datasets are selected from those.

We compared our model with the following prior works. They all use the same input data.

- Vector space model (VSM) (Tsinakos, A., & Kazanidis, I., 2012): The method transformed exercises into vector space and compared the cosine similarity between them. We chose it as the baseline model for comparison.
- Support vector machine (SVM) (Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. A., 2012): SVM model has a great ability to measure high dimensional input features and suitable for the classification of exercises with dense concepts.
- XGBoost (Chunamari, A., Yashas, M., Basu, A., et al, 2022): XGboost is the scalable machine learning models which is built up on the tree boosting. It shows great performance in many kinds of classification problems.
- Neural network (NN) (Patikorn, T., Deisadze, D., Grande, L., et al, 2019): Neural network is one the classic machine learning model and shows promise performance in the classification of exercises job.
- Long short-term memory network (LSTM) (Tong, W., Tong, S., Huang, W., et al, 2020): LSTM is one of the deep learning methods and can help to memorize the relationship of each input features.

The experiment results showed in the table 2. According to the results, our model F1-score of 0.7897 and Accuracy of 0.7957 for the matching of ITS exercises with knowledge components. For the others, the accuracy of VSM is 0.5579, SVM is 0.7073, XGBoost is 0.7287, NN is 0.7287, LSTM is 0. 7409. The F1-score of VSM is 0.5559, SVM is 0.6864, XGBoost is 0.7137, NN is 0.7127, LSTM is 0.7382. This performance significantly outperformed then prior state-of-art works.

Table 2. Experiment of Our Proposed Model

Evaluation	Model					
	VSM	XGBoost	SVM	NN	LSTM	WE-KE
Accuracy	0.5579	0.7287	0.7073	0.7287	0.7409	0.7957
Macro F1-score	0.5559	0.7137	0.6964	0.7127	0.7382	0.7897

In summary, our model outperforms then other model across junior high school's Japanese math exercises datasets. The results show that effectively combing both semantic meaning extracting from word embedding based methods and Key concepts comparison can help model to improve the performance. In the next session, we will do some ablation study to see the performance of each single methods.

4.2 Ablation study

In order to better analysis the contribution of each part of our model. We conduct the ablation study to evaluate, the results show in table 3.

According to the results, both methods have promising performance. It can show that word embedding based methods can extract the meaningful semantic information and Keyphrase extraction can effectively organize the exercises by keywords. It should be noticed that the experiment also shows that Keyphrase can effectively show the main topic of exercises, so that the rule-based model performance can be comparable with machine learning method. Moreover, the ensemble model can better analyze both model results and outperform then each single model.

Table 3. *Ablation Study*

Evaluation	Model		
	WE	KE	WE-KE
Macro F1-score	0.7137	0.7350	**0.7897**
Accuracy	0.7287	0.7195	**0.7957**

Besides, we also conduct the experiments of different word vectorization methods by using TF-IDF (Aizawa, A., 2003). This method related on the term frequency and inverted document frequency. The results of experiment showed in table 4. According to the results, the accuracy of VSM is 0.7165, SVM is 0. 8140, XGBoost is 0. 8140, NN is 0. 8171, LSTM is 0.8065. The F1-score of VSM is 0.7129, SVM is 0.8095, XGBoost is 0.8100, NN is 0.8156, LSTM is 0.8025. Our model the F1-score is 0.8364 and accuracy is 0.8384. Our model performance with the TF-IDF vectorization also outperforms others and higher than using Word2Vec. Due to the feature's length of TF-IDF is 866 which almost same as the size of experiment data. Therefore, the model is easily overfitting. We keep using Word2Vec in our main experiment. In conclusion, the ensemble of two methods can provide stable results and outstanding performance in the matching exercises with KCs task.

Table 4. *Experiment of Model with TF-IDF*

Evaluation	Model					
	VSM	XGBoost	SVM	NN	LSTM	WE-KE
Macro F1-score	0.7129	0.8100	0.8095	0.8156	0.8025	**0.8364**
Accuracy	0.7165	0.8140	0.8140	0.8171	0.8065	**0.8384**

4.3 Human test

In this session, we will evaluate our model with expert to check whether our model can effectively help them to save time for checking. Because this work is the fundamental of other ITS implementation such as automatically recommendation and knowledge tracing etc. Therefore, it is important to guarantee the correctness of matching exercises with KCs. It is inevitable to invite the expert into this process. For better filter the available knowledge components, we showed the top 3 results in each selection box according to the model prediction as showing in figure 4. Table 5 showed the comparison of Accuracy@K.

Table 5. *The Accuracy of top k*

Evaluation	Model					
	VSM	XGBoost	SVM	NN	LSTM	WE-KE
Accuracy@2	0.6829	0.8445	0.8445	0.8994	0.8598	**0.8994**
Accuracy@3	0.7378	0.9146	0.9132	0.9421	0.9177	**0.9543**

According to the results, our model Accuracy@2 is 0.8994 and Accuracy@3 is 0.9543 both of them are out perform then others. The result also indicates that the true results are most probably included in the top-ranking results. It can effetely save user time for searching.

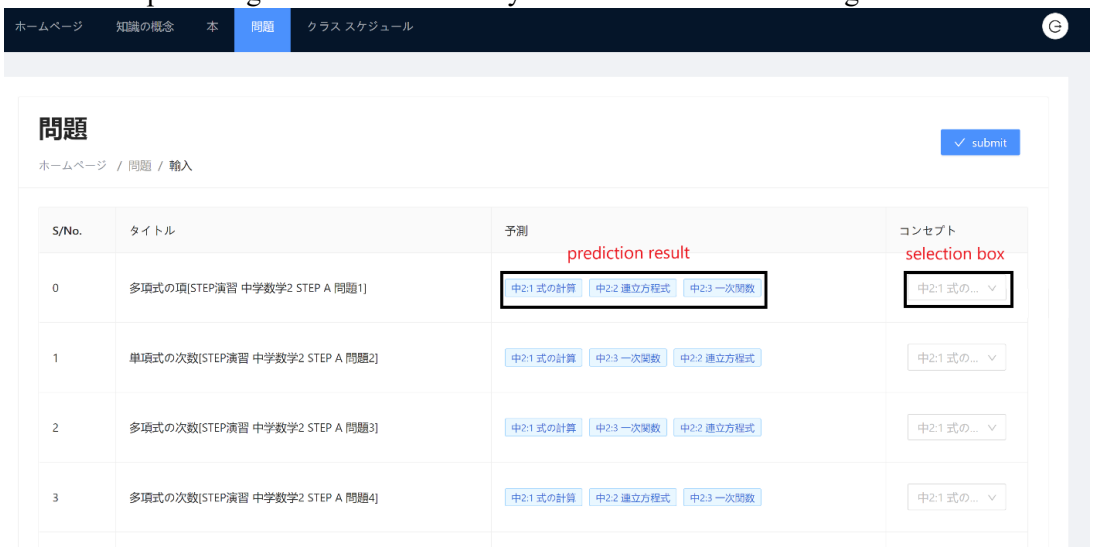


Figure 4. Shortcut of Knowledge Management System

In the figure 4, we develop the simple knowledge portal interface to allow user upload the data and receive the model prediction results. In the "prediction" column, it lists the top three related knowledge components according to the model result. It can help users quickly to find the available choice. For each selection box, the knowledge components are ranked according to the prediction probability and the highest one is default selected by the system. If without the model predictions, all the knowledge components rank by default orders. We conduct preliminary experiment based on it. We invited 6 participants who are the Ph.D. or master's degree students whose researching area is about the education and with good math ability. Then, randomly assign them to different experiment groups and they will not know which group they are in. One group provides the model prediction results and the other require them to assign the KCs to exercises by hands. We split the Junior high school datasets as 6:4 for training datasets and testing datasets than randomly extracted 50 exercises from the testing dataset for the human test.

The results showed in table 6. According to the result, with the help of our prediction results, expert took average 6 minutes less than no prediction function assist. We also calculate the recall as whether the true label in the filtered option list and it has around 0.98. It suggested that almost all the true labels were included in the filter option list and no need for searching. Furthermore, the precision@1 means how many true labels were ranked in the first option. The result was 0.84 and most of true labels were selected as default results. In conclusion, our matching exercises with KCs portals can effectively help experts saving their time for checking the classification results.

Table 6. *The Evaluation Result of Human Test*

Evaluation	Model		
	Time(m)	Recall	Precision@1
Prediction	**6**	0.98	0.84
No-prediction	8	-	-

5. Discussion

Although currently work shows promising performance when matching the exercises with the knowledge components, there still has many challenging and problems needed to be solved. First, parse the pdf file may lose more information then HTML format file. For example, the function information can be completed extracted from HTML file, but hard for pdf file. In many math exercises, the function information can also helpful for extracting the KCs information (Patikorn, T., Deisadze, D., Grande, L., et al, 2019). Second, the geometry is one of the important part for math course and the graph of those exercises are also delivered the important information (Liu, Q., Huang, Z., Huang, Z., et al, 2018). Therefore, heterogenous input features should also be concerned. Moreover, The datasets in our experiments are not plenty enough for comprehensively showing our models performance and the performance of real application. We will try to include more data into our experiments. Besides, our KCs labeled are not enough in the real educational tasks . We should also need to consider the overfitting problem described in the (Patikorn, T., Deisadze, D., Grande, L., et al, 2019). Since, there are a large number of near-identical problems. The model may only remember some specific words, but not the true features of exercises. In collusion, in the future work, we will try to expand our datasets with more KCs and datatype. We will also consider how to utilize the heterogenous data information. For the human test, it also needs to invite more expertise and teachers for the experiment test. We will provide more experiment result and detailed data in the future research report.

6. Conclusion

In this paper, we proposed to use an ensemble method for automatic labeling exercises with specific knowledge component. Comparing with the past researches, our method can maintain the high-performance on the small data size. We try to combine the advantages of word-embedding based model which has great ability for understanding the semantic meaning of texts and Keyphrase extraction model which built up on the domain knowledge and can be easily interpretable. According to the experiment result, the ensemble model outperforms the baselines such as: XGBoost, SVM, LSTM etc. Moreover,

it also better than each single model performance. For better prove the practical effectiveness, we develop the knowledge portal system which embedded with the proposed method and provided simple interface to help experts check and update the model results and relevant knowledge components information. Then, conduct the preliminary study of the human test. The results show that with the help of model, it can save the experts checking time.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, NEDO JPNP20006 and JPNP18013.

References

Aizawa, A. (2003). An information-theoretic perspective of tf–idf measures. *Information Processing & Management, 39*(1), 45-65.

Alzaidy, R., Caragea, C., & Giles, C. L. (2019, May). Bi-LSTM-CRF sequence labeling for keyphrase extraction from scholarly documents. In *The world wide web conference* (pp. 2551-2557).

Chau, H., Labutov, I., Thaker, K., He, D., & Brusilovsky, P. (2021). Automatic concept extraction for domain and student modeling in adaptive textbooks. *International Journal of Artificial Intelligence in Education, 31*(4), 820-846.

Chunamari, A., Yashas, M., Basu, A., Anirudh, D. K., & Soumya, C. S. (2022). Quora Question Pairs Using XG Boost. In *Emerging Research in Computing, Information, Communication and Applications* (pp. 715-721). Springer, Singapore.

Church, K. W. (2017). Word2Vec. *Natural Language Engineering, 23*(1), 155-162.

Conde, A., Larrañaga, M., Arruarte, A., Elorriaga, J. A., & Roth, D. (2016). litewi: A combined term extraction and entity linking method for eliciting educational ontologies from textbooks. *Journal of the Association for Information Science and Technology, 67*(2), 380-399.

Del Solato, T., & Du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education, 6*, 337-378.

Desmarais, M. C. (2012). Mapping question items to skills with non-negative matrix factorization. *ACM SIGKDD Explorations Newsletter, 13*(2), 30-36.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal, 10*(4), 469-484.

Hage, H., & Aimeru, E. (2006, March). ICE: A system for identification of conflicts in exams. In *IEEE International Conference on Computer Systems and Applications, 2006*. (pp. 980-987). IEEE.

Hou, Y., Zhou, P., Wang, T., Yu, L., Hu, Y., & Wu, D. (2016). Context-aware online learning for course recommendation of MOOC big data. *arXiv preprint arXiv:1610.03147.*

Hou, Y., Zhou, P., Xu, J., & Wu, D. O. (2018, April). Course recommendation of MOOC with big data support: A contextual online learning approach. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 106-111). IEEE.

Huang, T., & Li, X. (2021). An Empirical Study of Finding Similar Exercises. *arXiv preprint arXiv:2111.08322*.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Karlovčec, M., Córdova-Sánchez, M., & Pardos, Z. A. (2012, June). Knowledge component suggestion for untagged content in an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems* (pp. 195-200). Springer, Berlin, Heidelberg.

Liu, Q., Huang, Z., Huang, Z., Liu, C., Chen, E., Su, Y., & Hu, G. (2018, July). Finding similar exercises in online education systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 1821-1830).

Matayoshi, J., & Lechuga, C. (2020). Automated Matching of ITS Problems with Textbook Content. In *iTextbooks@ AIED* (pp. 17-28).

Nagisa (0.2.7). (2020, July 7). Retrieved April 25, 2022, from: https://github.com/taishi-i/nagisa

Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019, October). Graph-based knowledge tracing: modeling student proficiency using graph neural network. In *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)* (pp. 156-163). IEEE.

Pardos, Z. A., & Dadu, A. (2017, July). Imputing KCs with representations of problem content and context. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (pp. 148-155).

Patikorn, T., Deisadze, D., Grande, L., Yu, Z., & Heffernan, N. (2019, June). Generalizability of methods for imputing mathematical skills needed to solve problems from texts. In *International Conference on Artificial Intelligence in Education* (pp. 396-405). Springer, Cham.

Pdftotext (2.2.2).(2021, November 24). Retrieved April 25, 2022, from: https://github.com/jalan/pdftotext

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. *Advances in neural information processing systems, 28*.

Shen, J. T., Yamashita, M., Prihar, E., Heffernan, N., Wu, X., McGrew, S., & Lee, D. (2021, June). Classifying Math Knowledge Components via Task-Adaptive Pre-Trained BERT. In *International Conference on Artificial Intelligence in Education* (pp. 408-419). Springer, Cham.

Siddiqi, S., & Sharan, A. (2015). Keyword and keyphrase extraction techniques: a literature review. *International Journal of Computer Applications, 109*(2).

Tiana, Z., Zhengc, G., Flanaganb, B., Mic, J., & Ogatab, H. BEKT: Deep Knowledge Tracing with Bidirectional Encoder Representations from Transformers.

Tong, S., Liu, Q., Huang, W., Huang, Z., Chen, E., Liu, C., ... & Wang, S. (2020, November). Structure-based Knowledge Tracing: An Influence Propagation View. In *2020 IEEE International Conference on Data Mining (ICDM)* (pp. 541-550). IEEE.

Tong, W., Tong, S., Huang, W., He, L., Ma, J., Liu, Q., & Chen, E. (2020, November). Exploiting Knowledge Hierarchy for Finding Similar Exercises in Online Education Systems. In *2020 IEEE International Conference on Data Mining (ICDM)* (pp. 1298-1303). IEEE.

Tsinakos, A., & Kazanidis, I. (2012). Identification of conflicting questions in the PARES system. *International Review Of Research In Open And Distributed Learning, 13*(3), 297-313.

Vie, J. J., & Kashima, H. (2019, July). Knowledge tracing machines: Factorization machines for knowledge tracing. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 750-757).

Wong, W., Liu, W., & Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR), 44*(4), 1-36.

Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012, May). Probase: A probabilistic taxonomy for text understanding. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data* (pp. 481-492).

Zhou, Z. H. (2012). Ensemble methods: foundations and algorithms. CRC press.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Using Unsupervised Machine Learning to Model Taiwanese High-School Students' Digital Distraction Profiles Concerning Internet Gaming Disorder

Yu-Lin HO[a], Chien CHOU[b], Chen-Hsuan LIAO[b] & Jiun-Yu WU[b]*
[a]*Arete Honors Program, National Yang Ming Chiao Tung University, Taiwan*
[b]*Institute of Education, National Yang Ming Chiao Tung University, Taiwan*
**jiunyu.rms@gmail.com/jiunyuwu@nycu.edu.tw*

Abstract: Digital distraction is cognitive attention wandering or being directed to digital sources other than the main learning task. Adolescents with digital distractions may also be addicted to Internet gaming or even suffer Internet gaming disorder (IGD), severely harming their physical and mental development and learning performance. Digital distraction is related to low self-esteem, which is one of the antecedents of IGD. Therefore, this study investigates the association between digital distraction and IGD. We collected responses from 793 Taiwanese senior high students. Results showed that students utilize digital devices for entertainment when learning. Two-stage clustering classified students into four groups concerning their digital distraction constructs: perceived attention problems (PAP) and self-regulation strategies (SRS). The IGD-suspected participants were in the groups with strong PAP profile. We found that digital distraction would be associated with IGD. To mitigate IGD, we suggest early digital distraction screening and provide self-regulation strategies for high schoolers to mitigate their attention and IGD issues.

Keywords: Machine learning in Education, Internet Gaming Disorder, Digital Distraction, Self-Regulation Strategies, Learner and Student Modeling

1. Introduction

Internet Gaming Disorder (IGD) has already been included in the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) since 2013. Paulus, Ohamann, Gontard, and Popow (2018) and Saunders et al. (2017) found that IGD occurs more frequently in teenagers, especially in Eastern countries. Research has indicated multiple upstream factors of IGD, such as genes, social requirements, harmful gaming motivations, and low self-esteem (Laconi, Pires, & Chabrol, 2017; Paulus et al., 2018). Despite the fruitful achievement, investigating IGD through educational perspectives has remained almost untreated.

As digital devices become part of our daily lives, some crucial issues and the impact of digital distraction gradually intensify. Specifically, in the online learning context, students are highly likely to be distracted by advertisements, social media, or online games (Wu, 2017). Wu and Cheng (2019) indicated that digital distraction induces low self-esteem, a crucial antecedent of IGD (Paulus et al., 2018).

However, much more research is needed to validate the link between digital distraction and IGD. This study attempts to extend previous digital distraction work by examining how IGD-suspected schoolers' would demonstrate their digital distraction profiles. The study results may provide insights into digital distraction and IGD, improving teaching and learning practices.

2. Literature Review

2.1 Internet Gaming Disorder

IGD was included in DSM-5 by American Psychiatric Association and was defined in terms of nine diagnostic criteria in 2013 as follows:

- Preoccupation with gaming
- Withdrawal symptoms when gaming is taken away or not possible (sadness, anxiety, irritability)
- Tolerance, the need to spend more time gaming to satisfy the urge
- Inability to reduce playing, unsuccessful attempts to quit gaming
- Giving up other activities, loss of interest in previously enjoyed activities due to gaming
- Continuing to the game despite problems
- Deceiving family members or others about the amount of time spent on gaming
- The use of gaming to relieve negative moods, such as guilt or hopelessness
- Risk, having jeopardized or lost a job or relationship due to gaming

Although the exact cause of IGD has not been found yet (Gentile et al., 2017), some research has found that the prevalence of IGD has significant differences in sex, age, culture, game genre, and gaming motivation. The prevalence among males is about 11.9% compared with females, which is 2.9% (Laconi, Pires, & Chabrol, 2017; Paulus et al., 2018). Besides, most IGD patients are teenagers (Hawi, Samaha, & Griffiths, 2018; Paulus et al., 2018). Saunders et al. (2017) found that the prevalence in Eastern countries is generally higher.

Many scholars also studied the influences of IGD. Saunders et al. (2017) unified the physical and psychological effects and burdens of IGD, including sleep deprivation, day-night reversal, dehydration, etc. Although these physical and psychological effects do not seem highly life-threatening, some sporadic deaths are caused by prolonged sitting in front of the computer (Lee, 2004).

2.2 Digital Distraction and Self-Regulation Strategies

In a personal learning environment (PLE), most learners would likely use digital devices as learning tools. Although digital learning positively affects learning performance and motivation (Lin et al., 2017), it also shows adverse effects, such as learning procrastination (Wu & Cheng, 2019). When required to use online resources for learning, learners may also visit irrelevant websites, such as social media and internet games, simultaneously (Wu & Xie, 2018) and would be interrupted and keep browsing behaviors unrelated to classes (Taneja, Fiore, & Fischer, 2015). This situation negatively impacts learning performance (Duncan, Hoekstra, & Wilcox, 2012).

However, if students have low learning performance or attention discontinuity, it is not entirely because of the use of digital devices or social media; it might result from their inadequate self-regulation strategies for coping with digital distraction instead (Ho, Liao, Weng, Cheng, & Wu, 2021). Hence, teachers and parents can guide students to establish appropriate self-regulation strategies to regulate distractions and improve their autonomous learning self-efficacy.

3. Method

3.1 Participants

This study utilized data from an extensive national survey. Seven hundred and ninety-three participants were Taiwanese high school or vocational students from grades 10 to 12.

3.2 Ten-Item Internet Gaming Disorder Test (IGDT-10)

This study uses the Ten-Item Internet Gaming Disorder Test (IGDT-10), adapted from Király et al. (2017), which based on the symptoms defined by DSM-5. It is scored on a three-point Likert scale, namely “frequently," “sometimes," and “never.” A participant who chooses “frequently” in five out of ten questions can be determined as an IGD patient. The diagnostic accuracy rate of this cut-off point is as high as 99%, which is also applicable to Taiwanese adolescents (Ko et al., 2014).

3.3 Online Learning Motivated Attention Regulation and Strategies Scale (OL-MARS)

This study uses the Online Learning Motivated Attention Regulation and Strategies Scale (OL-MARS) to assess digital distraction (Wu, 2015). The OL-MARS is scored on a five-point Likert scale.

Wu (2015) indicated that there are two constructs of digital distraction: Perceived Attention Problems (PAP) and Self-Regulation Strategies (SRS). On the one hand, PAP consists of Social Media Notification (SMN) and Lingering Thoughts (LT). SMN indicates how students perceive whether they are susceptible to social media or digital devices' prompt messages, sounds, or vibrations. LT indicates the anticipation and alertness to events unrelated to learning.

On the other hand, SRS consists of Behavioral Strategy (BS) and Outcome Appraisal (OA). BS indicates using external behavioral strategies to control Internet use. OA means how students appraise their online learning outcomes and relate to their emotions to direct their attention mentally.

3.4 Two-stage Clustering Analysis

Clustering analysis, including hierarchical and non-hierarchical clustering analysis, is a statistical method for grouping participants according to the commonality of attributes. Although non-hierarchical clustering analysis is more widely used, its disadvantage is that it is more sensitive to initial distance. However, the agglomeration process of hierarchical clustering analysis is usually less critical. It is currently inclined to use two-stage clustering analysis to address both shortcuts (Milligan & Sokol, 1908). The first stage uses hierarchical clustering analysis. Since the data are all continuous variables, this study selects Euclidean distance, which is the most suitable for continuous variables and the best way to express the distance between two points, and Ward's minimum variance method, to minimize the variance of each cluster. The second stage uses non-hierarchical clustering analysis/unsupervised machine learning. In this study, *k*-means clustering is selected, the number specified in the first stage is used as a *k* value, and the final clustering results are named by their characteristics. We took the sum of SMN and LT as the PAP score and the sum of BS and OA as the SRS score. The scores of PAP and SRS were used as the clustering variables to form heterogeneous groups of PAP and SRS.

4. Results

4.1 Descriptive Statistical Analysis

Among the data from 793 students, 48% were males (52% females), and 79% were senior high school students (21% avocational students). Twenty-seven students were suspected of having IGD, and only five were female. The prevalence was about 3.4%, a little higher than the reported world average, but was in line with the prevalence in Taiwan (Chiu, Pan, & Lin, 2018).

In this study, we found that the three most common reasons for high school students to use digital devices were surfing social media (89.3%), browsing video websites (86.8%), and playing online games (82.8%).

We used the *R* (version 4.0.4) to compute the descriptive statistics of the two scales (i.e., PAP and SRS). The results showed that the data is normally distributed with skewness and kurtosis distributed between ±3. In addition, both scales are reliable (Cronbach's $\alpha \geq .80$), and the data is suitable for cluster analysis, as shown in *Table 1*.

Table 1. *Item statistics of online learning motivated attention regulation and strategies scale (OL-MARS) and ten-item Internet gaming disorder test (IGDT-10)*

Scale/Item	*M*	*SD*	*α*	*Range*		*Skewness*	*Kurtosis*
				MIN	*MAX*		
OL-MARS							
Perceived Attention Problems (PAP)							
SMN1	2.89	1.16	.86	1	5	0.07	-0.73

SMN2	2.85	1.16		1	5	0.09	-0.74
SMN3	2.59	1.14		1	5	0.38	-0.57
LT1	2.61	1.05	.85	1	5	0.28	-0.20
LT2	3.01	1.15		1	5	-0.04	-0.64
LT3	3.09	1.01		1	5	-0.05	-0.13
LT4	2.86	1.02		1	5	0.22	-0.18
Self-Regulation Strategies (SRS)							
BS1	2.52	1.26	.86	1	5	0.44	-0.81
BS2	3.15	1.30		1	5	-0.11	-1.04
BS3	3.10	1.25		1	5	-0.12	-0.91
BS4	3.10	0.97		1	5	-0.02	-0.11
BS5	3.09	1.00		1	5	-0.03	-0.19
BS6	3.28	0.99		1	5	-0.18	-0.04
OA1	3.45	1.17	.85	1	5	-0.38	0.63
OA2	3.20	1.20		1	5	-0.19	-0.83
OA3	3.93	1.06		1	5	-0.68	-0.29
IGDT-10							
IGD1	1.66	0.62	.86	1	3	0.37	-0.67
IGD2	1.35	0.54		1	3	1.21	0.47
IGD3	1.35	0.57		1	3	1.39	0.94
IGD4	1.39	0.58		1	3	1.21	0.45
IGD5	1.20	0.46		1	3	2.27	4.48
IGD6	1.50	0.64		1	3	0.90	-0.27
IGD7	1.36	0.57		1	3	1.33	0.76
IGD8	1.91	0.68		1	3	0.11	-0.85
IGD9	1.13	0.38		1	3	3.03	9.03
IGD10	1.47	0.61		1	3	0.94	-0.14

4.2 Unsupervised Machine Learning: Two-stage Clustering Analysis

4.2.1 Ward's Minimum Variance Method

Regarding the PAP and SRS scores, this study utilized Ward's minimum variance method and Euclidean distance at the first stage of hierarchical clustering analysis. The results are shown in *Figure 1*, suggesting that four groups were the most suitable solution.

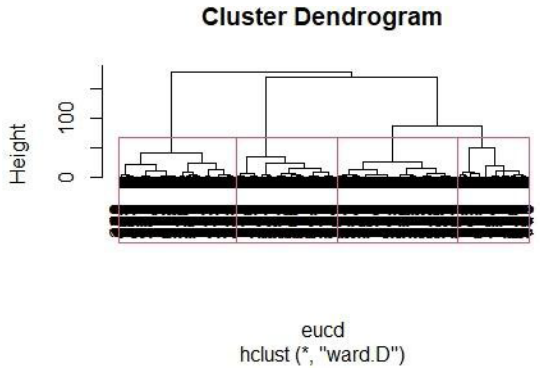


Figure 1. Hierarchical clustering analysis results in a dendrogram

4.2.2 k-means

Based on the results from the first stage, this study set the *k* value as 4 for *k*-means analysis. The clustering results are shown in *Figure 2*.

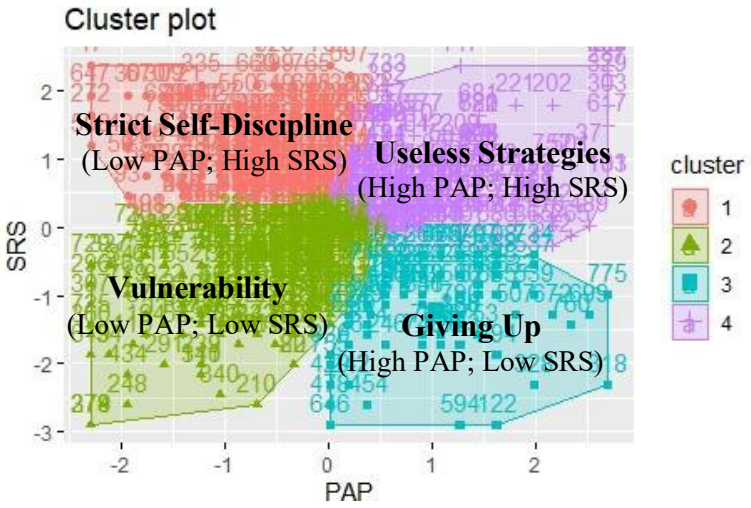


Figure 2. k-means cluster plot

Participants were distinctly divided into four groups: high PAP and high SRS, high PAP and low SRS, low PAP and high SRS, and low PAP and low SRS. Accordingly, these groups were named: Useless Strategies, Giving Up, Strict Self-Discipline, and Vulnerability. We performed further post-hoc with the *Scheffe test* for the results of clustering. The results showed that the SRS scores were significantly different among the groups, which can be expressed as follows:

- Strict Self-Discipline>Useless Strategies>Vulnerability>Giving Up

The PAP scores among the groups can be expressed as follows:

- Useless Strategies≒ Given Up>Vulnerability>Strict Self-Discipline

We further checked the grouping status of the twenty-seven participants who are suspected of having IGD. The results showed that 70.37% were in Useless Strategies (*N*=19), 22.22% were in Giving Up (*N*=6), and only 7.41% were in Strict Self-Discipline (*N*=2).

5. Discussion

This study analyzed the influence of digital distraction on IGD for Taiwanese high school students. We evidenced that Internet has become indispensable in teenagers' daily lives, and playing online games is a prevalent entertainment. We effectively classified participants into four groups by analyzing digital distraction (i.e., PAP and SRS) with a two-stage clustering analysis. Participants suspected of having IGD tend to have higher PAP. Therefore, future instructors should teach students more appropriate self-regulation strategies to alleviate IGD.

6. Conclusion

Easily distracted by digital devices and being unable to stop playing online games are signs of IGD. As digital devices are more prevalent in young learners' daily lives, they are more susceptible to IGD and digital distraction than ever before. This study demonstrates the relationship between digital distraction and IGD from an educational perspective. The results reveal that students with higher PAP are vulnerable to IGD, and appropriate self-regulation strategies may prevent students from having IGD. Hence, it is possible to prevent IGD if instructors can identify students with digital distraction issues early and intervene.

Although we obtain promising results on the early identification of IGD through analyzing digital distraction, this study is not without limitations. This study only used questionnaire scores for analysis. More rigorous experiments (i.e., neuroscience research) should be combined to explore the causal relationship between digital distraction and IGD.

Acknowledgements

This work was partially supported by the Higher Education Sprout Project from the Ministry of Education (MOE) and by the research grant from the Ministry of Science and Technology (MOST 110-2525-H-A49-001-MY4, MOST 108-2511-H-009-011-MY4), Taiwan.

References

Chiu, Y. C., Pan, Y. C., & Lin, Y. H. (2018). Chinese adaptation of the Ten-Item Internet Gaming Disorder Test and prevalence estimate of Internet gaming disorder among adolescents in Taiwan. *Journal of Behavioral Addictions*, *7*(3), 719-726.

Duncan, D., Hoekstra, A., & Wilcox, B. R. (2012). Digital Devices, Distraction, and Student Performance: Does In-Class Cell Phone Use Reduce Learning? *Astronomy Education Review, 11*, 010108.

Gentile, D. A., Bailey, K., Bavelier, D., Brockmeyer, J. F., Cash, H., Coyne, S. M., ... Young, K. (2017). Internet gaming disorder in children and adolescents. *Pediatrics*, *140*(Supplement 2), S81-S85.

Hawi, N. S., Samaha, M., & Griffiths, M. D. (2018). Internet gaming disorder in Lebanon: Relationships with age, sleep habits, and academic achievement. *Journal of behavioral addictions*, *7*(1), 70-78.

Ho, Y. L., Liao, C. H., Weng, M. J., Cheng, T. Y., & Wu, J. Y. (2021, December 3-4). *The preliminary analysis of association between Taiwan senior high school student's digital distraction, learning procrastination, and autonomous learning*. Paper presented at Taiwan Association for Educational Communications and Technology (TAECT 2021), Taipei, Taiwan.

Király, O., Sleczka, P., Pontes, H. M., Urbán, R., Griffiths, M. D., & Demetrovics, Z. (2017). Validation of the ten-item Internet Gaming Disorder Test (IGDT-10) and evaluation of the nine DSM-5 Internet Gaming Disorder criteria. *Addictive behaviors*, *64*, 253-260.

Ko, C. H., Yen, J. Y., Chen, S. H., Wang, P. W., Chen, C. S., & Yen, C. F. (2014). Evaluation of the diagnostic criteria of Internet gaming disorder in the DSM-5 among young adults in Taiwan. *Journal of psychiatric research*, 53, 103-110.

Laconi, S., Pires, S., & Chabrol, H. (2017). Internet gaming disorder, motives, game genres and psychopathology. *Computers in Human Behavior, 75*, 652-659.

Lee, H. (2004). A new case of fatal pulmonary thromboembolism associated with prolonged sitting at the computer in Korea. *Yonsei medical journal*, *45*(2), 349-351.

Lin, M. H., Chen, H. C., & Liu, K. S. (2017). A Study of the Effects of Digital Learning on Learning Motivation and Learning Outcome. *Eurasia journal of mathematics, science and technology education, 13*, 3553-3564.

Milligan, G. W., & Sokol, L. M. (1980). A two-stage clustering algorithm with robust recovery characteristics. *Educational and Psychological Measurement*, *40*(3), 755–759.

Paulus, F. W., Ohamann, S., Gontard, A. V., & Popow, C. (2018). Internet gaming disorder in children and adolescents: a systematic review. *Developmental Medicine & Child Neurology*, *60*(7), 645-659.

Saunders, J. B., Hao, W., Long, J., King, D. L., Mann, K., Fauth-Buhler, M., Rumpf, H.-J., Bowden-Jones, H.,Rahimi-Movaghar, A., Chung, T., Chan, E., Bahar, N., Achab, S., Lee, H. K., Potenza, M., Petry, N., Spritzer, D., Ambekar, A., Derevensky, J., … Poznyak, V. (2017). Gaming disorder: Its delineation as an important condition for diagnosis, management, and prevention. *Journal of Behavior Addictions, 63*(3), 271-279.

Taneja, A., Fiore, V., & Fischer, B. (2015). Cyber-slacking in the classroom: Potential for digital distraction in the new age. *Computers & Education*, *82*, 141-151.

Wu, J. Y. (2015). University students' motivated attention and use of regulation strategy on social media. *Computers & Education, 89*, 75-90.

Wu, J. Y. (2017). The indirect relationship of media multitasking self-efficacy on learning performance within the personal learning environment: Implications form the mechanism of perceived attention problems and self-regulation strategies. *Computers & Education, 106*, 56-72.

Wu, J. Y. & Xie, C. (2018). Using time pressure and note-taking to prevent digital distraction behavior and enhance online search performance: Perspectives from the load theory of attention and cognitive control. *Computers in Human Behavior, 88*, 244-254.

Wu, J. Y., & Cheng, T. (2019). Who is better adapted in learning online within the personal learning environment? Relating gender differences in cognitive attention networks to digital distraction. *Computers & Education*, *128*, 312-329.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Use of Professor Comments in Predicting Student Success

Timothy H. BELL[a*], Christel DARTIGUES[b], Florent JAILLET[c] & Christophe GENOLINI[d]
[a, b, c]*Université Côte D'Azur, CNRS, I3S, France*
[d]*Zébrys, France*
*timothy.bell@univ-cotedazur.fr

Abstract: During their studies students receive written notes and comments from their professors assessing their grades, attitudes, qualities, and lacuna. These characteristics reflect a more subjective approach as opposed to the typical grading system. This paper, through topic modelling and word vectorization approaches, uses textual data to predict at-risk students in their first year of university studies with a Random Forest model. First, we introduce the used methods and analyze the corpus at hand. Then we vectorize the data (by Latent Dirichlet Allocation and other vectorizing methods) to categorize it and use it in the classifier. We then propose adding a dynamic element to the prediction through linear regression when using our data as a time series. Finally, we will review the prediction accuracy and feature importance to assert if these professor comments do indeed reflect the student's scholar capacities. After comparing with the raw numerical grade data, we have better or as-good-as results by using our augmented textual data.

Keywords: Student Success, Random Forest, Allocation, Text Prediction, Time Series

1. Introduction

In France 60% of students in their first year of higher education fail the school year as seen in the MESR-SIES report (2013). Recently more and more research attack this problematic (Alyahyan & Dustegor, 2020). During the high school years, students get graded, and receive comments and notes throughout their studies. The latter, these short texts, allow for a better understanding regarding the student's grades and behavior. We place ourselves in a context where the student is accepted in a university and the teachers want to avoid failure of said student. In the community of educational mining and when predicting student success or detecting at-risk students, most studies mostly use previous grades and input them in Machine Learning models. Numerous works have shown good results (Alyahyan & Dustegor, 2020) especially with Random Forests and more generally ensemble methods (Vijayalakshmi & Venkatachalapathy, 2019; Balaji et al., 2021; Unal, 2020). This paper shows the potential of using Random Forests with transformed textual data.

There are many ways to transform our textual data into numeric vectors, we'll go through some methods relevant to the data used in this work. For a better prediction model, some of the data will be derived to create new variables, by combinations, linear regressions, and frequency differences.

2. Related Works

The work related to student success prediction is very extensive and has always been a pillar of research in educational data mining. The common prediction method is to use the student grades from high school or sociodemographic indicators. Other articles also take in consideration the student's interests and hobbies. A recent extensive state-of-the-art can be seen in Albreiki et al. (2021). Some papers such as Li et al. (2020) and Bell et al. (2021) show that data augmentation can improve prediction. Augmenting the data can be done by rearranging the data or performing various calculations on the data to obtain new features that might be more relevant for the model.

Regarding the use of textual data for student success prediction, there is little research done. The only work, to our knowledge, can be seen in Fateen et al. (2021) and Jayaraman (2020), the first using data from a cram school and the latter using advisor comments for student dropout prediction. Other studies do not use non-student written notes. The aim of this paper is to add to this near-empty research scope in educational research.

3. Dataset

The samples for our data are taken from 2 French university departments (Marketing and Management), we have data from their last 2 years before entering their university curriculum. All the data used in this paper has been entirely anonymized and sensitive data has been omitted. Our target data is whether the students passed or failed their first semester of university. Additionally, we also have absenteeism reported during the first semester. Having binary target data, we are therefore solving a binary supervised classification problem. During high school, students receive grades and professor comments in each subject. Our dataset has a comment/grade pair for each trimester and each year, this work focuses on the last 5 trimesters of high school. The average number of professor notes per student is 50. The data is very short with an average text length of approximately 7 words, the texts are very straight-to-point and often follow the same structure. First commenting on the success or failure of the trimester and then sometimes adding the student's attitude and/or sometimes giving advice. Some examples of sentences are "Unsatisfactory results this trimester" or "okay overall, can still progress" (in this article all the sentences and words will be translated into English).
After removing stop words, we have 3200 unique words in the dataset. Table 1 displays the top words, with "trimester" the most frequent, this word alone doesn't help us (the reader) or the algorithm. So, we also create k-grams up to 3 as we find sentences such as "very good trimester". In Table 1, the top k-grams are also shown revealing terms such as: "very good trimester" and "fairly good trimester". Some k-grams might seem a bit odd such as "trimester good student", this is because stop-words and others are removed during the preprocessing operations.

Table 1. *Top k-grams in the Corpus.*

1-grams	2-grams	3-grams
trimester	good trimester	very good trimester
work	very good	fairly good trimester
results	individual work	trimester good student

Table 2. *Top 1-grams per Class.*

Class	Top 3 1-grams	Class Frequency Difference
Failed	trimester	0.0001
	work	0.00068
	result	0.00054
Passed	trimester	-0.0001
	good	0.00082
	serious	0.00041

If we divide into the passing and failing classes, the top 3 terms for each class are seen in Table 2 with the class frequency difference being simply $\Delta f_w = \frac{f_w}{|C_D|} - \frac{f_w}{|C_{D'}|}$, where f is the mathematical frequency, w is a word or k-gram, C is the corpus. D is a subset of samples, and D' is the subset of samples obtained from the set subtraction $C - C_D$, D corresponds to the current class, for instance the failing class may be the subset D and the succeeding class the subset D'. Due to the size of the corpus, there is variability in the vocabulary, it is therefore expected to have word frequencies close to 0.

4. Methods

The aim of this paper is to focus on the "thumbs up, thumbs down" aspect of these texts and make them as explicit as possible to feed our model. A simple way to pick through our data is by counting occurrences of certain words, words selected from the word rankings of each class. For instance, counting the number of times a sample student has the word "disruptive". So, a student that has the word "disruptive" appear many times will have more probability of belonging to a certain class if this word is decisive in any manner. But more complex methods exist, the next 2 Sections go through various operators that take transformed text features and create scalar values that describe the features' class.

1.1 Latent Dirichlet Allocation

LDA (Blei et al., 2003) is an automatic topic modelling algorithm, often used in NLP, that attributes a topic score for each word. In Figure 1, after preprocessing the raw data in (a) and obtaining vectors of k-grams in (b) for each document (a document here is one single comment or teacher note), we input all the data into an LDA model, and we obtain in (c) a topic distribution matrix containing the k-grams in their probable topic. LDA is found suitable in our case, as we have short sentences, limited vocabulary (teachers tend to reuse the same words), and words that will often appear in similar word contexts and orders of words (obtaining frequent k-grams). The typical way of classifying our documents is using the per-document-per-topic probability γ value of each document (also called Θ by Blei) and each topic.

We explore two other ways to use the values obtained from LDA. Firstly, where v_w is the topic probability vector obtained for each k-gram w in a document c, each value of the vector is the probability of the k-gram to be associated with a topic. A document can be represented as a set of vectors $c'' = v_1, .., v_n$ for n number of k-grams in the document. To reduce the dimensions of c'' to be able to input into our predictive model, we use $c''' = p(c'')$. With p the pointwise product of the k-gram vectors that return a vector the same dimension as v (of size n the inferred number of topics). The resulting vector c''' is split in t_n number of features, t_n is the number of inferred topics. This is illustrated simply in Figure 1 with (d) the set of vectors c'' (or matrix) of size (number of topics in LDA model) ∗ (number of k-grams in document). (e) are the newly obtained features.

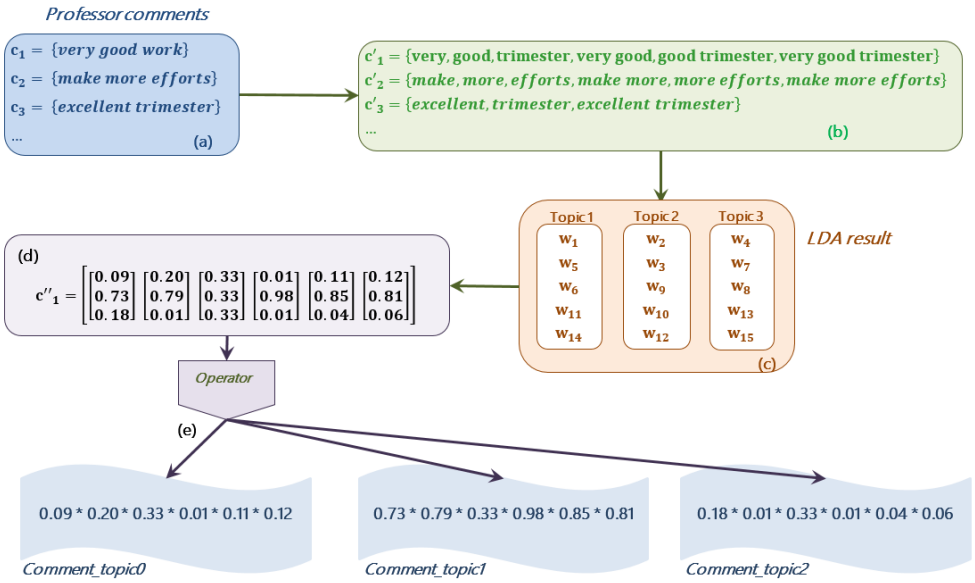


Figure 1. Summary of the RF Models Creation.

Additionally, to create a wider gap between the topics, so that our predictive model can better split depending on how strongly the vectors reflect the association of the k-grams with certain topics,

we perform a pointwise product on the vectors. The resulting scalar is used as a coefficient on indicator functions for each topic. Obtaining a set of features: $\{I(T(c) = t_1).c'''1, \dots, I(T(c) = t_n).c_n'''\}$ with n the number of inferred topics, c_n''' elements of vector c''', and the indicator function using the γ values for each topic to establish whether the class vector is the main document topic.

Another method of transforming our textual data is to take the results obtained in Table 2 and follow the same process as in Figure 1. Therefore, each sample feature would become a set of two features $A_+(c) = \prod_{w \in c} \Delta f_w\,,\ if\ \Delta f_w > 0$ and $A_-(c) = \prod_{w \in c} \Delta f_w\,,\ if\ \Delta f_w < 0$ where *c* is a document and *w* the k-grams of the document.

1.2 Dynamic Aspect

As found in Bell et al. (2021) and Li et al. (2020) we can find relevant features for an accurate prediction by taking account of the progress of a student over time. By doing linear regression on the values throughout the trimesters we obtain the regression coefficient β of the evolution of the student. As it may be more relevant to look at how the student improves at specific combinations of trimesters. For each subject (e.g., Mathematics), we would multiply the number of each feature by $\sum_{i=2}^{5} \binom{5}{i}$ extending our feature space by 26 for the yearly professor notes and (number of subjects) $*$ 26. In our dataset we have 9 different subjects that carry information through the 2 years. The data will therefore get 260 newly added features for each method of transforming the documents into numerical data. For the increase coefficient over the span of trimesters 1,4, and 5, we will note $\beta_{1,4,5}$.

1.3 Random Forest

We chose Random Forests (RFs) (Breiman, 2001) as it has proven its use in student success prediction (Zhang et al., 2021; Zeineddine et al., 2021). RFs are an ensemble method consisting of multiple decision trees. Each decision tree is trained on a random collection of samples as well as a random collection of features. As we have limited samples, we will cross-validate 10 RFs. We use accuracy and the F1-score to compare models. Each RF model is trained on detecting the passing class and the failing class. In our RF model we use the Gini index (Zhu et al., 2014) for the tree splits. Having this index gives us a built-in feature importance model. In our results section we will discuss the top features selected when training the model, helping us compare the advantages of adding certain sets of features.

We will create RFs for each feature group, and also one RF combining all feature groups, and then we'll compare the results with each other and with a prediction using only the numerical data. The compared groups are the k-gram occurrence, the γ value per-document-per-topic, the pointwise product of the LDA vectors of each document, the pointwise product as a coefficient, the class frequency product, the combination of all feature groups.

5. Results

After trialing, we decided that 3 topics should be informed to our LDA model. Dividing our word bank into a first category with k-grams like "Good Trimester" and "Good", another with words like "Difficult" and "Barely", and finally, "Suitable" and "Suitable Results" for the last category. With all the augmentation done, we obtain a total of approximately 6000 features, mostly features discussed in 4.2. Most of the features are irrelevant to our prediction and do not help in splitting the nodes in our RF. Ideally, work done in this paper should encourage feature selection to get rid of useless features.

1.4 Performance Review

LDA topic modelling works quite well with this data, probably since the texts are quite short and contain little noise. Also, the teacher wants to convey positive or negative assessments in a clear-cut fashion (e.g., "Serious student. Good grades."). Therefore, when looking at the topics as a matrix (illustrated in

(e) in Figure 1) we can see that the "bad" topic contains k-grams such as "need work", "bad", etc. But some k-grams may not be placed in the "good" topic, such as "serious overall" or "satisfying good work" which are put into the "neutral" topic.

In Table 3 we display accuracies for the augmented data groups, the group with the lesser performance is counting element occurrences, with an accuracy quasi-identical to the passing rate (56%). The Class Frequency Product doesn't perform much better either. The rest perform similarly and show sufficient results in our use case, additionally, when using all feature groups in a single predictive model we obtain good results.

Table 3. *Performance of feature groups.*

	Accuracy	Failing class f1-score	Passing class f1-score
Element occurrence	0.58	0.39	0.76
LDA γ value	0.65	0.36	0.82
LDA pointwise product	0.62	0.35	0.80
LDA coefficient pointwise product	0.65	0.41	0.78
Class frequency product	0.60	0.35	0.75
All combined	0.69	0.53	0.86

Predicting with only grades achieves an accuracy of 70% in the marketing department and 73% in the management department. The difference between this accuracy and what we get with our model (using the best accuracy) is 1% only for the marketing department and 6% for the other, both in favor of our model. This big difference may simply be since marketing in this specific establishment may be more demanding (only 56% pass).

1.5 Most Decisive Features

Ultimately, we are predicting at-risk students in their first year of university. Knowing what features describe this "failing" class of students is necessary. The best predictor for the marketing class is the β value of trimesters 3, 4, and 5 of the pointwise vector multiplication of the LDA vectors. As a note, in referral to the pointwise product illustrated in Figure 1, we tried different calculations (non-extensively) for step (e) and found that only the product can generate useful features. The k-gram occurrence counts only showed one feature present in the top 100 predictive features and it has a low importance score. Among the different subsets of data in our dataset, element occurrence counts are never in the top splitting. In the top 20 features for each department, 18% of features are LDA per-document-per-topic probabilities, 31% are coefficient based LDA pointwise products, 14% simple LDA pointwise products, the rest are mainly β values with combinations of trimesters. Less than 2% is of the class frequency product type.

A note on absenteeism. When using the model to predict the absenteeism obtain good results, for a proportion of 17% of reported absent students we have an 89% percent accuracy in prediction. When using the purely numerical data, this characteristic is unpredictable. The advantage of using teacher notes and comments for predicting things like absenteeism is quite clear. For this prediction, the occurrence of k-grams relating to "absent" or "disruptive" are features that appear in the top 20 most decisive features. The β value of these k-gram occurrences do not appear however to be decisive in any split of our RF model, this can be explained by the low presence of the said k-grams in some trimesters over the years, and therefore no trend can be extrapolated through linear regression.

When comparing with the raw grade prediction or other predictions from papers mentioned throughout this article, the accuracy of using only textual data shows that we obtain similar or sometimes better results. Attesting the not-so-subjective characteristics of theses short texts.

6. Conclusion

Over the last years, student success prediction has been a core field of study for the community. Typically, grades, and other meta data are used to predict student success, some ideas on possible

predictors can be found in Bell et al. (2021) and Zhao et al. (2020). In this paper we used comments and notes that describe the student's behavior among other things. These short texts carry information that cannot be conveyed through the previously mentioned features, they allow the teachers to gain more insight into the behavioral aspects of students.

Our goal in this study was to explore if written comments and notes from teachers can give out more information than using grades, as we believe that as well as discussing student grades and quality of work, these short texts carry information on other elements such as behavior or absenteeism. In this paper we found that using the textual data can carry more relevant information in the prediction model, as we see that augmenting the data by calculations or a dynamic approach, results in slight improvement in prediction. We also saw that predicting absenteeism has shown promising results, as information regarding this can only be extracted through written notes as in most datasets used for predicting student performance there are no metrics relating to absenteeism. Understanding information like this can be used as a tool for teachers to anticipate student needs as soon as the first semester of higher education.

In future works, we firstly hope to explore different methods of using the textual data in the hope of achieving better performance and then find ways to include numerical data to the model to get a more accurate prediction in the end. We would also like to—in the same way as absenteeism—predict "implication" of a student during his classes. Another way to further extend this study would be to perform sentiment analysis on the comments which may give us a different dimension to the text.

References

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences, 11(9*).

Alyahyan, E., & Düştegör, D. (2020). Predicting academic success in higher education: Literature review and best practices. *International Journal of Educational Technology in Higher Education, 17(1)*, 3.

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review. *Applied Sciences, 11(21)*.

Bell, T., Dartigues-Pallez, C., Jaillet, F., & Genolini, C. (2021). Data Augmentation for Enlarging Student Feature Space and Improving Random Forest Success Prediction. In *Artificial Intelligence in Education: 22nd International Conference*, Proceedings, Part II. 82–87. Springer-Verlag.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet Allocation. *J. Mach. Learn. Res., 3*, 993–1022.

Breiman, L. (2001). Random Forests. *Mach. Learn., 45(1)*, 5–32.

Fateen, M., & Ueno, K., & Mine, T. (2021). An Improved Model to Predict Student Performance using Teacher Observation Reports. In *Proceedings 29th International Conference on Computers in Education Conference*. 31-40.

Jayaraman, J. (2020). "Predicting student dropout by mining advisor notes," in *Proceedings of The 13th International Conference on Educational Data Mining (Ifrane),* 629–632.

Li, H., Ding, W., Yang, S., & Liu, Z. (2020). Identifying at-risk K-12 students in multimodal online environments: A machine learning approach. In *Proceedings of The 13th International Conference on Educational Data Mining*. 137–147.

MESR-SIES. Réussite et échec en premier cycle. Note d'Information Enseignement supérieur Recherche. (2013). Retrieved from https://www.enseignementsup- recherche.gouv.fr/sites/default/files/ imported_files/documents/NI_MESR_13_10_283447.pdf.

Unal, F. (2020). Data mining for student performance prediction in education. In *Data Mining, chapter 9*. IntechOpen, Rijeka.

Vijayalakshmi, V., & Venkatachalapathy, K. (2019). Comparison of Predicting Student's Performance using Machine Learning Algorithms. *International Journal of Intelligent Systems and Applications*, 11, 34-45.

Zeineddine, H., Braendle, U.C., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. *Comput. Electr. Eng.*, 89, 106903.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational Data Mining Techniques for Student Performance Prediction: Method Review and Comparison Analysis. *Frontiers in Psychology*, 12.

Zhao, Y., Xu, Q., Chen, M., & Weiss, G. (2020). Predicting Student Performance in a Master of Data Science Program Using Admissions Data. In *Proceedings of The 13th International Conference on Educational Data Mining*. 325-333.

Zhu, W., Feng, J., & Lin, Y. (2014). Using Gini-Index for Feature Selection in Text Categorization. In *Proceedings of the 2014 International Conference on Information, Business and Education Technology*. 76-80. Atlantis Press.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Learning Algorithm Implementation Structures for Multilabel Classification via CodeBERT

Karl Frederick ROLDAN[a], Gerd Lowell JANA[b], John Kenneth LESABA[c] & Joshua MARTINEZ[d]
[a]*Undergraduate, Ateneo de Naga University, Philippines*
[b]*Undergraduate, Ateneo de Naga University, Philippines*
[c]*Undergraduate. Ateneo de Naga University, Philippines*
[d]*Thesis Advisor, Ateneo de Naga University, Philippines*
[a]kroldan@gbox.adnu.edu.ph

Abstract: Task constraint feedback is the collective name for any kind of feedback system that checks whether problem-defined constraints were fulfilled by students upon submission of work. This can be as simple as checking if certain programming constructs exist, or if a specific algorithm or data structure required by the problem is fulfilled. Most of these systems use static analysis (Fischer, 2006; Gotel, 2008) or natural language processing techniques (Lane, 2005) to generate feedback. A transformer is a neural network for sequence processing, such as natural languages. Previous work has shown that transformers can be generalized for programming language tasks such as code summarization. In this study, we used the CodeBERT transformer to classify or tag algorithms implemented in some code snippets to check constraint satisfaction. Using a custom dataset containing source code aiming to implement algorithms, we show that CodeBERT is capable of learning structures of how code is implemented regardless of how a programmer names the code. Averaging each label's f1-score, the model was able to obtain an average of 0.85, which showed promising results in the dataset.

Keywords: Deep Learning, Programming Language Processing, Sequence Models, Multilabel Classification, Transfer Learning

1. Introduction

Code summarization is one of the heavily researched areas in programming language processing. (Alon, 2019; Yahav, 2019) It aims to give a semantic description of the input code snippet. For example, a *merge sort* function may be described by a code summarization model as a *sort*. This paper deals with a subset of code summarization that the study calls *algorithm detection*, which focuses on classifying functions more specifically. Given a code snippet, an algorithm detection model attempts to find all the included algorithms and label them as such. This model will form the basis for developing a part of an intelligent tutoring system (ITS) for Project CodeC, a competitive programming application currently under development at Ateneo de Naga University. Specifically, a task-constraint based feedback system that analyzes an input whether it contains the required algorithms by the problem setter.

In this paper, the researchers aim to develop a sequence model that acts as a multilabel classifier to tag seven introductory algorithms and two introductory data structures from code snippets. The included programming languages in the study are *Python*, *Java*, and *JavaScript,* which are covered by the ITS being developed. Furthermore, the included algorithms and data structures would be *insertion sort*, *merge sort*, *selection sort*, *quick sort*, *bubble sort*, *linear search*, *binary search*, *linked lists*, and *hash maps*. Since this paper focuses on algorithm detection, evaluation of real-world student submissions will not be studied, but hopes to serve as a foundational study for a future ITS.

2. Review of Related Literature

2.1 Programming Language Processing

Sequence models are machine learning algorithms that are designed to input and/or output sequential data such as time series, audio, signals, texts, and videos. Multiple-sequence neural networks have been proposed with varying performances on different tasks in the past decades. In this paper, the researchers make use of the transformer that is currently the state-of-the-art sequence model. (Vaswani, 2017) Transformer models do not process the input sequentially, which renders it easily parallelizable. The auto-encoding transformer BERT (Jacob Devlin, 2019) was pre-trained on a corpus of 16GB of data, while RoBERTa (Liu, 2019) had a larger corpus of 160GB and achieved better results than the former. Recent breakthroughs have shown that sequence models can be generalized to programming languages as well. Central themes in research include code summarization (Rui Xie, 2021) which can help in automatically commenting on a snippet of code, code-clone detection (Hui-Hui Wei, 2017) which can be used for finding functional similarities for two code snippets that look distinct from each other, and code completion (Alon, 2020) which can be used for assistance when coding. Furthermore, these models can learn programming language semantics by leveraging the Abstract Syntax Trees (AST), (Lili Mou, 2016; Long Chen, 2019) encoding AST paths into tokens which can be processed by traditional networks (Alon, 2019; Yahav, 2019) and by Hoare triples of a program. (Piech, 2015). Furthermore, the CodeBERT autoencoding transformer is a replica of RoBERTa but trained on an entirely different dataset, which is composed of *Java*, *Python*, and *JavaScript* snippets. (Feng, 2020)

2.2 Task-Constraint Feedback

Task-constraint feedback refers to systems that propagate feedback on whether some required tasks have been met by the student. Some of the research done includes WeBWork, which attempts to perform tests that mimic software quality assurance testing (Gotel, et al., 2008) and ProPL, which is a dialogue-based system that aims to help programmers learn program planning by asking questions (Lane, H. C. 2005). Both of these systems give a response on whether the student has correctly satisfied the given constraints. This paper aims to provide a system that is capable of detecting algorithms in a submitted snippet that can be used for teaching algorithms and related classes.

3. Technical Background

3.1 CodeBERT

A transformer is an encoder-decoder sequence model that is non-recurrent, that is, it takes the input sequence without the need for loops like Recurrent Neural Networks would use. (Vaswani, 2017) The BERT family of transformers are categorized as autoencoders, that is, they only contain the encoder part, which can be used for compressing the input data. (Devlin, 2019)

The CodeBERT transformer is an exact replica of RoBERTa trained in six different programming languages. (Feng, 2020) Furthermore, CodeBERT was trained using unimodal data, consisting solely of code snippets, and bimodal data, comprising a pair of a code snippet and an English language description of the snippet. Since CodeBERT is already a trained transformer, this study utilizes the learned features to make better predictions from a small dataset. To optimize the model, the researchers fine-tuned the last three out of twelve encoding layers of CodeBERT. This enables CodeBERT to learn the dataset gathered by the researchers, instead of relying solely on the artificial neural network built on top of CodeBERT.

4. Methodology

4.1 Dataset and Labeling

Since there are no publicly available datasets on programming language processing that aim to classify algorithms, the researchers gathered this study's dataset from public GitHub repositories containing Python, Java, and Javascript source codes which aim to implement included algorithms and data structures. Figure 1 shows that *linear search* is the most-gathered algorithm in the custom dataset and *binary search* has the fewest. Most of the files in the dataset implement only one of the algorithms in the study. After labeling, features that do not contribute to code semantics, such as extra whitespaces, comments, and docstrings were removed. The researchers achieved an almost perfect agreement of 0.829 during labeling according to Kraemer's extension of the Kappa coefficient. (McHugh, 2012)

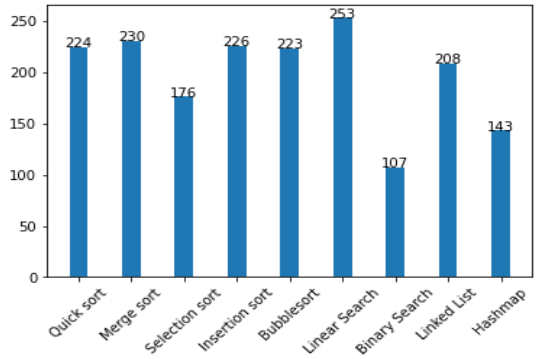


Figure 1. Number of Labels per Algorithm/Data Structure.

4.2 Training

Initially, the model was trained with frozen CodeBERT weights. That is, CodeBERT was used as a feature extractor under the assumption that it had learned some basic algorithms during its pre-training tasks. However, this was not the case, and the model had poor performance. Two main problems were found. First, CodeBERT did not learn these algorithms at all during the pre-training. Second, the model merely memorized identifier names and function orders.

For the first problem, the researchers fine-tuned the first three encoding layers of CodeBERT. This enabled CodeBERT to learn algorithm implementations on top of programming languages. The output is used as an input to a feedforward neural network. For the second problem, the dataset was augmented as described in the next subsection.

4.3 Data Augmentation

To address the problem where the model memorized identifier names, the dataset was augmented by swapping identifier names (function names, class names, and variable names) with randomly generated strings of lengths ranging from 1 to 3. The short length maximizes the probability that CodeBERT will not tokenize the identifier as unknown. Otherwise, the model might have a bias towards unknown tokens. The researchers took steps to ensure that the mappings are bijective. Each file was augmented between 0 to 6 times to ensure randomness in identifier names. This will ensure that the model used in the study learns the structure rather than names. However, language keywords and common functions were not augmented as these are usually crucial to how the algorithm and the programming language work.

Another problem found was that the model was memorizing the ordering of functions in the file. To mitigate this problem, the researchers concatenated two files of different programming languages that do not have common algorithms implemented and added both concatenation orders into the dataset under some probability. The resulting files were then augmented further with random strings from the first augmentation technique discussed

4.4 Evaluation

Precision, recall, and f1-score were recorded for each label for model evaluation purposes. In this paper, the researchers maximized the precision to ensure that an unmet task will not pass the constraint checker.

5. Results

5.1 Experimental Setup

The first experiment result shown in this paper combines the use of augmented and original datasets for the train set. The test set's augmentation does not use random identifiers. The second experiment removes all the unaugmented code snippets from the train set and moves them to the test set. Hence, the distribution of data between the train set and test set in the second experiment is completely different. The goal is to have the model generalize well.

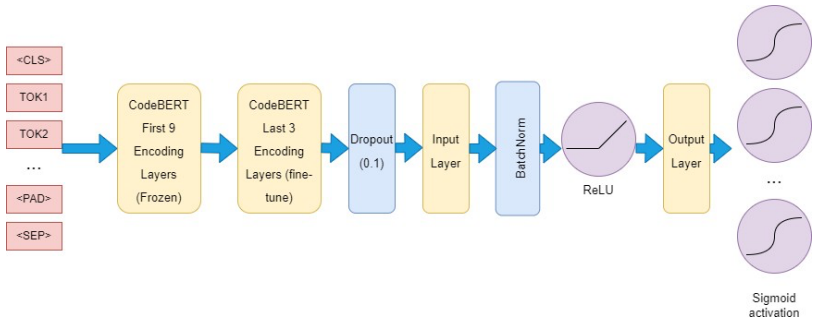


Figure 2. Model architecture.

Figure 2 illustrates the model architecture. To prevent overfitting the dataset, a dropout layer was used prior to feeding the output of CodeBERT into the artificial neural network. Preliminary experiments had all CodeBERT weights frozen. However, this proved ineffective and incapable of overfitting even the training set. Fine-tuning the last three layers proved to be enough to learn the training set and generalize well on the test set. The output layer is activated by nine sigmoid functions, each pertaining to one class. The researchers used binary cross-entropy loss as the loss function.

5.2 Training results

The first experiment makes use of traditional data augmentation techniques wherein the augmented files are used as additional data points for the dataset in the hopes that this will increase the dataset size. Table 3 shows that the model was able to achieve very good results, achieving almost perfect f1 metrics. However, upon experimentation, one problem that was found was implementing an empty function named after one of the algorithms that the model incorrectly classifies. This shows that the model still uses identifiers to make predictions instead of relying purely on code structure.

Table 1. *Evaluation Metrics. (COM) refers to combined augmented and unaugmented dataset during training. (AUG) refers to having only the augmented dataset for training.*

	Recall		Precision		F1 Score	
	COM	AUG	COM	AUG	COM	AUG
Quick Sort	0.938	0.78	0.988	0.957	0.963	0.86
Merge Sort	0.936	0.74	0.994	0.962	0.964	0.836
Selection Sort	0.935	0.81	0.944	0.85	0.939	0.829
Insertion Sort	0.931	0.816	0.974	0.926	0.952	0.867
Bubble Sort	0.919	0.824	0.989	0.836	0.953	0.83
Linear Search	0.936	0.792	0.986	0.82	0.961	0.806
Binary Search	0.914	0.911	0.991	0.886	0.951	0.899
Linked List	0.93	0.876	0.981	0.838	0.955	0.906
Hash Map	0.939	0.779	0.963	0.881	0.951	0.827

To mitigate the bias towards identifier names, the unaugmented dataset was removed from the training set and moved to the test set. Therefore, CodeBERT is forced to learn only the structural information since only files with random identifiers remain in the training set. As shown in Table 1, the model was able to generalize well despite having different data distributions between the train set, composed of augmented files, and the test set, composed of unaugmented files.

Table 1 shows that the first experiment yielded better metrics. However, note that this is due to CodeBERT using identifier names to make predictions and cannot generalize well on experimental inputs. The second experiment shows slightly lower metrics. This is because the model does not make use of identifier names to create predictions but solely leverages on the function structure. Hence, the model was confused for some of the function names but still displayed good predictive power as demonstrated by the high precision value.

5.3 Examples

To experimentally evaluate each trained model, the researchers created a simple web-based application that tags algorithms present in a code snippet. To demonstrate the sanity of the model, Figure 3 shows a code snippet that implements a selection sort algorithm but is named "binary search". Since the model was trained on random identifiers, it was able to correctly identify "selection sort" on structure alone.

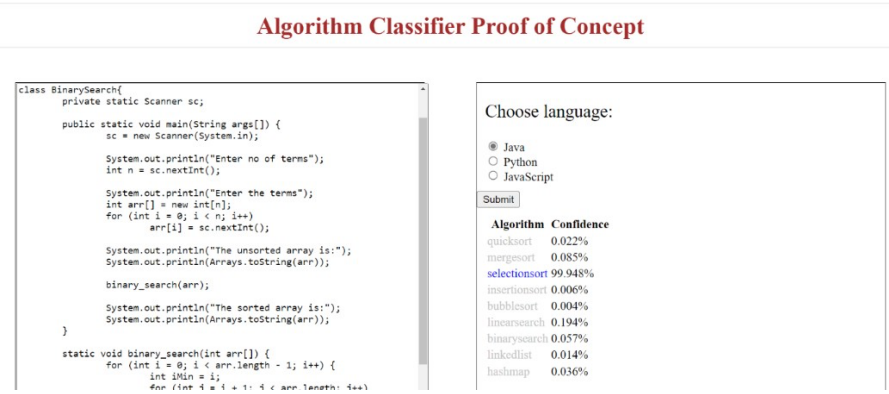


Figure 3. Implementing *selection sort* but *binary search* as identifier.

6. Conclusion

CodeBERT was originally trained for tasks such as code summarization and code search. However, the study was able to demonstrate that CodeBERT is not only capable of learning the meaning of a given code snippet, which can be used for automatic commenting, but can also learn structural information and make predictions from this. Furthermore, the model was able to generalize well and will mostly lead to correct predictions if the input code snippet is in the same distribution as the test set.

Since the study has proven that CodeBERT is capable of learning structural information, a more enhanced model with more classes and datasets geared towards student submissions can be used for algorithm tagging. This model can eventually be used for a task-constrained feedback system. Both the code (Roldan, 2022a) and the dataset (Roldan, 2022b) for this study can be freely accessed on GitHub.

6.1 Recommendations

This study only investigates whether CodeBERT can learn code structure without the aid of abstract syntax trees or graphs. This does not prove that the model can properly predict algorithms from student submissions. Therefore, the researchers recommend that this area be studied and how useful the model will be in aiding task-constraint feedback. To do this, the researchers recommend gathering the dataset from competitive programming websites, such as Project CodeC, to eliminate expected bias against student submissions. To mitigate CodeBERT limitations, such as the 512 max token length and limitations of language choices, other models such as graph neural networks can be investigated.

Acknowledgements

We would like to thank all the people who prepared and revised previous versions of this document.

References

Alon, U. S. (2020). Structural language models of code. *ICML'20: Proceedings of the 37th International Conference on Machine Learning*, 245-256.

Alon, U. Z. (2019). Code2Vec: Learning Distributed Representations of Code. *Proc. ACM Program. Lang.*

Feng, E. A. (2020). CodeBERT: A Pre-Trained Model for Programming and Natural Languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1536–1547.

Fischer, G. (2006). Distributed intelligence: extending the power of the unaided, individual human mind. *AVI '06: Proceedings of the working conference on Advanced visual interfaces*, 7-14.

Gotel, E. A. (2008). Global perceptions on the use of WeBWorK as an online tutor for computer science. *2008 38th Annual Frontiers in Education Conference*.

Hui-Hui Wei, M. L. (2017). Supervised deep features for software functional clone detection by exploiting lexical and syntactical information in source code. *IJCAI'17: Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3034–3040.

Devlin, (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171-4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Lane, H. C. (2005). Teaching the tacit knowledge of programming to noviceswith natural language tutoring. *Computer Science Education, 15*, 183-201.

Lili Mou, G. L. (2016). Convolutional neural networks over tree structures for programming language processing. *AAAI'16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 1287-1293.

Liu, Y. a. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* Retrieved from arXiv: https://arxiv.org/abs/1907.11692

Long Chen, W. Y. (2019). Capturing source code semantics via tree-based convolution over API-enhanced AST. *Proceedings of the 16th ACM International Conference on Computing Frontiers*, (pp. 174-182).

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica 22, 3*, 276-282.

Piech, C. H. (2015). Learning program embeddings to propagate feedback on student code. *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning*, 1093-1102.

Roldan (2022b). *Algorithm Multilabel Classifier dataset*. Retrieved from GitHub:n https://github.com/gerdiedoo/alg-dataset

Roldan (2022a). *Algorithm Multilabel Classifier source code*. Retrieved from GitHub: https://github.com/karlfroldan/algorithm-detection

Rui Xie, W. Y. (2021). Exploiting Method Names to Improve Code Summarization: A Deliberation Multi-Task Learning Approach. *2021 IEEE/ACM 29th International Conference on Program Comprehension*, 138-148. Vaswani, e. a. (2017). Attention is all you need. *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 6000-6010). Red Hook: Curran Associates Inc.

Yahav, U. A. (2019). code2seq: Generating Sequences from Structured Representations of Code. *International Conference on Learning Representations.*

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Micro- and Macro-Level Features of NLP-Based Writing Tools in Higher Education

Michael BURKHARD[a*], Sabine SEUFERT[a], Patcharin PANJABUREE[b], Chailerd PICHITPORNCHAI[c] & Christina NIKLAUS[d]
[a]*Institute for Educational Management and Technologies, University of St.Gallen, Switzerland*
[b]*Faculty of Education, Khon Kaen University, Thailand*
[c]*Institute for Innovative Learning, Mahidol University, Thailand*
[d]*Institute for Computer Science, University of St.Gallen, Switzerland*
*michael.burkhard@unisg.ch

Abstract: This paper reviews tools that use natural language processing (NLP) to support writing in higher education. These NLP-based writing tools usually include various features such as grammar and spelling suggestions, but also general feedback such as the overall readability or consistency of the text. In contrast to previous studies, these feedback features are analyzed and described in detail to give a more accurate overview of the current capabilities of these tools. In a systematic analysis, we present 14 NLP-based writing tools that provide different types of micro-level feedback (e.g., grammar, spelling), macro-level feedback (e.g., argumentative structure, rhetorical moves), as well as additional services (e.g., plagiarism checks, access to writing templates). In contrast to previous findings, the state-of-the-art analysis revealed that many writing tools provide feedback at the micro-level and the macro-level. However, macro-level feedback and further additional services are often only included in paid subscription plans. Overall, the presented features of NLP-based writing tools could be helpful in designing own use cases as they may serve as a checklist and guideline to develop tools that go beyond what already exists.

Keywords: Academic writing, natural language processing, writing tools, higher education

1. Introduction

Academic writing skills are imported 21st-century skills (van Laar et al., 2017; Oberländer et al., 2020) and help students convey their understanding and critical thinking. However, in many contexts, it is challenging to promote *academic writing skills* (Thaiss & Zawacki, 2006; Stevenson & Phakiti, 2014). For example, lecturers often do not have the time to provide detailed and individual feedback to each student. The sparse feedback students receive usually provides them with inadequate guidance on how to develop their writing skills further and leads to frustration among students (Seufert & Spiroudis, 2017).

To address this issue, recent advances in artificial intelligence (AI) and natural language processing (NLP) can be used to provide students with an additional, individual, and adaptive feedback channel to improve the overall quality of the text (Zhang, 2013; Rapp & Kauf, 2018; Strobl et al., 2019). NLP refers to the processing of natural language information with the help of a computer (Gu, 2020). For this purpose, elements from the field of linguistics are combined with various AI approaches (Ding, 2019; Vajjala et al., 2020). In the context of this paper, we define *NLP-based writing tools* as tools, programs, or apps that provide NLP-based feedback on an inserted text. These *NLP-based writing tools* usually include various *features* such as grammar and spelling suggestions, but also general feedback on specific aspects of the text (e.g., readability, persuasiveness of arguments, degree of formality, consistency).

Previous research has already examined *writing tools used in education* (e.g., Allen et al., 2016; Cotos, 2015; Stevenson & Phakiti, 2014). Allen et al. (2016) suggested a classification of three different computerized writer support functions, which they called *Automated Writing Evaluation* (AWE), *Automated Essay Scoring* (AES), and *Intelligent Tutoring Systems* (ITS). The literature review of Strobl et al. (2019) analyzed 44 writing tools used in an academic context, distinguishing between *process-oriented* and *product-oriented* tools. According to Strobl et al. (2019), *process-oriented* tools

are *interactive writing platforms* (IWP) that support learners' learning process using prompts and scaffolds. In contrast, *product-oriented* tools generally include *automated writing evaluation* (AWE) and can provide formative feedback on a specific text (by using NLP).

Within AWE, Strobl et al. (2019) further distinguish between *micro- and macro-level* feedback. *Micro-level* feedback is based on factual knowledge and provides feedback on individual words or sentences (e.g., grammar, spelling). Instead, feedback on the *macro-level* refers to the text as a whole and provides feedback on the overall context of the text (e.g., argumentative structure, rhetorical moves) (Strobl et al., 2019, p. 33). In their analysis, Strobl et al. (2019, p. 44) conclude that writing tools that provide *macro-level* feedback are still rare.

Although the literature review by Strobl et al. (2019) is very valuable, it looks at writing tools in secondary and higher education together, even though the writing needs for the two audiences are different in many ways. For example, the structure of scientific papers differs from that of other text forms, which are more commonly used in secondary education (e.g., narrative text types, poetry). Additionally, citation according to academic standards plays a more important role in higher education. Students often have to write their first academic paper without knowing academic writing requirements (Kruse & Chitez, 2014). For this reason, it might be meaningful to look at writing tools intended for use in higher education. Due to the different target groups, these tools may also have different features. Regarding the review of Strobl et al. (2019), a more detailed assessment of the feedback features offered by the different writing tools could also be conducted. This assessment would provide a more accurate overview of the current capabilities of these writing tools. Moreover, since NLP research has made great progress in recent years, it could well be that writing tools have improved significantly since 2019. Therefore, a revisited analysis in the more specific context of *higher education* might be helpful. Accordingly, the objective of the paper at hand is to describe and analyze *NLP-based writing tools used in higher education*, obtain a more nuanced picture, and reflect on the meaningful use and prospects of these tools. Moreover, in light of the identified research desideratum, the following research question should be addressed: *What are the state-of-the-art features of NLP-based writing tools used in higher education?*

From a theoretical point of view, the paper at hand contributes to a better understanding of the usefulness of *NLP-based writing tools in higher education*. From a practical standpoint, the presented features of *NLP-based writing tools* could help design use cases, as they may serve as a checklist and guideline to develop tools that go beyond what already exists.

To this end, section 2 provides more detailed information about the applied method and the conducted search procedure to identify *NLP-based writing tools used in higher education*. Section 3 describes the identified *writing tools* and their features. Section 4 discusses the findings critically and section 5 concludes with some final remarks.

2. Method

Based on the literature review by Strobl et al. (2019), the goal was to identify *NLP-based writing tools for higher education*. First, inclusion criteria were defined that all tools had to fulfill. The goal was to find state-of-the-art *writing tools* that are (1) still working today, (2) focus on higher education, and (3) provide NLP-based feedback on an inserted text (*micro-* or *macro-level*).

Figure 1 summarizes the review process. The 44 digital writing tools described in Strobl et al. (2019) review were taken as a starting point. Additional supplementary desk research revealed six further potentially interesting writing tools for higher education that had not already been included in the literature review by Strobl et al. (2019). In total, 50 potentially relevant writing tools were identified.

In a first step, an attempt was made to access each of the 50 tools to see if they still existed and could be used. Several tools had restricted access (e.g., behind a login wall), but many tools were no longer available. In total, 15 writing tools had to be removed because these writing tools did no longer exist or were at least no longer publicly available.

In a second step, the remaining 35 writing tools were assessed in more detail and examined concerning the defined inclusion criteria. In total, 21 writing tools had to be removed because they did not meet the above inclusion criteria. Particularly, writing tools were removed that did not provide

personalized NLP-based feedback (*micro-* or *macro-level*) but supported the writing process primarily with the help of scaffolding elements (see, e.g., EssayJack, n.d.; Rapp & Kauf, 2018).

In total, 14 writing tools remained for further fine-grained analysis. These tools will now be described in detail in the next section.

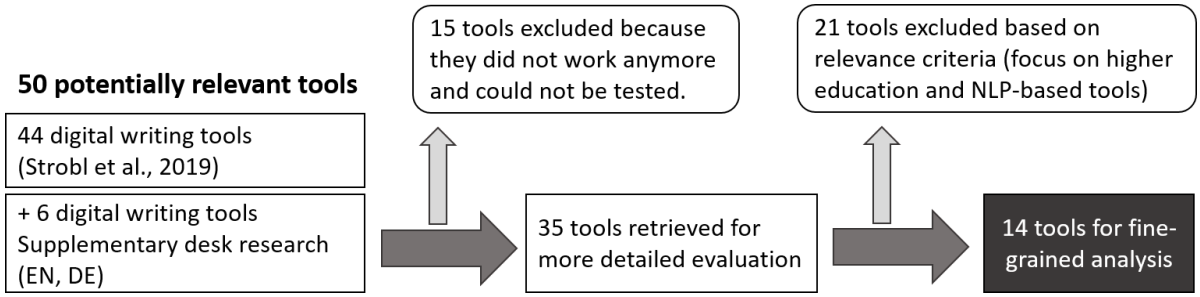


Figure 1. Review process.

3. Results: State-of-the-Art Analysis

Table 1 gives an overview of the 14 identified writing tools in higher education. The majority of the tools are freely accessible, intended to support English writing and can be tested free of charge. It is necessary for many of the tools to create a user account. The tools *AcaWriter* (1), *ArgueLearn* (2), and *CorpuScript* (3) are only accessible to students of the particular university. Through the information provided on the respective website, these tools nevertheless allow a transparent insight into their functionality and features.

Table 1. *NLP-based tools supporting student writing in higher education.*

Tool	Name	Link
(1)	AcaWriter	https://acawriter.uts.edu.au/
(2)	ArgueLearn	https://iwi.unisg.ch/en/projects/argulearn/
(3)	CorpuScript	https://www.schrijven.ugent.be/corpuscript
(4)	CorrectEnglish (Vantage Learning)	https://www.correctenglish.com/
(5)	Criterion	https://www.ets.org/criterion/about/
(6)	De-Jargonizer	http://scienceandpublic.com/
(7)	Enago	https://www.enago.com/researcher-hub/
(8)	Grammarly	https://app.grammarly.com/
(9)	Marking Mate (standard version)	http://writingtools.xjtlu.edu.cn:8080 /mm/markingmate.html
(10)	PaperRater	https://www.paperrater.com/
(11)	Scientific Writing Assistant (SWAN)	https://cs.joensuu.fi/swan/
(12)	Trinka	https://www.trinka.ai/
(13)	Turnitin (Revision Assistant)	https://www.turnitin.com/
(14)	Writefull	https://www.writefull.com/writefull-for-word

Most of the tools are web-based applications. However, some tools can only be used in the form of a Microsoft-Word-Plugin (*Writefull* (14)) or must be downloaded and installed like regular software (*Scientific Writing Assistant* (11)). The various writing tools offer different services, but they also have many things in common. Table 2 provides a summary overview of the different forms of feedback offered by these tools.

Table 2. *Features of NLP-based tools supporting student writing in higher education.*

Tool	Micro-level	Macro-level	Additional features
(1)	Writing suggestions (e.g., sentence	Overall reflective report and feedback with individualized advice	-

	length, style)		
(2)	-	Feedback on readability, coherence, and persuasiveness; feedback on the structure of arguments (claims and premises)	-
(3)	-	Check of word count; auto-generated suggestions for further reading	Plagiarism check
(4)	Spelling; word suggestions	-	Format documents for MLA or APA citation standards; plagiarism check
(5)	Diagnostic feedback (e.g., grammar, style)	-	-
(6)	Highlights jargon in writing	Suitability for general audience score	-
(7)	Grammar	-	Plagiarism check
(8)	Spelling; grammar; punctuation	Checks on conciseness, readability, inclusive language, fluency, level of formality	Plagiarism check
(9)	Grammar; punctuation	Academic style rating; overall score rating with individualized advice	Reference check
(10)	Grammar; writing suggestions	-	Plagiarism check
(11)	-	Evaluates the structure of a scientific paper; gives individualized advice on different sections (e.g., abstract)	-
(12)	Grammar	Overall consistency check	Publication readiness checks (e.g., journal scope match, technical and ethical compliance)
(13)	Grammar; writing suggestions	-	Plagiarism check; similarity check; reference and citation check
(14)	Language feedback for the scientific content	Exploration of writing patterns	Title generator; paraphraser; academic phrase bank

As shown in Table 2, almost all tools provide some forms of *micro-level feedback*. This form often involves general feedback on grammar and spelling combined with alternative wording suggestions. Moreover, many tools also provide *macro-level feedback* that tries to assess or relate to the overall quality of a text. Some types of *macro-level feedback* may be quite simple such as the overall check of the allowed word count (*CorpuScript* (3)). Other types of *macro-level feedback* are more sophisticated and provide different forms of overall score ratings. For example, *De-Jargonizer* (6) highlights the words in a text according to their frequency and calculates how suitable a text is for the general audience. Based on the feedback, text passages with rarely used words (that may be difficult to understand for the general audience) can be adjusted. *AcaWriter* (1), *Grammarly* (8), and *Marking Mate* (9) provide the user with a detailed overall report and individualized advice on how to improve the text further. The overall reports usually comprise several subcategories, whose recommendations can be followed in detail. *Marking Mate* (9) has a subcategory called *academic style rating* that checks for typical deficiencies such as the use of informal expressions or colloquial words. *Grammarly* (8) gives the user feedback on different aspects of the overall quality of the text, such as the perceived fluency or the level of formality.

The *Scientific Writing Assistant* (11) can even provide detailed feedback on the text structure. For example, the *Scientific Writing Assistant* (11) can point out that certain elements that are considered important in the text structure are not mentioned in the abstract; or that certain parts of the paper (e.g., literature background) are probably too short or too long, relative to the rest of the text. However, to precisely establish and analyze these connections, the individual text elements (abstract, title, introduction, conclusion, etc.) must be entered separately, which makes the handling somewhat more complicated. *ArgueLearn* (2) also addresses the overall structure of a text and checks how well the arguments are constructed. For this purpose, *ArgueLearn* (2) identifies and marks claims in the text along with any premises that support a particular claim. In this way, the user can review the arguments made and improve them if necessary.

In addition to *micro- and macro-level* feedback, most tools incorporate further *additional features* that may support the user in its writing. Often, this includes plagiarism and citation standard checks (see Table 2). *Writefull* (14) includes an academic phrase bank (where typical sentences used in the academic context can be used as a template), a paraphraser, and a title-generator feature. Trinka (13) can run several publication readiness checks, such as a journal scope match and technical and ethical compliance checks.

In addition to the free versions, there often exist additional paid subscription plans for many writing programs (*CorrectEnglish* (4), *Enago* (7), *Grammarly* (8), *PaperRater* (10), Trinka (12), *Writefull* (14)). These paid subscription plans promise enhanced features (e.g., unlimited usage, faster processing, ad-free, improved *macro-level* feedback, access to writing templates and external databases, publication readiness, and plagiarism checks) and usually cost between 8 and 24 USD per month. The length of the subscription (month or year) often affects the price.

Concerning Table 2, it is noteworthy that *micro-level* feedback is generally offered free of charge, while *macro-level* feedback and other functions such as plagiarism checks often have to be paid for (at least after a certain level of usage). Considering that Microsoft Word (www.microsoft.com/word) or OpenOffice (www.openoffice.org) can already provide relatively good feedback at the *micro-level* (e.g., regarding grammar), it seems plausible that only the provision of "basis" *micro-level* feedback services generally are for free.

4. Discussion

The analysis in the last section has revealed that today several writing tools exist in higher education that can provide feedback not only at the *micro-* but also at the *macro-level*. This analysis contributes to a more nuanced picture, as Strobl et al. (2019) concluded that only a few tools provide *macro-level* feedback. However, the various tools often have a modular nature and are composed of different (*micro-* and *macro-level*) features that can add user value.

The analysis of existing *NLP-based writing tools in higher education* has also shown that in contrast to institutional providers (e.g., universities), many private providers often pursue commercial interests. In the sense of a freemium business model, basic services are usually offered free of charge, but additional improved features have to be activated via paid subscription plans.

On the one side, this raises questions of fairness since not all students may afford to access improved writing tools, and their usage can be an advantage when writing texts. On the other side, however, the motivation of the commercial providers must also be questioned, who may be pursuing different interests with their writing tools than higher education institutions (who want to use writing tools primarily in a pedagogically meaningful way). Are commercial providers concerned about what writing tools are *best* for the user in terms of learning gains or rather about what is the *most convenient* for the user? As learners may not always know best (Kirschner & van Merriënboer, 2013), commercial providers could capitalize on this and encourage a dependency on their writing tools. New approaches may be needed to ensure that users have the competencies to use writing tools to their advantage without becoming dependent on them in the long run.

5. Conclusion

This paper systematically compared fourteen NLP-based writing tools used in higher education. The state-of-the-art analysis revealed that many writing tools provide feedback at the *micro-level* (e.g., grammar, spelling) and *macro-level* (e.g., argumentative structure, rhetorical moves). However, *macro-level* feedback and other enhanced features (e.g., plagiarism checks, access to writing templates) are often only included in paid subscription plans. Overall, the presented features of *NLP-based writing tools* could be helpful in designing own use cases as they may serve as a checklist and guideline to develop tools that go beyond what already exists. Future research in NLP-based writing tools should aim to develop applications that provide pedagogical value in multiple ways. Among other things, this could be achieved through the modular and simultaneous integration of different feedback elements at the *micro-* and *macro-level*, as they were presented here in this paper.

References

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). *Computer-Based Writing Instruction*. ERIC. https://eric.ed.gov/?id=ED586512

Cotos, E. (2015). Automated Writing Analysis for Writing Pedagogy: From healthy tension to tangible prospects. *Writing & Pedagogy, 7*(2-3). https://doi.org/10.1558/wap.v7i2-3.26381

Ding, F. (2019). *How to AI: Navigating the buzzwords of artificial intelligence*. Retresco. https://www.retresco.com/how-to-ai-natural-language-processing/

EssayJack (n.d.). *Wish academic writing was easier?*. EssayJack. https://www.essayjack.com/

Gu, M. (2020). *NLP Glossary for Beginners: Demystify Natural Language Processing with these terms*. Medium. https://medium.com/analytics-vidhya/nlp-glossary-for-beginners-c3093529ee4

Kirschner, P. A., & van Merriënboer, J. J. (2013). Do learners really know best? Urban legends in education. *Educational psychologist, 48*(3), 169-183. https://doi.org/10.1080/00461520.2013.804395

Kruse, O., & Chitez, M. (2014). Schreibkompetenz im Studium: Komponenten, Modelle, Messung. In Dreyfürst, S. & Sennewald, N. (Eds.). *Grundlagentexte zur Theorie, Didaktik und Beratung*, 107-126. Verlag Barbara Budrich.

Oberländer, M., Beinicke, A., & Bipp, T. (2020). Digital competencies: A review of the literature and applications in the workplace. *Computers & Education, 146*, 103752. https://doi.org/10.1016/j.compedu.2019.103752.

Rapp, C., & Kauf, P. (2018). Scaling Academic Writing Instruction: Evaluation of a Scaffolding Tool (Thesis Writer). *International Journal of Artificial Intelligence in Education, 28*, 590–615. https://doi.org/10.1007/s40593-017-0162-z

Seufert, S. & Spiroudis, E. (2017) Wissenschaftliche Schreibkompetenz in der Studieneingangsphase fördern: Erfahrungen mit einem Rückmelde- und Bewertungsraster (Rubric). *Zeitschrift Schreiben, 2*, 1–9.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007.

Strobl, C., Ailhaud, E., Benetos, K., Devitt, A., Kruse, O., Proske, A., & Rapp, C. (2019). Digital support for academic writing: A review of technologies and pedagogies. *Computers & Education, 131*, 33–48. https://doi.org/10.1016/j.compedu.2018.12.005

Thaiss, C., & Zawacki, T. M. (2006). *Engaged writers and dynamic disciplines. Research on the academic writing life*. Boynton/Cook.

Vajjala, S., Majumder, B., Gupta, A. & Surana, H. (2020). *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems* (1. Ed.). O'Reilly Media. https://www.oreilly.com/library/view/practical-natural-language/9781492054047/

van Laar, E., van Deursen, A.J. A. M, van Dijk, J. A. G. M., & de Haan, J. (2017). The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior, 72*, 577-588. https://doi.org/10.1016/j.chb.2017.03.010

Zhang, M. (2013). Contrasting automated and human scoring of essays. *Educational Testing Service R&D Connections,21*, 1–11. https://originwww.ets.org/Media/Research/pdf/RD_Connections_21.pdf

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Investigating Bloom's Cognitive Skills in Foundation and Advanced Programming Courses from Students' Discussions

Joel J. W. LIM, Swapna GOTTIPATI* & Kyong Jin SHIM
School of Computing and Information Systems, Singapore Management University, Singapore
joel.lim.2018, *swapnag,kjshim@smu.edu.sg

Abstract: Programming courses provide students with the skills to develop complex business applications. Teaching and learning programming is challenging, and collaborative learning is proposed to help with this challenge. Online discussion forums promote networking with other learners such that they can build knowledge collaboratively. It aids students open their horizons of thought processes to acquire cognitive skills. Cognitive analysis of discussion is critical to understand students' learning process. In this paper, we propose Bloom's taxonomy based cognitive model for programming discussion forums. We present machine learning (ML) based solution to extract students' cognitive skills. Our evaluations on compupting courses show that ensemble model performs better with an average F1-score of 76%.

Keywords: Discussion forum, bloom's taxonomy, cognitive skills, programming posts, machine learning

1. Introduction

Learning to write computer programs has never been an easy task largely due to abstraction, critical thinking, analytical abilities, etc. (Bosse & Gerosa, 2016). Foundational programming courses teach students how to develop console applications, web applications, and other applications with simple features. Whereas advanced programming courses such as Object Oriented Programming and Advanced Web Application Development, train students to develop more complex and robust software applications. Discussion forums for programming courses are part of collaborative learning that provide an interactive learning environment for cognitive participation in the form of questions and answers.

Cognitive analysis of discussions is critical to understanding students' learning processes (Garrison, 2003). Six levels of Bloom's taxonomy (Bloom, 1956) are widely used for evaluating the cognitive levels of learner's knowledge. The cognitive skills analysis in programming courses is not similar to non-programming courses due to the nature of posts and learning objectives (Cabo, 2015). To support cognitive analysis for programming courses, Fuller (2007) developed a Computer Science specific learning taxonomy. Analysing programming posts aid instructor to prepare effective teaching methods to improve student's performance (Dunlosky, 2013).

Manually analyzing discussions to understand the cognitive skills of students in programming courses is a tedious job. Several researchers have proposed automated content analyses of discussions using machine learning techniques (Ntourmas, 2019; SHAH, 2021). In this paper, we take a similar approach. First of all, we define Bloom's taxonomy based cognitive model on discussion posts for programming courses. Secondly, we develop a solution model based on machine learning to analyze the content. In particular, each post is classified into a Bloom's cognitive level using feature modelling and classification algorithms. Finally, we apply the solution model to the foundational and advanced programming courses to extract and compare students' cognitive skills and interactions. One key contribution of this research work is the framework that aligns Bloom's cognitive skills to programming discussion forums. Second contribution is ML based solution design for classifying discussion posts.

2. Related Work

Discussion forums are special platforms where students exhibit both cognitive and social skills. Schellens (2005) proposed a social media cognitive taxonomy based on the users' behaviour in discussion forums. It is based on task-related contributions and cognitive dimensions, and they are categorized as new information, explicitation and evaluation. Since these categories are limited and not very useful for several disciplines, several researchers used the idea of applying modified Bloom's taxonomy to discussion forums. For example, Stump et al. (2013) proposed modified Bloom's framework to analyze the cognitive skills of students in MOOC forums specific to the Engineering discipline - Circuits and Electronics. Wong et al. (2015) proposed a cognitive classification framework based on Bloom's taxonomy for discussion forums and applied it to a MOOC forum related to art discipline - Introduction to Art. Automated classification of forum posts can be achieved with machine learning models (Ntourmas, 2019, Shah, 2021, & Khodeir, 2021). In this paper, we propose a Bloom's taxonomy-based framework for discussion forums related to programming courses. To automatically classify the posts, we adopt machine learning techniques used by previous works. We further analyze posts and apply visual models for pattern discovery and comparison analysis.

3. Framework for Cognitive Analysis of Programming Discussions

Previous studies (Kovanović et al. 2016) have pointed to utilizing Bloom's taxonomy to evaluate students' understanding of topics in online discussion forums. Based on these studies, Wong et al. (2015) proposed a cognitive classification framework for discussion forums. We integrated both works and Table 1 shows our proposed framework to classify the programming discussion posts.

Table 1. *Cognitive classification of posts in programming discussion forums*

Label	Non-programming course forums	Programming course forums	Bloom Levels
0	Message that is irrelevant to the concepts such as acknowledgement or greeting, etc.	A post that is irrelevant to the concepts such as acknowledgement or greeting, etc.	Non-Cognitive
1	Message must translate, construe, interpret, or extrapolate information. Message must describe, list, or name factual information.	A question post with no detailed explanation. A reply post with the direct answer with no explanation or pointer to another post.	Understand & remember
2	Message must exploit information and put the resulting knowledge into action	A reply to the question using only theory without explanation but provides links to resources. A post with questions related to concepts and with clear explanations.	Apply
3	Message must appraise or relate information to the real world	A reply to the question with an attempt to explain, paraphrase and contextualize the concepts.	Evaluate
4	Message must deduce, scrutinize, or survey information	A reply that evaluates the pros and cons of the answer or solution in context to the question.	Analyze
5	Message must formulate, generate, restructure, or combine information.	A post that summarizes various concepts discussed with respect to question or related to course topic.	Create

In our preliminary data analysis, we observed non-programming related posts and these posts are still useful to understand the interaction levels of the students in a course. Therefore, we propose Level-0 to capture such posts. We also combined two levels, understand, and remember as, unlike programming assessments, in the context of discussion forums, it is difficult to distinguish both the

levels. To explain this challenge, we use example posts as shown in Table 2. For level 5, we used the example from developer forums since it is more detailed.

Table 2. *Cognitive levels and example posts*

Label	Bloom's Level	Programming course forums
0	Non-Cognitive	<@U0193N90XN0> ikr :joy: Post 1- Happy Birthday <@U017X1BHQP6>!:wink: Post 2- got it!\ Post 3- now in school with XXX practicing WAD
1	Understand & remember	Post 1- hi! does anyone know how to prevent the underline just for 'j'? Post 2- Perhaps you could try not including float for #side-bar
2	Apply	hi anyone know why such behaviour is displaying? I am currently using font-size:150% for the button which by right should display 1.5x bigger than the rest of the font and it looks fine when rendered as a website but when rendered in mobile version, the font shrinks and looks smaller in comparison to the rest of the text (Web Vs Mobile-Iphone 5)
3	Evaluate	you could try making the larger image visible only when the size is >md and making it hidden once size <md. likewise you can make the smaller image visible only when size is <md and making it hidden once size is more than md. you can refer to <https://getbootstrap.com/docs/4.0/utilities/display/> under hiding elements

4	Analyze	Perhaps, you can try adding an id to the and add in the style :
		li-design {
		text-decoration: none;
		border-bottom: dotted;
		}
		<li id = "li-design">Handsome
		So that the entire will be displayed with that specific style.
		Hope this helps!

5	Create	Xcode supports source code for the programming languages **C,** **C++**, **Objective-C**, **Objective-C**++, **Java**, **AppleScript**, **Python**, **Ruby**, ResEdit (Rez), and **Swift**, with a variety of programming models, including but not limited to **Cocoa**, Carbon, and **Java**. Third parties have added support for GNU Pascal,[5] Free Pascal,[6] Ada,[7] C#,[8] Perl,[9] and D.[10] Source : h ttps://en.wikipedia.org/wiki/Xcode Which one would you want to add ? I've never ventured in those areas. But there are XCode extensions. https://developer.apple.com/documentation/xcodekit/creating_a_source_editor_extension And a few implementation tutorials https://www.vadimbulavin.com/xcode-source-editor-extension-tutorial/ https://code.tutsplus.com/tutorials/how-to-create-an-xcode-source-editor-extension--cms-26772

4. Solution Design

The solution design overview consists of two key stages: Data cleaning and, Post classification.

Data Cleaning: Instead of the traditional discussion forums in Learning Management Systems the faculty used Slack which is a popular messaging application. It was originally built for businesses. However, by March 2020, over 3,000 colleges and universities had adopted Slack in their classrooms (Slack, 2020). We first parsed JSON files from Slack into an excel workbook and then extracted specific fields; student message, reply Message, Replying target, Discussion ID, Links, date, and time of post. As the first step of normalizing the post, we tokenized it, converted it into lowercase, removed emojis and removed stop words using NLP library and regex (Kulkarni, 2019). We removed posts with a single word. We consider both textual contextualized information and post characteristics such as length, to classify posts into cognitive categories.

Discussion Post Classification: The second stage of the solution design is categorizing the posts into Bloom's cognitive levels. We first generate the post features using Linguistic Inquiry and Word Count (LIWC) (Tausczik, 2010) which provides seven categories in relation to cognitive ability. The post types range from a set of simple language to complex words. Dale-Chall readability score provides the measure of comprehension difficulty when reading text (McClure, 1987). We then train and test the

classifiers on the post and choose the classifier with the best performance, in this case, F1-score. The key features used in training the classifier are post tokens, LIWC cognitive dimensions and Dale–Chall readability score. This study explores the use of BERT (Devlin et al. 2019), AdaBoost (Schapire, 2013) and XGBoost (Chen & Guestrin, 2016) algorithms to build the classifier. We also use the ensemble method to ensemble all three algorithms.

5. Experiments and Results

In this section, we first describe the data preparation followed by the results and analysis of the solution model on discussion forums from foundation and advanced level programming courses.

5.1 Background of the Courses

In our school, all students must complete two web application development courses. Web Application Development course (WAD I) is a foundation course that equips students with the knowledge and skills to develop database-driven web applications using HTML, PHP, and MySQL. Upon successful completion of WAD, I, the advanced course, Web Application Development course (WAD II) teaches students how to develop well-styled and responsive web applications that provide rich user experiences using HTML, CSS, Bootstrap, JavaScript, and Vue.js.

5.2 Cognitive Levels Classification

The team collected 816 posts from both the courses, and we labeled random 773 posts (300 for training and 443 posts for testing). For labeling, the members used rule-based approach to assign cognitive levels to each post (Anderson et al. 2005) and we took the common label. The team used cognitive framework described in Table 1. Three coders were trained on how to label the discussion posts. Table 3 shows F1-score results of four classification models and Ensemble produced best performance.

Table 3. *Classifier performance - F1-Score comparison on testing data*

Course	BERT	XGBoost	AdaBoost	Ensemble
WAD I	0.37	0.66	0.61	**0.8**
WAD II	0.15	0.59	0.54	**0.72**

Most of the posts in Slack consist of short replies which create an impediment that prevents the BERT from fully contextualizing and associating words. The wide range of topics discussed in the Slack forum further dilutes word associations and creates further difficulty for the model to effectively pick keywords. As such the overall accuracy of the Slack models has fallen to 37% and 15% for WAD I and WAD II respectively. While the BERT is largely successful in its performance in multiple NLP-related tasks, its performance is significantly dependent on support from a large dataset.

For WAD I, the overall accuracy of the XGBoost and AdaBoost are 66% and 61% respectively. There is a notable 29% and 24% increase in accuracy from the XGBoost and AdaBoost compared to the BERT. For WAD II, the overall accuracy of the XGBoost and AdaBoost model are 59% and 54% respectively. There is a notable 44% and 39% increase in accuracy from the XGBoost and AdaBoost compared to BERT. The overall accuracy of the ensemble model for each forum dataset is higher than the attained accuracy from each individual three machine learning classifiers. Figure 1 shows the precision, recall and F1-score comparisons.

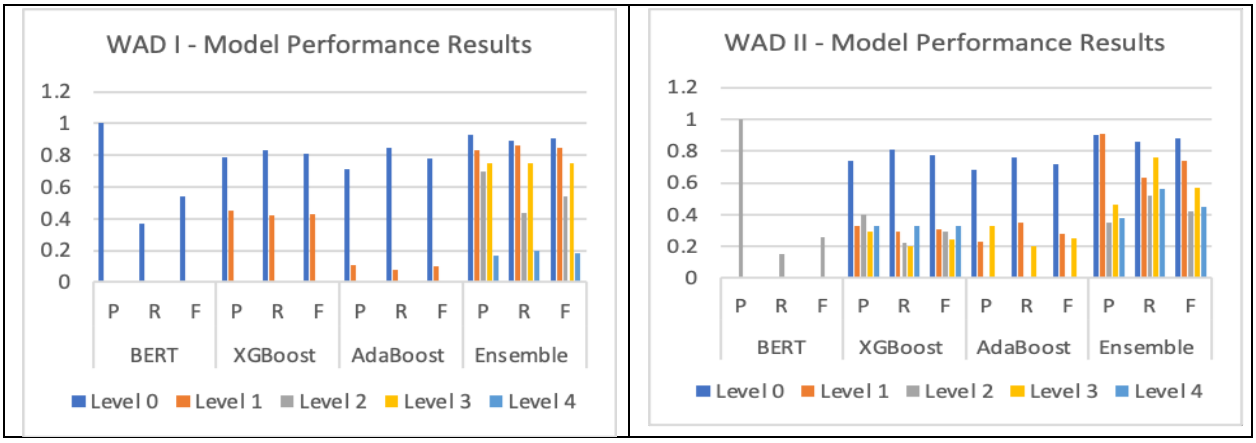


Figure 1. Precision, recall & F1-score comparison of the classifiers for each level.

From Figure 1, we observe that the ensemble model has performed better on the levels. We also notice that level 5 is not extracted. This is due to the smaller data sets for this level which we will explain next. In the final step of creating a cognitive classifier, we use ensemble models to label the remaining data. In total, we have 816 posts which we use for comparative analysis in the next section.

5.3 Comparison Analysis between Foundation and Advance Programming Courses

Figure 2 depicts the cognitive analysis among the participants in both the forums.

Level	WAD I	WAD II
0	156	242
1	98	177
2	19	40
3	25	42
4	6	11

(a) Comparison by number of posts

Level	WAD I	WAD II
0	**51.3%**	47.3%
1	32.2%	**34.6%**
2	6.3%	**7.8%**
3	8.2%	8.2%
4	2.0%	**2.1%**

(b) Comparison by percentage of posts

Figure 2. Cognitive analysis comparison on full dataset.

From Figure 2, we observe that although the students from advanced programming post extensively compared to the foundation level, the cognitive levels are statistics by proportion are very similar. We observe that most simple questions and answers (level 1) are provided for both courses. From the comparative analysis, we observe that the higher order thinking skills are not evident in the posts which agrees with the findings by Johnson & Fuller (2006). We have noticed that in developer forums, higher order cognitive levels are evident and this is due to the type of the questions. Therefore, the instructors should post questions that require higher order skills so that the students will try to answer and instructors can analyse the gaps in the knowledge.

5.4 Limitations

Our proposed solution worked well for the discussion forum within the chosen information systems programming courses. Our experiments show that ensemble model performs better than the three machine learning models. However, this may not be true for other courses. Therefore, we can improve the solution model to consider the code related keywords to improve the performance of the models and to generalize across all programming courses. The second limitation is the discussion forum settings. We used Slack which provides data extraction in JSON format which outputs more clean data. This may not be true for other discussion forums and the extraction process may end up with noise in the data. To handle the noise (e.g. HTML tags), the researchers need to adopt extensive cleaning techniques. Further, the data used in this study is a limitation for the model performance for higher order cognitive skills classification. To overcome this, we can enhance the data by using the programming posts from Stack Overflow which is a popular forum among software developers. An interesting future work is to align the cognitive levels to the teaching concepts. The summarization based on the teaching concepts

and cognitive statistics enables the teaching faculty to understand which programming concepts the students are finding difficult. We are currently working on the cognitive level summaries and student-level personalized journey dashboard that can be shared with the students.

6. Conclusion

Discussion forums play a key role in collaborative learning and aid in students' cognitive skills development. This paper presents a Bloom's taxonomy-based framework for discussion forums related to programming courses. Our solution design adopts existing machine learning techniques to automatically classify discussion forum posts. We evaluated the solution model on discussion posts of foundation and advanced programming courses from the computing curriculum. We analysed students' cognitive levels for both courses. The main findings show that students post both cognitive and non-cognitive messages, and most of the posts are at the lower cognitive level.

Acknowledgements

We would like to thank all the students who participated in this research.

References

Anderson, L. W. (2005). Objectives evaluation and the improvement of education. *Studies in Educational Evaluation*, 31(2): 102-113.Availableat: https://doi.org/10.1016/j.stueduc.2005.05.004

Bloom, B. S. Taxonomy of Educational Objectives: The Classification of Educational Goals. New York: Longmans, Green, 1956. Print.

Bosse, Y. & Gerosa, M. A. (2016). Why is programming so difficult to learn? Patterns of Difficulties Related to Programming Learning. *ACM SIGSOFT Software Engineering Notes*. 41(6).

Cabo, C. (2015). Quantifying student progress through Bloom's taxonomy cognitive categories in computer programming courses. *ASEE*. 122. 10.18260/p.24632.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv, abs/1810.04805.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. Psychological Science in the Public interest, 14(1), 4-58.

Fuller, U. & Johnson, C. & Ahoniemi, T. & Cukierman, D., Hernán-Losada, I., Jacková, J., Lahtinen, E., Lewis, T., Thompson, D., Riedesel, C. & Thompson, E. (2007). Developing a computer science-specific learning taxonomy. *ACM SIGCSE Bulletin*. 39. 152-170. 10.1145/1345375.1345438.

Garrison, D. (2003). Cognitive presence for effective asynchronous online learning: The role of reflective inquiry, self-direction and metacognition. *Elements of Quality Online Education: Practice and Direction*. 4.

Johnson, C. & Fuller, U. (2006). Is Bloom's taxonomy appropriate for computer science?, *presented at Baltic Sea Conference on Computing Education Research: Koli Calling*.

Khodeir, N. A. (2021) Bi-GRU Urgent Classification for MOOC Discussion Forums Based on BERT. *IEEE Access*, vol. 9, pp. 58243-58255, 2021, doi: 10.1109/ACCESS.2021.3072734.

Kulkarni, A., & Shivananda, A. (2019). Natural language processing recipes. Apress.

McClure, G. (1987). "Readability formulas: Useful or useless. (an interview with J. Peter Kincaid.)". IEEE Transactions on Professional Communication. 30: 12–15. S2CID 13157772.

Ntourmas, A., Avouris, N., Daskalaki, S., & Dimitriadis, Y. (2019) Comparative Study of Two Different Mooc Forums Posts Classifiers: Analysis and Generalizability Issues. *IISA* 1-8, doi: 10.1109/IISA.2019.8900682.

Schellens, T. & Valcke, M. (2005) Collaborative learning in asynchronous discussion groups: What about the impact on cognitive processing? *Computers in Human Behavior*, 21(6), 957-975.

Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37–52). Springer.

Shah, U., Rambhia, R., Kothari, P., Ramesh, R., & Banerjee, G. (2021) Automatic Classification of MOOC Forum Messages to Measure the Quality of Peer Interaction. *ICCE, 2021*.

Shim, K. J., Gottipati, S., & Lau, Y. M. (2021). Flip & slack – Active flipped classroom learning with collaborative slack interactions. *ICCE 2021*. November 22-26. 121-131.

Slack. (2020, March 16). Distance learning thrives in Slack. https://slack.com/blog/collaboration/distance-learning-in-slack

Stump, G. S., DeBoer, J., Whittinghill, J., & Breslow, L. (2013, December). Development of a framework to classify MOOC discussion forum posts: Methodology and challenges. *In NIPS Workshop on Data Driven Education,* 1-20.

Tausczik, Y. R. & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*. 29, 24–54.

Wong, J.-S. & Pursel, B. & Divinsky, A. & Jansen, J. (2015). Analyzing MOOC discussion forum messages to identify cognitive learning information exchanges. *Proceedings of the Association for Information Science and Technology*. 52. 1-10. 10.1002/pra2.2015.145052010023.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

I will help you, but will you help me? How the Perception of a Teachable Agent May Influence Performance

Eva-Maria TERNBLAD [a*] Magnus HAAKE [a] & Betty TÄRNING [a]
[a] *Dept of Cognitive Science, Lund University, Sweden*
* eva-maria.ternblad@lucs.lu.se

Abstract: Learning by teaching someone else has proven to be beneficial in both human-human and human-agent interaction. Instructing someone else, that is, taking the role as a tutor, has a series of positive effects on students' learning and performance. For example, the fact that someone else is in need of help seems to affect students to put more effort into the task at hand, trying harder, being more thorough and persisting longer. This has inspired researchers to design educational software that uses teachable agents (TAs). The use of such software has shown to improve students' – not the least lower-achieving students' – learning. Designing teachable agents is, however, a delicate matter since the personality and capability of the agent may affect students' behaviors and performance. This study, which includes data from 156 6th grade students who used an educational game in history during three lessons, contributes to research on how students' perceptions of a neutral TA (without any specific personality and behavior) may influence their performance. The focus is on whether and how students' learning gains are influenced by the extent to which students perceive their TA as someone who does need their help. In addition, we explored whether such potential effects would differ between lower- and higher-achieving students. Results were that students' perceptions about the TA's need for help was a significant predictor of high performance (based on in-game performance and post-test scores) – independent of their general achievement level (grounded in the students' reading proficiency). In other words, all students, whether lower- or higher-achieving, benefitted from being convinced of the agent's need for help. This finding is somewhat different from previous studies, where TAs mainly have been found to be beneficial for lower-achieving students. In relation to this, the present study adds a novel piece of information by suggesting that for TAs to be beneficial in educational software, their need for help should be clearly communicated and emphasized. This, on the other hand, may make them useful for all students, independently of how well they usually perform in school.

Keywords: Teachable agents, protégé effect, educational games, learning by teaching, achievement level

1. Introduction

Pedagogical agents are thriving in a broad range of different pedagogical contexts, in roles as experts, motivators, or mentors (e.g., see Kim & Baylor, 2016). In this paper we will focus on pedagogical agents that themselves can learn, referred to as *teachable agents* or *TAs* (Chase, Chin, Oppezzo & Schwartz, 2009). In the context of digital educational applications, a TA exploits the pedagogical tradition of 'learning by teaching' in that it takes the role as a tutee and learns from the student, while the student him- or herself learns from instructing the TA. Benefits from educational software using TAs have been shown in a number of studies (Biswas, Leelawong, Schwartz, Vye, & The Teachable Agents Group at Vanderbilt, 2005; Chase et al., 2009; Okita & Schwartz, 2013; Sjöden & Gulz, 2015; Tärning & Silvervarg, 2019). One of several documented benefits is the so called *protégé effect* (Chase et al, 2009), i.e., that students put more effort into their own learning (measured as spending more time on tasks, preparing in more detail, etc.) when they learn in order to instruct someone else, compared to when they learn for themselves. As of today, the large bulk of studies on TAs targets learning benefits from using different types of TAs. In contrast, the aim of this study is to investigate how a more neutral

TA is perceived (as a sibling, a classmate, or as an assistant) and whether a student's perception of the TA as someone in much or little need of help affects performance and learning.

1.1 Design features of teachable agents

Even though a TA's role is always that of a tutee, the character can be portrayed with different personalities, and several prior studies have examined the effectiveness of certain agent design features on specific learning outcomes. Ceha, Lee, Nilsen, Goh and Law (2021) for example found that a TA that tells jokes and funny stories for the amusement of others can increase motivation and effort. Further, Silvervarg and Månsson (2018) found that the way a teachable agent is introduced can affect how it is perceived, which, in turn, can influence students' perseverance. In their study, two groups of students interacted with a TA in an educational game. One group only received the introduction to the agent that was built into the software, whereas the other group also received an additional verbal introduction from the research leaders. Students in the latter group were more inclined to perceive the agent as someone with a desire to learn. They also gave higher ratings in self-reports on their experiences of putting in an effort and not giving up. Other studies by Tärning, Gulz, and Haake (2017) and Tärning, Silvervarg, Gulz, and Haake (2019) have shown that a TA portrayed as having low self-efficacy in a specific domain can increase students' performance and self-efficacy in that domain, and especially so for students that themselves have low self-efficacy. The authors here propose that a TA with low self-efficacy might come across as someone in more need of help than a TA with high self-efficacy.

In a similar study, Silvervarg, Kirkegaard, Nirme, Haake and Gulz (2014) investigated how a challenging TA was perceived among students with higher and lower self-efficacy (measured by self-reports on the students' teaching abilities). The researchers found that, in general, students were inclined to go along with the TAs challenging proposals rather than holding on to their own initial response, both when their initial response was correct but to a higher extent when it was incorrect. They also found that students with lower self-efficacy experienced that the TA challenged them too often. This was not the case for students with higher self-efficacy.

Taken together, there is evidence that the way the TA is portrayed can have an impact on how students perceive it, and this also has a potential to affect their learning.

1.2 Helping a teachable agent

As mentioned above, several studies have investigated possible effects of TAs with respect to different roles or different personalities. For instance, Tärning and Silvervarg (2019), analyzed chat dialogues between students and their TA in an educational game, and found that TAs that expressed low self-efficacy with respect to their own ability to learn math received more positive and encouraging comments about their intelligence and competence. They also found that the students responded more frequently to the feedback from such a TA, as compared to a TA that expressed high self-efficacy. Similar to Tärning et al. (2017; 2019) the authors here argue that a TA with low self-efficacy might appear as someone in more need of help and therefore receives a better treatment compared to a TA with high self-efficacy (which might be perceived as someone that can take care of itself, without help).

Furthermore, Ogan et al. (2012) explored students' interaction with a TA (Stacy) in a math game (using the SimStudent platform) in terms of social or non-social conversation. Results were that the students who treated Stacy as a peer and for example spoke directly to her, were more successful in their learning task. This in contrast to students who talked to Stacy in a third person perspective and made fewer social comments, which was correlated with negative learning gains. Evidently, it appears that *how a student perceives their TA* can influence *how they treat it* and that a better treatment, for example in terms of being more polite, can be beneficial for the students' own learning.

Strong positive learning benefits from using TAs have so far been most pronounced in lower-performing students (Chase et al., 2009; Sjöden & Gulz, 2015; Pareto et al., 2009, Tärning et al., 2017; Tärning et al., 2019), and a number of explanatory mechanisms for this have been proposed. One possible explanation is that the TA always is inferior and more ignorant than the student and hence, the student has the opportunity to take the lead as someone more knowledgeable – a situation lower-performing students may not be familiar with in a school context. Seemingly, taking the lead as someone

competent, able to teach someone else, can improve student's self-efficacy. Hence, experiencing that someone else is in need of your help can be a potential incitement for learning.

However, students' individual opinions about a TA may also play a role. The present study therefore addresses possible variations in students' perception of a more neutral TA (i.e., a TA not designed to have any distinctive personality or behavior) and to which extent students perceive the TA as someone in need of help. Furthermore, we also investigate whether this, in turn, has an influence on their performance.

1.3 Research questions

Although the benefits and effects of TAs have been evaluated in several studies, students' attitudes towards their TA and the relation between these and students' learning and performance has been less studied. It has been shown that TAs are beneficial for lower-performing students *as a group* and likewise for students with lower self-efficacy *as a group*, but we know less about individual variability *within* such groups. The way a student perceives her agent may also tell us something about the student's self-awareness of her role as a teacher. Therefore, we are not only interested in students' characterization of their agent as such (for example as a sibling or a classmate), but also in their ideas about the agent's need for help. To our knowledge, no previous study has explored these issues. Consequently, we end up with the following research questions (RQs):

RQ.1: How do students characterize their (teachable) agent (when they can choose among a sibling, a classmate, and an assistant) and to what extent do they think that their agent is in need of help?

RQ.2: Will students with different achievement levels characterize their teachable agent differently?

RQ.3: Will student's opinion of their teachable agent's need of help influence their performance, and will such a potential impact differ between higher- and lower-achieving students?

2. Method

The research data in this study originates from a larger study on feedback engagement performed in spring 2019. A presentation of the participants and the experimental setup, together with a description of the material and a definition of the parameters relevant for this particular study, follows below.

2.1 The educational game

The stimulus in the study consists of an educational game in history where the students make time-travels to historical scenes and scientists, search for text-based information, and solve tasks (constructing a concept map, setting up a timeline, sorting statements into categories, or answering multiple-choice questions). To be able to continue and progress in the game, the tasks (six in total), must be completed one at a time with a sufficient degree of correctness (80%). The students have, however, unlimited attempts to revise the tasks, and they are allowed to make as many time-travels as they want.

The game contains two agents: Professor Chronos (a middle-aged wizard-like gentleman) and the time-elf Timy (an androgyn childlike character). Professor Chronos, who is the current 'Guardian of History', is about to retire and needs a successor. The educational narrative can then be set to one of two conditions. (i) *Learning for oneself;* The student's mission is to become Professor Chronos' successor and hence needs to prove his or her proficiency in history by making time-travels and solving tasks. In this condition, Timy has a role of a narrator. (ii) *Learning by teaching*; The time-elf Timy wants to become Professor Chronos' successor, but since he/she is suffering from time-travel sickness, the student has to make the time-travels and instruct Timy on how to solve the tasks. Timy then presents the solutions to Professor Chronos. In this condition, Timy has the role of a TA.

In the first condition, Timy in the role of a narrator has a less prominent role compared to the second condition where Timy has the role of a TA (Figure 1). In the first condition, Timy welcomes the

student to the game world and presents the game narrative without engaging in any dialogue with the student. After that, Timy's role in the first condition is restricted to introducing the tasks and instructing the student on how to solve them, e.g., dragging items to boxes or putting figures on a timeline. In the TA-condition, on the contrary, Timy also interacts with the student throughout the game by means of shorter dialogues. Figure 2 below shows Timy introducing the game (left) and having a dialogue with the student in his/her role as TA (right).

Timy as a narrator	Timy the time-elf introduces the game. No dialogues.	Timy presents the tasks.	The student presents the solutions of the tasks to Professor Chronos without Timy's presence. No dialogues.
Timy as a TA	*Timy the time-elf introduces the game as well as her/himself as a teachable agent. A shorter dialogue between the student and Timy.*	*Timy presents the tasks.*	*Timy is present during the tasks where the student instructs him/her. Timy then presents the solution to Professor Chronos. Some shorter dialogues between the student and Timy.*

Figure 1. Narrative of the game and the role of the time-elf in the two versions of the game – representing the two conditions in the study.

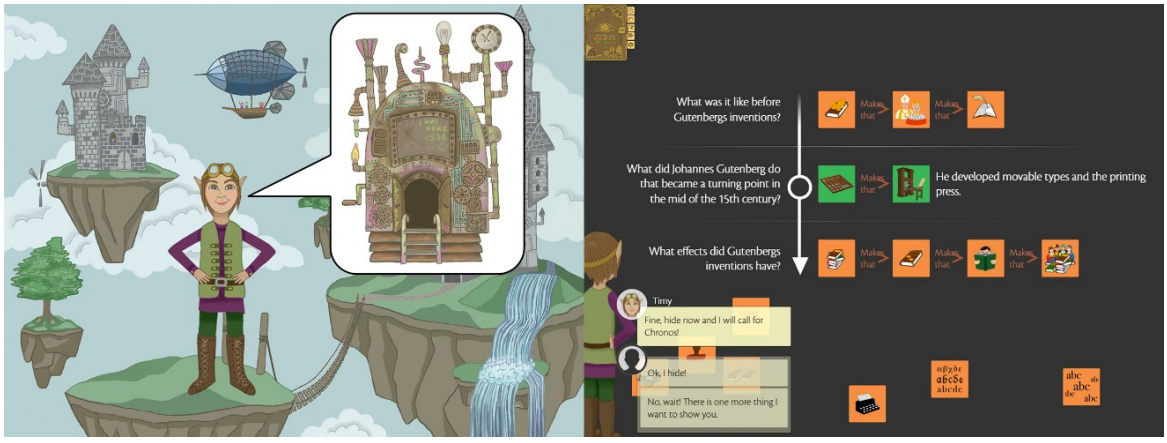


Figure 2. Left: Timy introducing the game narrative (both conditions); right: Timy having a dialogue with the student when being instructed on a task (TA-condition).

2.2 Participants, materials and procedure

In the original study, 285 Swedish 6th grade students in eleven classes from six different schools in southern Sweden participated (see Silvervarg, Wolf, Blair, Haake & Gulz, 2021). But, due to schedule difficulties and shortage of time, only 156 students were able to respond to our 'agent opinion' questionnaire (described below) and report on their personal experience of the game and how they perceived their TA. This data makes up the basis for this study.

The participants played the game during 3 separate sessions, each lasting approximately 60 minutes. During each session, two experimenters were on site, helping the students with questions and technical issues regarding the game. All students were randomly distributed to one of the two conditions ('Timy as TA' or 'Timy as narrator'), and each class was therefore divided into two groups (one for each condition), working in separate classrooms. At the first session, the participants were introduced to the game through a short film with Timy presenting the game narrative (see Figure 2, left).

Since the ability to process text-based information has a strong impact on a student's overall performance in the game, the teachers provided assessments (low, mid, high) of each student's reading proficiency. However, when observing the students solving their first task, it became clear for the researchers that the ability for the 'mid-achievers' to make use of the information in the game were quite diversified (since some of them solved the first task – a simple sorting task – quite easily, while

others failed repeatedly). This heterogeneity was not perceived within the other two groups. As a consequence, based on their initial in-game performance (more precisely: the number of revisions they needed for solving the very first task), the mid-achieving students were split up and categorized in either of the two other groups: higher- and lower-achieving students. The final distribution of the students in different conditions is presented in the Results section below.

All students played the game at their own pace. Some finished all tasks before the end of the third lesson and others did not complete all tasks. At the end of the last lesson the students performed a post-test, consisting of six multiple-choice questions (see Silvervarg et al., 2021). After that, the students were asked to complete the 'agent opinion' questionnaire (previously used in Silvervarg and Månsson (2018)), assessing their opinion of Timy by answering four statements: The time-elf… (i) is like a classmate, (ii) is like an assistant, (iii) is like a younger sibling, and (iv) needs my help. Each statement could be answered on a continuous bipolar scale ranging from 'completely disagree' to 'completely agree'. The last question, whether Timy needs help or not, relates to the protégé effect, i.e., that you put more effort into a task when teaching someone else compared to when learning for oneself (Chase et al., 2009).

The 'agent opinion' questionnaire was distributed to participants in both conditions ('Timy as TA' and 'Timy as narrator'), wherefore the results also serve as a validation of the intervention by providing information on the extent to which the time-elf was perceived as intended.

3. Results

The results are based on the dataset of 156 students described above. All of these students completed at least one historical mission in the game, responded to the post-test, and filled in the 'agent opinion' questionnaire. For the analyses, the students were distributed into four groups based on their achievement level (higher/lower, described in the method section above) and the experimental condition they were assigned to (Timy as Narrator; Timy as TA), see Table 1.

Table 1. *Distribution of students within different conditions*

Achievement Level	Experimental Condition	*N*
Lower	Timy as Narrator	37
Lower	Timy as TA	39
Higher	Timy as Narrator	37
Higher	Timy as TA	43
	Σ	156

In the following sections we first investigate the students' characterization of the time-elf (Timy), and how they evaluated Timy's need for help. After that comes an evaluation of the condition with 'Timy as a TA' (in comparison with 'Timy as a narrator'), and finally we investigate the relationship between the individual student's opinion about Timy's need for help and his or her performance.

3.1 RQ.1: The students' characterization of the time-elf Timy and their thoughts about Timy's need for help

The students' opinion of Timy, given by their ratings of the four statements (The time-elf… (i) is like a classmate, (ii) is like an assistant, (iii) is like a younger sibling, (iv) needs my help), was evaluated by measuring the participants' responses in millimeters along a 100 mm bipolar analogous scale, and then transforming this measure to a value between 0 and 1. The transformed results of the responses, presented in Table 2 below, reveal that the students' opinion of Timy differed significantly between the two conditions for all statements except 'classmate'. This is especially notable for the students' ratings on the statement "The time-elf needs my help" ('Timy as TA': $M = 0.68$, $SD = 0.31$, 'Timy as Narrator': $M = 0.38$, $SD = 0.33$), (Welch t-test: $t(150.5) = -5.88$, $p << .001$, $d = 0.95$)), where the means from the two conditions ended up on different sides of the 0.50 mid-value on the rating scale. This suggests that Timy's need for help worked on a conscious level and supports the notion that the students in the TA-condition were aware of the TA's role as a tutee and that they were supposed to help him/her.

The students' perception of the time-elf as an 'assistant' also differed significantly between conditions although both means were below the 0.5 mid-point, indicating that none of the groups considered this statement to be an appropriate description of Timy. When it comes to the other two categories: 'classmate' and 'younger sibling', there was no or only a small significant effect between conditions. Both statements also fell below the 0.5 mid-point; especially so the 'younger sibling' statement. Clearly, the time-elf Timy did not fit well with neither of these three statements.

Table 2. *The students' opinion, Mean (M; Range: 0–1) and Standard Deviation (SD), about the time-elf Timy for the two conditions × four statements (above) and Welch two sample t-tests between conditions for the four statements (below).*

Condition	***M* (*SD*) for Timy as…**			
	classmate	**assistant**	**younger sibling**	**in need of my help**
Timy as narrator	0.37 (0.27)	0.43 (0.28)	0.16 (0.24)	0.38 (0.33)
Timy as TA	0.32 (0.32)	0.26 (0.29)	0.28 (0.32)	0.68 (0.31)
Welch two sample *t*-test				
df	150.7	152.0	147.4	150.5
t-value	1.05	3.68	−2.55	−5.88
p-value[1]	0.30	< 0.001 ***	0.024 *	<<0.001 ***

[1] *p*-values Holm corrected for family-wise error rate.

3.2 The impact of the teachable agent on performance

To ensure that the TA had a significant effect on learning, we started out with a comparison of performance levels for different student groups and conditions (even though this is not articulated as a research question). This was done by classifying the students as 'good performing' or 'not-as-good performing' (depending on their results on tasks in the game and scores on the post-test) and setting up a logistic regression model to predict the probabilities for each one of the students to fall into the 'good performing' category. The following variables were used:

- **Good performing**: *GoodPerf* (binary dependent variable). Defined as having at least 50% correct on the post-test and completing at least three tasks; classified as 1 if the student passed this threshold, otherwise 0.
- **Student achievement level**: *AchLev* (categorical independent variable). Based on teacher assessments on reading proficiency an initial in-game performance. Two levels: higher and lower [Higher, Lower]
- **Time-elf condition**: *CndElf* (categorical independent variable). Two conditions: Timy as a TA and Timy as Narrator [TA, Narrator].
- **Interaction effects**: It was hypothesized that there might be an interaction effect between the two independent variables.

The following binomial logistic regression model predicting *GoodPerf* was then set up:

$$\text{logit}(GoodPerf) = \beta_0 + \beta_1 AchLev + \beta_2 CndElf + \beta_3 CndElf{:}AchLev$$

The best model fit was found by backward and forward stepwise elimination, revealing that all predictors had a significant contribution. This model performed significantly better than an intercept-only base line model ($\chi^2(3)$: 34.0, $p < .001$), having a reasonable fit (C-statistics: 0.77, Somers' D_{XY}: 0.54). The logistic regression analysis reveals that both the student achievement level (*AchLev*) and the interaction effect between *AchLev* and *CndElf* significantly predicted students as being 'good performing' (*GoodPerf*). See Table 3 for more details and statistics. The predicted probabilities for students within different conditions and achievement levels being 'good performing' are presented in Figure 3 below.

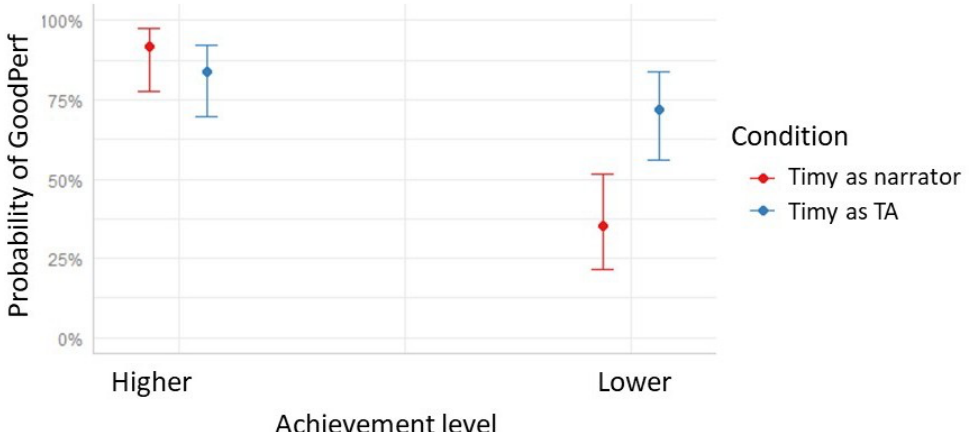


Figure 3. Probability of students being 'good-performing' (based on in-game performance and post-test scores) in relation to condition (*Timy as a TA* and *Timy as Narrator)* and achievement level (*Higher/Lower*).

Table 3. *Summary of the binomial logistic regression model fitted to predict students as 'Good Performing' (N = 156)*

Predictors	Coeff.	Odds.Ratio	Std. Err.	z-value	Pr (>|z|)	Sign.
Intercept	2.43	11.36	0.60	4.03	< 0.001	***
AchLev[Lower]	−3.04	0.05	0.69	−4.83	< 0.001	***
CndElf[TA]	−0.79	0.45	0.73	1.08	0.28	
AchLev[Lower]:*CndElf*[TA]	2.34	10.38	0.88	2.64	< 0.01	**

Model statistics. AIC: 161, C-value: 0.77, Somers' D_{XY}: 0.54, Likelihood ratio test: $\chi^2(3)$: 34.04, $p < .001$

Significance codes: . $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

3.3 RQ.2: Opinions about Timy as a TA for students with different achievement levels

Since the TA, in line with previous studies, had a significant impact on lower-achieving – but not higher-achieving – students (Silvervarg et al., 2021), we wanted to investigate if this could be related to differences in the students' opinions about their TA. A first evaluation was done by, for each student in the 'Timy as TA' group ($N = 82$), assigning Timy to one of the four statements in the 'agent opinion' questionnaire. That is, the statement that received the largest value was assumed to correspond to the individual student's 'best' description of Timy, as long as it was above the mid-point value of 0.5. Otherwise, Timy was classified as a 'NoType'. The results clearly suggest that within this condition, higher- and lower-achieving students have similar opinions about Timy (Chi-squared test: $\chi^2(9, N = 82) = 4.27$, $p = .37$).

Table 4. *Characterization of Timy within the group 'Timy as a TA' (N = 82) for higher- and lower achieving students*

Achievement Level	Tutee	Classmate	Assistant	Younger Sibling	No Type	*n*
Higher	27	5	1	2	8	43
Lower	19	5	5	1	9	39

Thereafter, we compared the two student groups with respect to the 'agent opinion' questionnaire measure for 'The time-elf [Timy] needs my help' (Higher: $M = 0.74$, $SD = 0.25$, Lower: $M = 0.63$, $SD = 0.36$). Here again there weren't any significant difference between the ratings from students with different achievement levels (Welch t-test: $t(67.0) = 1.54$, $p = 0.13$).

3.4 RQ.3: The impact of the students' opinion of the teachable agent's need of help on performance

To investigate Research Question 3: "*Will the student's opinion of the teachable agent's need of help influence their performance, and will such a potential impact differ between higher- and lower-achieving students?*", the students' ratings of 'Timy's need for help' (here referred to as *ElfHelp*) in the 'Timy as TA'-condition was used as a possible predictor in the following logistic regression model:

$$\text{logit}(GoodPerf) = \beta_0 + \beta_1 AchLev + \beta_2 ElfHelp + \beta_3 AchLev{:}ElfHelp$$

As described in the method section above, *ElfHelp* is a continuous independent variable, ranging between 0 and 1. The other two variables (*GoodPerf*[0, 1] and *AchLevf*[Higher, Lower]) have been described in section 3.2 above. The best model was fitted to the dataset by a backward and forward stepwise elimination and is presented in Table 5. Results show that only the variable *ElfHelp* contributed to the model and significantly predicted *GoodPerf* ($p < 0.01$) for the students (in the condition where Timy played the role as a TA). Neither achievement level (*AchLev*) nor the interaction effect had any significant contribution. The model is visualized in Figure 4 below:

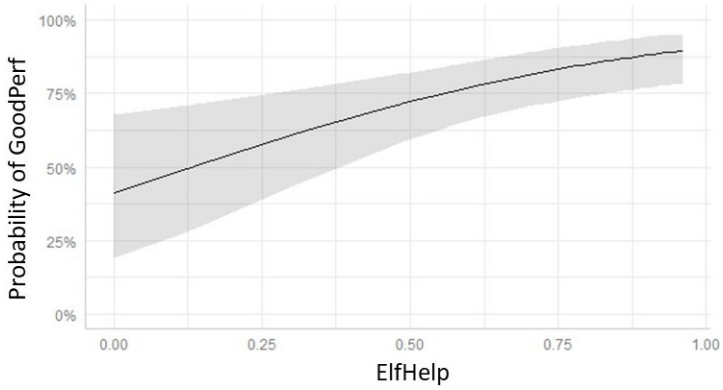


Figure 4. Probability of students in the 'Timy as TA'-condition being 'good-performing' (based on in-game performance and post-test scores) in relation to the measure *ElfHelp*.

Table 5. *Summary of the binomial logistic regression model GoodPerf ~ $\beta_0 + \beta_1 ElfHelp$ for students in the 'Timy as TA'-condition (N = 82).*

Predictors	**Coeff.**	**Odds.Ratio**	**Std. Err.**	**z-value**	**Pr (>|z|)**	**Sign.**
Intercept	−0.35	0.17	0.56	-0.62	0.54	ns
ElfHelp	2.61	13.60	0.85	3.07	< 0.01	**

Model statistics. AIC: 80, C-value: 0.72, Somers' D_{XY}: 0.44, Likelihood ratio test: $\chi^2(1)$: 10.17, $p < .01$

Significance codes: . $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4. Discussion

Even though a series of studies have investigated the effects of different TA personalities, students' opinions on the extent to which their TA needs their help is rarely examined or questioned. TAs are generally presumed to signal to the students to teach them, and since teaching someone else may have an effect on learning, this effect is assumed to relate to students' awareness of the TA's need of being instructed. In this study, we were interested in scrutinizing this relation and see if: (i) students' conception of an agent differ depending on whether the agent acts as a TA or not, (ii) students with different achievement levels think differently about their TAs' need for help, and (iii) the students' evaluation of the TA's need for help may have an impact on performance and learning – and if this relationship differed between higher- and lower-achieving students. In this pursuit, we ended up with

some interesting results, whereof some are in line with previous research and others are more unexpected.

First, characterizing a teachable agent – or perhaps any agent – is not straightforward. In our case, the students in the 'Timy as TA'-condition did not think of Timy as a classmate or a younger sibling, even though both of these characters can represent someone in need of help. This can be contrasted to the results in Silvervarg and Månsson (2018) where students in the experimental condition (who received additional verbal instructions about Timy's role as a tutee) were more inclined than students in the control group to characterize their TA as a classmate or a younger sibling. It might be, that by emphasizing Timy's needs of help in the beginning, the students came to care more about him/her, building a personal bond and thus also being more inclined to categorize him/her as a close relative or a friend. Interestingly, the students in the 'Timy as narrator'-condition were more inclined to rank Timy as an assistant (although the average value was below 0.5), even though s/he only introduced the game and did not provide any further help.

Our results clearly show that Timy's role in the game affected the students' opinions and thoughts about him, and the students in the condition where Timy acted as a TA also stated that he/she needed help. This was not the case if Timy only took the role as a narrator in the game. Noticeably, the students in the 'Timy as a TA'-condition were aware of their responsibilities and roles as teachers.

Not surprisingly, the condition 'Timy as TA' had a clear positive effect on performance and learning for students with lower achievement levels (assessed by combining teacher assessments on reading proficiency with initial in-game performance). This finding is in line with previous research (Chase et al., 2009; Sjöden & Gulz, 2015; Pareto et al., 2009; Tärning et al., 2017; Tärning et al., 2019). Figure 3 above also shows a minor tendency for higher-achieving students to perform less well when Timy was acting as their TA, but this effect is not significant. However, having a TA evens out the differences between lower- and higher-achieving students' learning gains.

More interestingly, the characterization of Timy and the self-report measure for Timy's need for help (within the 'Timy as TA'-condition) didn't differ significantly between students with different achievement levels. Even if our results reveal a tendency for lower-achieving students to report a slightly lower 'Timy is in need of help'-value than higher-achieving students, the difference between the groups is not significant. The reason for the different effects of the TA on learning for different student groups can thus not originate entirely from differences in opinions and thoughts about Timy as a TA.

Moreover, looking at ratings on the TAs need for help and relating them to the individual student's learning outcomes (measured by a combination of in-game performance and post-test-scores), it becomes clear that the only significant predictor of being a 'Good Performing' student is the student's thoughts about Timy's need for help (*ElfHelp*). Neither the achievement level (*AchLev*) nor the interaction effect between *AchLev* and *ElfHelp* was significant. This finding is new (and to our knowledge, not previously investigated). Since higher-achieving students often are less affected by TAs in general (Chase et al., 2009; Sjödén & Gulz, 2015; Pareto et al., 2009; Tärning et al., 2017; Tärning et al, 2019) we could assume that they also would be less affected by their personal opinions of them (in this case, how much help the time-elf Timy needs). However, this was not the case. *All* students, independently of achievement level, benefitted from perceiving the TA as someone in need of help. This could also be formulated as '*for a TA to have an effect on learning, it has to be designed in a way that the student takes its needs seriously'*. As a design consequence, the need for help ought to be clearly communicated and emphasized. In addition, taking into account the slight possibility that higher-achieving students may actually perform *worse* when working with a TA, these students also need to be convinced about the TA's needs.

Finally, the study has limitations. First of all, the ratings from the 'agent opinion' questionnaire were all gathered at the end of the experiment, after the gameplay. This means that the students might have felt differently about Timy when actually playing, and that these feelings might have varied during the game, perhaps due to various obstacles and difficulties. It may also be the case, that the self-report measure of Timy's need for help is an after-construction, and a sort of validation of the individual student's success. That is, students with stronger progression and result on the post-test, might have attributed the success to their responsibilities towards Timy and his need for help. Students reporting a lower value for Timy's need for help might, on the other hand, have 'blamed' Timy for their lack of success, avoiding taking full responsibility for their teaching (and learning).

The results of this study point to potential learning gains from designing teachable agents so that their need for help comes forth clearly. How this should be done is something to be investigated further. Should the agent, for example, specifically be asking for help and/or should more subtle behavioral patterns be exploited to communicate this? Further, the results imply that someone in need of help can trigger learning, but why is this so? Do students see the TA as a safety net in that the TA performs and takes tests, or would they prefer to play on their own, including taking the tests – and are such preferences similar or different for different groups of students?

Acknowledgements

The presented research has been funded by the Marianne and Marcus Wallenberg Foundation.

References

Biswas, G., Leelawong, K., Schwartz, D., Vye, N., & The Teachable Agents Group at Vanderbilt. (2005). Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, *19*(3- 4), 363-392.

Ceha, J., Lee, K. J., Nilsen, E., Goh, J., & Law, E. (2021. Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes? In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-14). New York, NY: ACM.

Chase, C. C., Chin, D. B., Oppezzo, M. A., & Schwartz, D. L. (2009). Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of science education and technology*, *18*(4), 334-352.

Kim, Y., & Baylor, A. L. (2016). Research-based design of pedagogical agent roles: A review, progress, and recommendations. *International Journal of Artificial Intelligence in Education*, *26*(1), 160-169.

Okita, S. Y., & Schwartz, D. L. (2013). Learning by teaching human pupils and teachable agents: The importance of recursive feedback. *Journal of the Learning Sciences*, *22*(3), 375-412.

Ogan, A., Finkelstein, S., Mayfield, E., D'Adamo, C., Matsuda, N., & Cassell, J. (2012), "Oh dear stacy!" social interaction, elaboration, and learning with teachable agents. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 39-48). New York, NY: ACM.

Pareto, L., Schwartz, D. L., & Svensson, L. (2009). Learning by guiding a teachable agent to play an educational game. In V. Demitrova, R. Mizoguchi, B. du Boulay, & A. C. Graeseer (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 662-664). IOS press.

Silvervarg, A., Kirkegaard, C., Nirme, J., Haake, M., & Gulz, A. (2014). Steps towards a challenging teachable agent. In T. Bickmore, S. Marsella, & C. Sidner (Eds.), *Lecture notes in computer science: Vol. 8637. Intelligent Virtual Agents* (pp. 410-419). Springer, Cham.

Silvervarg, A., & Månsson, K. (2018). How do you introduce an agent? The effect of introduction type on how a teachable agent is experienced by students. In A. Bogdanovych, D. Richards, & S. Simoff (Eds.), *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (pp. 29-34). New York, NY: ACM.

Silvervarg, A., Wolf, R., Blair, K. P., Haake, M., & Gulz, A. (2021). How teachable agents influence students' responses to critical constructive feedback. *Journal of Research on Technology in Education*, *53*(1), 67-88.

Sjödén, B., & Gulz, A. (2015). From learning companions to testing companions: Experience with a teachable agent motivates students' performance on summative tests. In C. Conati, N. Heffernan, A. Mitrovic, & M.F. Verdejo (Eds.), *Lecture notes in computer science: Vol. 9112. Artificial Intelligence in Education* (pp. 459-469). Berlin/Heidelberg, Germany: Springer-Verlag.

Tärning, B., Gulz, A., & Haake, M. (2017). Supporting low-performing students by manipulating self-efficacy in digital tutees. In G. Gunzelmann, A. Howes, T. Tenbrink, & E.J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1169-1174). Austin, Texas: Cognitive Science Society.

Tärning, B., Silvervarg, A., Gulz, A., & Haake, M. (2019). Instructing a teachable agent with low or high self-efficacy – does similarity attract? *International Journal of Artificial Intelligence in Education*, *29*(1), 89-121.

Tärning, B., & Silvervarg, A. (2019). "I Didn't Understand, I′ m Really Not Very Smart" – How Design of a Digital Tutee's Self-Efficacy Affects Conversation and Student Behavior in a Digital Math Game. *Education Sciences*, *9*(3), 197-218.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

‘I know that I clicked but not if I read’: An Exploratory Study Comparing Data Traces and Self-Reports on Feedback Engagement

Eva-Maria TERNBLAD[a*], Agneta GULZ[a,b] & Betty TÄRNING[a]
[a] *Dept of Cognitive Science, Lund University, Sweden*
[b] *Dept of Computer and Information Science, Linköping University, Sweden*
*eva-maria.ternblad@lucs.lu.se

Abstract: The use of educational digital tools – both within and outside the classroom – has opened up for novel ways to investigate and pursue research on learning. For instance, the built-in intelligence that many applications have may also log the learner's actual activities, transforming them into models and statistics. Research reveals that such data-traces sometimes point in other directions than traditional self-reports, where students are asked about their learning activities or their use of a certain application or system. We wished to pursue this line of inquiry and explore if there are some types of estimations that correlate with objective measures, and hence if some types of estimations are better carried out through objective measures than through self-reports. Our study compared data-logged activities (clicks and eye-tracking measures) with self-reports on game-behavior for two types of estimations (requests for more feedback and reading feedback). To our knowledge, this kind of inquiry, comparing different types of data on feedback management, is rare. The results reveal that even if the students were quite good at estimating and reporting their requests for more feedback (in terms of clicks), they were substantially poorer at evaluating the extent of feedback messages they had read. These findings suggest that estimations of tasks that require more complex inner cognitive processes (e.g. reading) are more difficult for students themselves to report appropriately – whereas more procedural processes or concrete behaviors (i. e. clicking) are easier. These results are further discussed together with possible limitations in the measures used.

Keywords: Educational Games, Feedback, Reading behavior, Log files, Self-reports

1. Introduction

The digitalization of learning environments during the last decades has transformed education and learning science in fundamental ways. Not only may intelligent systems individualize learning by, for instance, providing automatic feedback, hints, or tailored instructions to a learner. In addition, log files may be collected and transformed into fine-grained statistics that may reveal how learners engage in a certain task, what resources they use, or if they engage in self-regulated learning. Such data traces, it turns out, often differ from self-reports, where students are asked to make judgements – often retrospectively – about their own applied strategies, actions, and efforts (Perry & Winne, 2006; Winne & Perry, 2000).

In fact, the validity of self-report instruments for evaluating learning activities is today broadly questioned, and in the field of self-regulated learning there are ongoing discussions of why self-report data may not be reliable indicators of the tactics learners actually use while they are studying (Hadwin, Nesbit, Jamieson-Noel, Code & Winne, 2007; Cho & Yoo, 2017). This might be particularly true for younger students, since they – due to less developed metacognitive capacities than those in youths and adults – have great difficulties when it comes to self-judgments (Demetriou & Kazi, 2006), and often strongly overestimate their own performances and abilities (Demetriou & Efklides, 1989). Younger children also tend to confuse the perceptions of academic ability with perceptions of appropriate social behavior (Paris & Newman, 1990). However, making proper judgements about one’s own learning strategies is considered important for gaining new knowledge (Zimmerman, 2001), and previous studies

show that metacognitive skills also often correlate with general performance (Bransford, Brown & Cocking, 2000).

One alternative to subjective reports is, of course, 'objective' ones. And, as stated above, today's digitized learning environments are developing rapidly, not at least in the field of learning analytics. Faster computers, better data storage together with sophisticated statistical tools make it possible to gather fine-grained and real-time digital patterns and feed this back to students, teachers, or researchers in a series of ways. And many researchers point to the possibility of not only evaluating students' active choices and strategies, but also their inner cognitive states and processes (Azevedo et al, 2013; Koedinger, d'Mello, McLaughlin, Pardos & Rosé, 2015) – all in the pursuit of optimizing teaching and learning situations and supporting students to become effective and self-regulated learners. One of the added values with gathering behavioral data in learning contexts is also to predict learning outcomes.

In this quest, it is of great importance to understand more about the pros and cons regarding objective versus subjective data, and to learn more about how such data best should be combined. However, before throwing ourselves into the big complex pond of learning analytics and student modelling, it may be of interest to start at a more fundamental level – not at least since many learning processes still remain to be explicitly defined. For instance, what are the relations between traceable interactive behaviors (such as clicking, reading) and the students' own thoughts about their learning processes (i. e. requesting/wanting information and/or processing it)? And what differences between self-reports and data logs may we expect when we focus on such processes and strategies? The purpose of this paper is to contribute to this line of research by looking at data from a specific kind of SRL-strategy, namely *feedback management*. By investigating two different types of behaviors – namely 'requesting' feedback and 'reading' feedback messages – we hope to spread some light over how self-reports and data traces may reveal different things, not only about the processes in general, but also about the student as a learner. To our knowledge, this kind of inquiry, comparing different types of data on feedback management, is rare.

1.1 Subjective and objective measures for evaluating SRL

When it comes to evaluating differences between self-reports and digital traces, most studies – at least those focusing on SRL – have mainly been concerned with the measures' appropriateness for modelling learning profiles or predicting learning gains. Here, Cho and Yoo (2017) found that subjective measures from the classical MSQL-questionnaire (Pintrich, Smith, Garcia & McKeachie, 1991) were less appropriate for predicting students' achievement levels (in form of final grades) than data traces from an online learning management system (Blackboard). However, van Halem, van Klaverena, Drachslerbc, Schmitzd, & Cornelisza (2020) received partially different results when comparing MSQL-reports with logs from another LMS (Canvas) accompanied by logs from a learning tool in statistics. Here, van Halem et al (ibid.) concluded that while some MSLQ-reports could explaine a substantial *proportion* of academic performance, other variables based on trace data instead might reveal a substantial *variation* in performance. And in another study by Ellis, Han and Pardo (2017), it was shown that the *combination* of self-reports and observational data provided a better predictive model of students' learning outcomes than any of the measures alone.

Fewer studies have made direct and straight-forward comparisons between self-report measures and data logs. However, Hadwin et al (2007) did an exploratory case study on eight students using gStudy, where they paired MSQL-reports with digital online behaviors (such as making questions, finding ideas, linking information, summarizing ideas, making lists and defining lack of knowledge). In general, self-reports correlated poorly with data logs, although they were slightly more comparable for some actions, such as summarizing important information. In another study, Kia, Hatala, Baker & Teasly (2021) added pop-up windows to an existing LMS (Canvas), where students answered multiple-choice questions about their ongoing behaviors (stated like 'which of the statements below best describes what you are doing on this assignment at this moment?'). These responses were then evaluated and compared to data logs categorizing different SRL-behaviors (task definition, planning, enactment, and adapting). It was shown that the correspondence between students' self-reports and categorized data logs varied between 32% and 73%. The most difficult behaviors to self-assess (or to interpret digitally, depending on what is thought of as the most correct) were 'planning' and 'adaption', while 'task definition' and 'enactment' (i.e., activities corresponding with working on the assignment) were easier.

Evidently, self-evaluating one's own behaviors is not straightforward, but even if self-reports have limitations, the weaknesses of data traces also need to be emphasized. As an example, a study by Tempelaar, Rienties & Nguyen (2020) confirms the individual biases in self-reports but lifts forth that the very biases also may be used as predictors for student performance. In addition, they point to the fact that data logged behaviors, such as a 'high level of learning activity' may have a series of different causes. A 'high level of activity' may signal an engaged and well performing student, but may also indicate a conscientious student, or a student with low proficiency in need of extra learning efforts. The authors conclude that especially trace data of 'process type' (e. g. is, number of attempts to solve an exercise, time on tasks, number of assignments completed) should be used with caution, whereas data traces of 'product type' (e. g. proportion of exercises correctly solved) could be better trusted, and especially so if complemented with self-report data. These conclusions are in line with other studies (Tempelaar, Rienties, Giesbers, 2015; Tempelaar, Rienties, Mittelmeier & Nguyen, 2018).

To summarize, even if some measures may be more reliable and/or useful than others for some specific purposes, it is not clear why and how subjective and objective measures of different phenomena tell us different things. However, in their review of 14 SRL-studies between 2003 and 2015, Rovers, Clarebout, Savelberg, de Bruin & van Merriënboer (2019) conclude that *granularity* is an important concept when comparing different measures of SRL, They also conclude that while self-report questionnaires may give a rather accurate insight into students' global level of self-regulation, behavioral objective measures are more accurate when evaluating specific SRL-strategies. This, on the other hand, seems like a simplification when considering the findings of Tempelaar et. al (2020).

1.2 Subjective and objective measures of feedback engagement and reading

Students' engagement with feedback is an important aspect SRL. For efficient learning to occur, you should attend to feedback, read it, understand it and act on it properly. When it comes to measuring students' thoughts of and use of feedback, this has historically been done through surveys, interviews or observations (Kerr, 2017; Eriksson, Björklund Boistrup & Thornberg, 2022). A classical setup is here to ask students what kind of feedback they like or dislike, if and how they make use of the feedback and why, and how the feedback ought to be formulated. On the other hand, when it comes to interpreting data logs for analyzing feedback behavior – which has been done in a series of studies during the last decades – the students' subjective opinions about the feedback (and their handling of it) have been set aside. Instead, the goal with the measurements has been to gather 'true' feedback engagement by, for instance, counting requests for feedback (Ternblad & Tärning, 2020), looking at if the feedback has been used when revising tasks (Silvervarg, Wolf, Blair, Haake & Gulz, 2021) or when the students engage with it (Chen, Breslow, DeBoer, 2018). Such data traces have often been gathered through the use of specifically designed digital learning applications or educational games (Tärning, Lee, Andersson, Månsson, Gulz, & Haake, 2020; Biswas, Roscoe, Jeong & Sulcer, 2009), but are, as far as we know, rarely analyzed in more open-ended learning systems.

To our knowledge, subjective and objective measures of feedback engagement have never been compared. This might be due to the simple structures of the used variables (such as clicking for getting more feedback) and that the digital footprints in this case seem to speak for themselves. To ask the students if they requested more feedback or not does in this case appear superfluous (if they don't remember clicking, they are simply mistaken, since they obviously did click). However, the student's judgements about their own behavior (that is, if they think of themselves as individuals wanting feedback, or if they are convinced that they behave appropriately even if they do not) may still be of interest. And equally, it may also be of value to evaluate if such judgements differ in respect to what is measured. Perhaps some behaviors are easier to self-evaluate than others? This is certainly the case when it comes to more complex measures, as Hadwin et. al (2007) and Kia et. al (2021) have pointed out.

One important aspect of feedback engagement is to properly attend to and read the feedback. To evaluate if and how students have read a text is mostly done by asking questions about its content. That means equating reading comprehension with reading (which may, of course, be the same thing). However, paying attention to a text and *trying* to read it, but without fully understanding it, is not the same thing as not attending to it at all. And if looking at these two behaviors in an SRL-context, not paying enough attention is far worse. So, how to measure and assess 'reading' without confounding it with 'understanding'?

One way of evaluating students' reading behaviors is, of course, to plainly ask them. This was typically done for evaluating reading habits and reading comprehension strategies during the late 1900's, revealing that students in all age groups have difficulties to self-report how and what they actually have read (Baker & Cerro, 2000). Not only do self-report measures concerning reading strategies lack a theoretical basis, but they are also highly context- task- and goal dependent (Hadwin et. al, 2007). Moreover, Cromley and Azvedo (2006) conclude, studying ninth graders readings of texts on American history, that one of the problems with self-reports is that they often are collected outside of the learning context (after of before the actual reading takes place). They compared prospective self-reports on specific reading behaviors with think-aloud protocols and a concurrent control measure of actual strategy use (gathered by multiple choice-questions). When trying to relate these measures to one another – as well as to a test on reading comprehension – they stated that the prospective self-reports were the only ones not correlating with any of the other measures.

A different, and today more common approach is to evaluate students' reading behaviors by using eye-tracking measurements. Eye-tracking is a non-intrusive measurement technique based on the Eye-mind hypothesis originating from Just and Carpenter (1980), which states that visual fixations on objects (or words) is strongly correlated with cognitively processing them. And even if it's important to bear in mind that the detection of eye-movements does not *directly* correspond to underlying cognitive processes (we may, for instance, attend to something that is not in our visual focus), it still constitutes a reliable instrument for evaluating capacities like information processing and visual attention (Carter & Luke, 2020). When detecting reading behaviors in eye-tracking studies, scanpath patterns are normally used together with durations of fixations and recordings of blinks and saccades. By using well proven algorithms (often based on Hidden Markov Models) it is then possible to predict whether a given sequence of events reflects a reading or not (see for instance Kollmorgen & Holmqvist, 2007; Simola, Salojärvi & Kojo, 2008; Wengelin et. al, 2009). And it is of course also possible to evaluate if some specific areas or objects have been attended to by merely looking at fixations. To our knowledge, eye-tracking measures capturing reading have rarely been compared to self-reports. Instead, they have been evaluated through reading comprehension tests and free recall.

1.3 The scope of the present study

This study has two purposes: First, to contribute to the line of research comparing subjective and objective measures important for SRL, and see if we, with our specific dataset and our specific subject (that is, feedback engagement), get similar results as previous studies, and if not, discuss why. Second, to illustrate how subjective and objective measures on feedback engagement may differ due to what is specifically measured. For this paper we looked at two different estimations students did after playing an educational game in history. On the one hand, we investigated fifth grader's retrospective self-evaluations of (i) how large a proportion of automatically presented feedback messages they read while playing an educational game, and (ii) how large a proportion of optional feedback messages they did request while playing. These self-reports were then compared to the students' gaze behavior (using eye-tracking) and logged actions (in terms of clicks), searching for correlations, deviations or similarities.

Of special interest to us was whether the subjective report of a more cognitive demanding and also more ambiguous task like reading (ambiguous in that way that it can mean different things to different people) would correlate less well with data-logging (i.e., eye-tracking measures) whereas self-reports for a more concrete task (estimating the number of clicks) would correlate better with the objective measures (i.e. data-logging). If such a discrepancy turned up, we were interested to learn what it could tell us about a) the different content and representability of the measures, and b) the students' metacognitive capacities of judging their own behavior. More specifically, our research questions were:

i) How good are middle-school students at judging their own 'clicking' and 'reading' behavior? And are there any differences between estimating 'reading' and estimating 'clicking' when comparing self-reports with data traces?

ii) Can potential differences between self-reports and data traces be used to predict performance?

Notably, the research questions specifically target students' engagement with feedback messages. And since engaging with feedback is an important part of self-regulated learning, we were also interested in discussing the possible benefits of using both types of measures and if and how comparisons between them may be used for evaluating students' SRL-skills.

2. Method

2.1 Participants, procedure and materials

45 5th graders (24 girls and 22 boys) from two classes from a school in southern Sweden participated in a study on feedback behavior by playing an educational game in history. In the game, students visited historical scenes and persons, searched for text-based information and solved tasks (either by constructing a concept map, categorizing statements, or pairing historical figures and events on a timeline) (see also Tärning et al, 2020). After having completed a task and having it corrected by a "correction machine", the students also received informative text-based feedback on selected errors and hints on how to correct them (see Figure 1 for a time-line task and examples of feedback on possible errors). The feedback was designed to be critical and constructive, scaffolding the students to do better when they revised the task. After the first feedback message, the students were offered to click on other marked errors to receive additional pieces of feedback. The students had to be at least 80% correct on a task to be allowed to start on a new one, and many students needed many trials to be able to progress in the game.

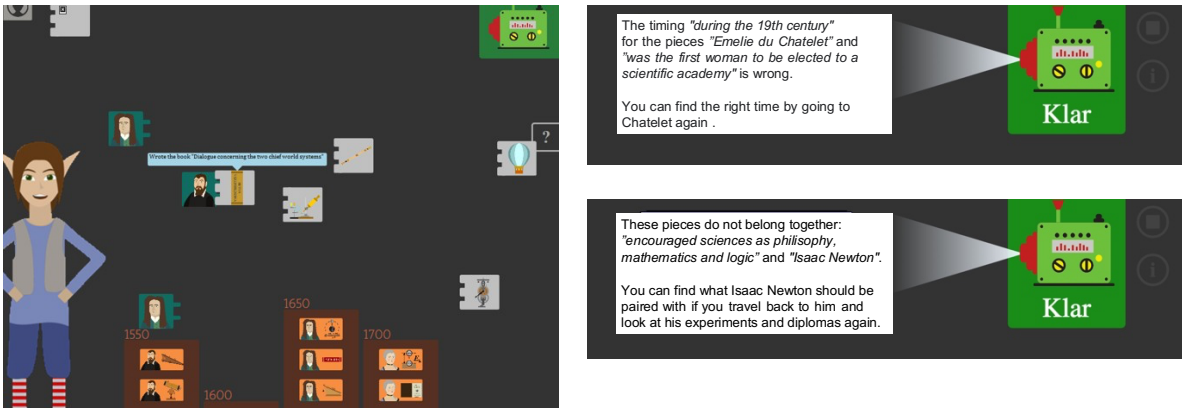


Figure 1. Timy constructing a time-line with historical persons and their scientific contributions (left), and examples of feedback from the correction machine (right).

In total, the students worked with the game during three one hour sessions. During the third session (the data from which we base the present analysis on), the students used computers with integrated eye-trackers situated in a classroom-like laboratory at Lund University. The experiment ended with the students filling in a questionnaire with questions regarding their efforts and behavior when playing. One of these questions targeted to what extent the students had read the feedback provided to them, while another asked to what extent they had clicked to receive more feedback on some other errors or mistakes. For both questions there were five possible answers: "always", "most often", "sometimes", "not very often" and "never".

2.2 Eye-tracking measurements and data logs

All screen activities were logged together with eye movement events (fixations, saccades and blinks) using an integrated SMI REDm eye tracking camera with SMI iViewX and SMI BeGaze 3.6 software with a sampling rate of 120 Hz. The eye-tracking recordings were used for evaluating if the students *noticed* the feedback or not, and if they *read* the feedback or not. For measuring if the feedback message was noticed, we used the *fixation based AOI hit*, which states for a fixation that its coordinate value is inside the area of interest (AOI) (Holmqvist et al., 2011). An AOI is defined as a region in the stimuli

in which the researcher is interested in gathering data. Consequently, in the present study, the areas containing feedback boxes were defined as AOIs. After verifying an AOI hit, every feedback instance was categorized as either 'noticed' or 'ignored', and the proportion of noticed feedback messages was calculated for each participant.

'Reading behavior' was identified by a learning algorithm in form of a support vector machine (SVM) (Lee, 2017). This was calibrated for each participant, using three eye movement measures: duration of fixation, saccadic amplitude, and regression. Every feedback text was then classified as 'read' or 'not-read' based on an intrinsic threshold determined in a pilot study (Lee, 2017). The reason for not using existing standardized models (which are based on adult readers) was due to the students' age and large variability in gaze behavior. Consequently, by simplifying the eye-tracking measure and adjusting it to each individual student, we ensured capturing a sufficient degree of 'reading-like' events.

Finally, we also calculated an in-game *performance measure* by using the logged scores (percent correct) for each task trial and averaging these values over all tasks and trials.

3. Results

To be able to compare self-reports with eye-tracking data, the five levels of 'clicked' and 'read' feedback from the questionnaires were translated into numbers: 100%, 75%, 50%, 25% and 0%. Data from 42 students was included in the final analysis, and the results are described in the passages below. It should be noted, that even if statistical significance levels are calculated and presented, no corrections for multiple analyses have been performed. The analysis is explorative – thus the results should be interpreted as indicative, mainly pointing in directions that could be further investigated.

3.1 Comparison between self-reports and data traces targeting feedback clicks

We start with an evaluation of the students' requests for additional feedback (here referred to as 'clicks'). The self-reported extent of feedback clicks (35% in average) was similar to the logged one (38% in average). A Mann-Whitney U-test revealed no significant difference between measures ($W=855$, $p = 0.67$). Studying the relationship between individual values by calculating Pearson's correlation coefficient, the objective and subjective measures were found to be significantly positively correlated: $r(34) = 0.40$, $p < 0.01$ (see Figure 2).

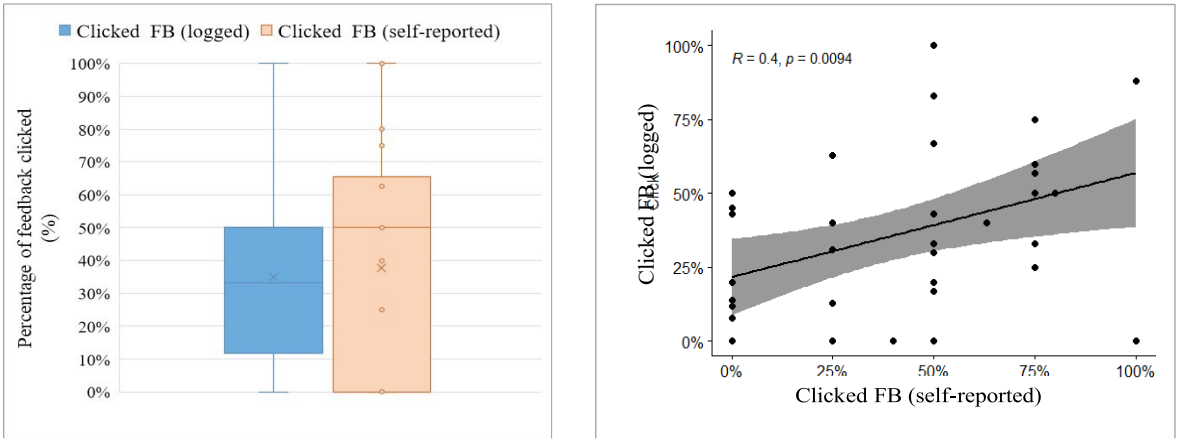


Figure 2. Comparison between logged and self-reported feedback requests (clicks). Left: Boxplot. Right: Scatterplot with Pearson's Correlation.

3.2 Comparison between self-reports and data traces targeting readings of feedback

When it comes to the students' reports of feedback readings, the self-reported extent of feedback read (59% in average) was larger than the one based on eye-tracking measurements (47% in average). A Mann-Whitney U-test here revealed a significant difference between the two measures ($W=486$, $p< 0.01$), indicating that the students, in general, overestimated the extent to which they read the messages.

In this case, the subjective and objective measures were not found to be correlated: r(34) = .05, p = .77. See Figure 3 for visualizations of the results.

To evaluate whether the students confused reading the feedback with merely 'glancing at it', the students' self-reported extent of feedback read (59%) was also compared to the logs of feedback 'noticed' (72%). A Mann-Whitney U-test revealed a weak but still significant difference between the two measures (W=777, p = 0.02), indicating that the students, in general, only self-reported a part of the noticed feedback messages as 'read'. These two measures were not found to be correlated either: r(34) = .04, p = .81 (see Figure 4).

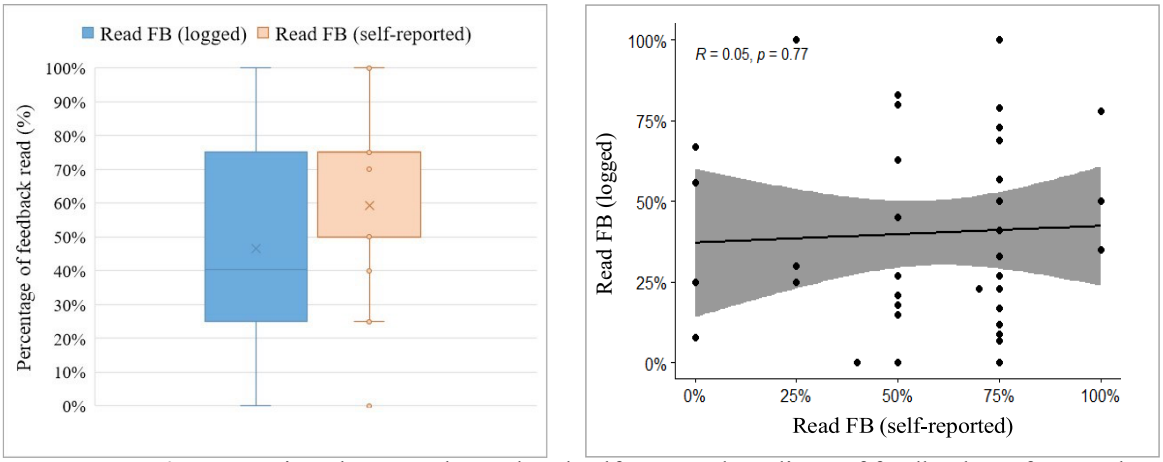


Figure 3. Comparison between logged and self-reported readings of feedback. Left: Boxplot. Right: Scatterplot with Pearson's Correlation.

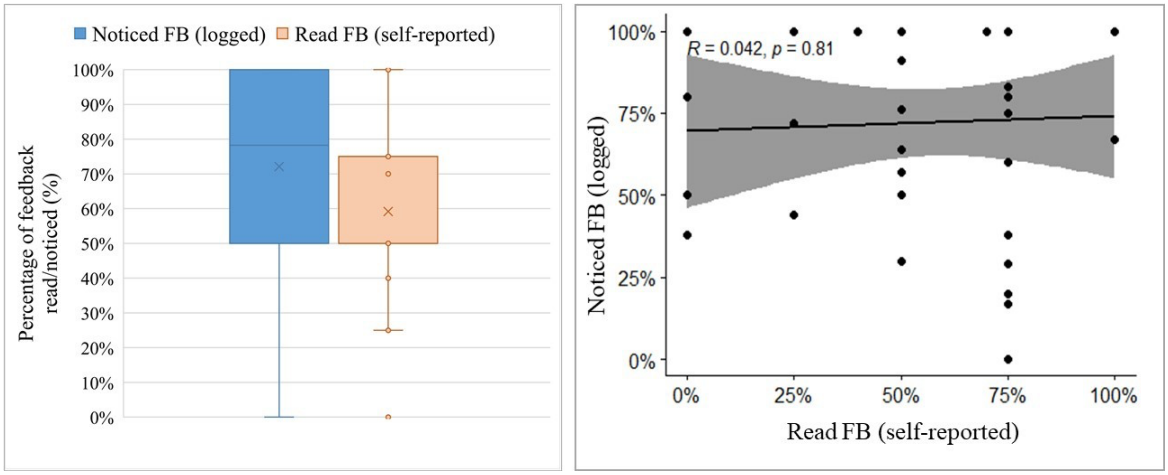


Figure 4. Comparison between self-reported readings of feedback and logs of feedback noticed. Left: Boxplot. Right: Scatterplot with Pearson's Correlation.

One might expect that the self-reported readings of feedback messages would lie somewhere in between the logs of 'noticed' and 'read' messages, and this was also true for 47% of the participants. However, 31% underestimated their reading, reporting to have read a lower extent of the messages than what was logged as 'read'. On the other hand 22% reported to have read a higher extent of messages than what was logged as 'noticed' – clearly making an improper judgement about their reading behavior, since it is impossible to read what you don't notice.

3.3 Comparisons of (mis)judgments and performance

As stated above, metacognitive capacities may correlate with academic performance. We therefore decided to investigate if the incorrectness of the individual student's judgement would correlate with their performance in the game, here focusing on feedback readings (since these were difficult to estimate). We could expect a negative correlation between this incorrectness (measured as the absolute

difference between the self-reports and logged feedback readings) and the average performance level (measured as % correct over all tasks and trials). The result is presented in Figure 5 below, revealing that the inaccuracy in the self-reports did not have any significant correlation with the students' average performance levels: r(34) = .17, p = .33.

As of yet, the self-reports on how much feedback the students had read haven't corresponded to any other measure, and the value of asking students about their reading behavior appears, so far, to be questionable. However, one final analysis, addressing if the self-reports could reflect other desirable and productive strategies (such as really trying to solve the tasks or performing well) is in its place. Hence, we decided to do a comparison between the self-reports on feedback readings and in-game performance. The result is presented in Figure 5 below, showing that the two measures had a positive and significant correlation: r(34)= 0.41, p=0.01.

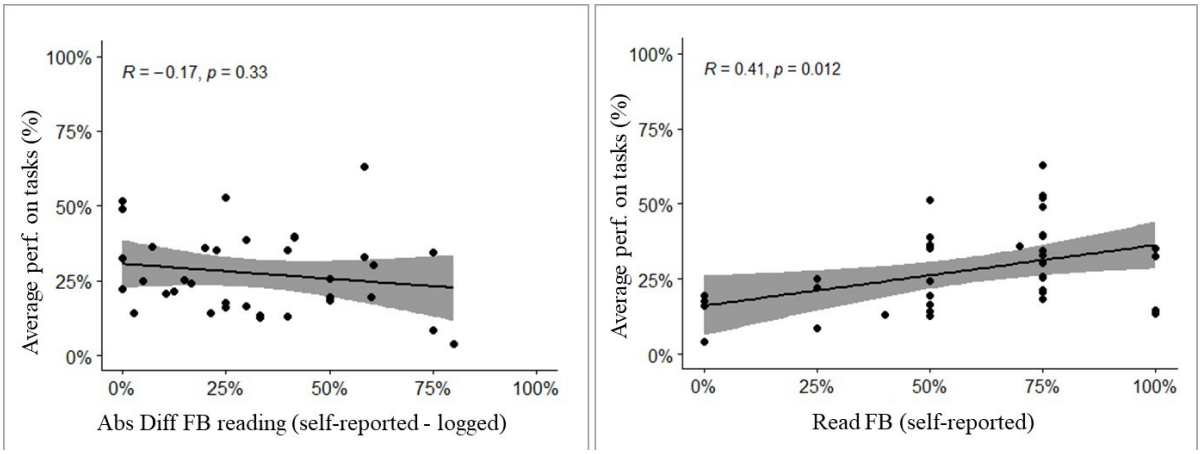


Figure 5. Scatterplots comparing in-game performance levels with the absolute differences between self-reports and data traces (left) and with the self-reports on feedback readings (right).

4. Discussion

Evidently, in the present study the students' ability to correctly judge to what extent they had 'read' a feedback text was substantially weaker than their ability to estimate to what extent they had made requests for additional feedback texts (by clicking). Even though some students overestimated (but also underestimated) their inclination to 'click', this behavior was less difficult to evaluate correctly compared to the cognitive, and quite complex, behavior of 'reading'. This is in line with previous research, showing that reading is difficult to self-assess (Baker & Cerro, 2000; Hadwin et. al, 2007). The difficulty in observing one's own reading behavior could be due to the fact that attention directed towards a text does not necessarily involve deeper engagement. The text could be 'noticed', 'glimpsed at' or 'skimmed' – without any proper decoding or understanding. However, the self-reports on feedback readings didn't correspond to the logged eye-tracking measures of having 'noticed' the feedback either. Indeed, self-evaluating this process appears to be hard.

Since reading is an important aspect of self-regulated learning, judging it properly is essential. However, many automatic processes that learners apply, and reading is one of them, often operate beneath the threshold of attention, making them hard to retrieve from memory (Perry & Winne, 2006). The self-report concerning 'reading the feedback' may therefore instead be an after-construction, related to performance, effort, or desirable behavior in school (a similar discussion regarding relations between retrospective-self-reports and achievement is done by Veenman and van Cleef (2019)). Hence, depending on if the students' feedback engagement led to a success or failure, the interpretation of 'having read the feedback' might differ.

In addition, the students' ability to correctly judge how much feedback they read (by looking at the differences between self-reports and data traces from eye-tracking measurements) did not correlate with in-game performance, questioning the actual benefits with comparing these two types of measures. Taken together, the findings suggests that 'reading' might be a typical process suitable for measuring objectively, without involving the student's own (and often biased) beliefs and wishes. Still, another

possibility is that the students only remembered reading (or not reading) the *latest* of the presented feedback messages, and thereby did not evaluate their behavior for the entire session. If this is the case, the problem is not the lack of accurate self-esteem, but flaws in memory retrieval. Consequently, questions about students' learning strategies should probably be posed when the students in fact are studying, and not after (or before, as Cromley and Azvedo (2006), also pointed out).

Before too strongly advocating eye-tracking for evaluating reading processes, the accuracy of the 'reading'-measure used in this study can, of course, also be questioned. Setting a different threshold would most certainly have led to a different outcome. And it might also be, that 'partly glancing at' or 'partly reading' the feedback messages could be more beneficial than not noticing them at all, and that 'reading' shouldn't be classified in a binary way. Unfortunately, at the same time as inner cognitive processes and operations are difficult to self-evaluate, they are also the most difficult ones to catch and quantify by intelligent algorithms and measurements. This is especially true for children, who behave not only differently, but also in a less stable and predictable way than adults.

It should finally be emphasized that students' self-reports may bring invaluable information about the student's own perceptions, attitudes, and beliefs about studying and learning. Hence, it might not be a question about which measures to use, but about *how* and *when* to register them and/or combine them. Gaining more knowledge about differences and similarities between measures is therefore essential, not at least before conducting large surveys about students' learning habits, or before turning data traces into advanced student models using sophisticated data mining techniques.

Acknowledgements

The presented research has been funded by Marianne and Marcus Wallenberg Foundation.

5. References

Azevedo, R., Harley, J., Trevors, G., Duffy, M., Feyzi-Behnagh, R., Bouchet, F. & Landis, R. (2013). Using Trace Data to Examine the Complex Roles of Cognitive, Metacognitive, and Emotional Self-Regulatory Processes During Learning with Multi-agent Systems. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 427-449). New York, NY: Springer.

Baker, L., & Cerro, L. C. (2000). Assessing Metacognition in Children and Adults. In G. Schraw & J. C. Impara (Eds.), *Issues in the Measurement of Metacognition* (pp. 99-145). Lincoln, NE: Buros Institute of Mental Measurements.

Biswas, G., Roscoe, R., Jeong, H. & Sulcer, B. (2009). Promoting Self-regulated Learning Skills in Agent-based learning Environments. In S. C.,Kong, H. Ogata, H. C.Arnseth, C. K. K. Chan, T. Hirashima, F. Klett, J. H. M. Lee, C. C. Liu, C. K. Looi, M. Milrad, A. Mitrovic, K, Nakabayashi, S. L. Wong, S. J. H. Yang, (Eds.) (2009). *Proceedings of the 17th International Conference on Computers in Education*. (pp. 67-74). Hong Kong: Asia-Pacific Society for Computers in Education.

Bransford, J.D., Brown, A.L. & Cocking, R.R. (2000). *How People Learn; Brain, Mind, Experience, and School (Exp. ed.). Washington, D.C:* National Academy Press.

Carter, B. T. & Luke, S. G. (2020). Best practices in eye tracking research. *International Journal of Psychophysiology, 155*, 49-62,

Chen, X., Breslow, L., DeBoer, J. (2018). Analyzing productive learning behaviors for students using immediate corrective feedback in a blended learning environment. *Computers & Education, 117*, 59-74.

Cho, M-H., Yoo, J. S. (2017). Exploring online students' self-regulated learning with self-reported surveys and log files: a data mining approach. *Interactive Learning Environments, 25*(8), 970–982.

Cromley, J. G. & Azevedo, R. (2006) Self-report of reading comprehension strategies: What are we measuring? *Metacognition Learning, 1*, 229–247

Demetriou, A., & Kazi, S. (2006). Self-awareness in g (with processing efficiency and reasoning). *Intelligence, 34*(3), 297–317.

Demetriou, A., & Efklides, A. (1989). The person's conception of the structures of developing intellect: early adolescence to middle age. *Genetic, Social, and General Psychology Mono- graphs, 115*(3), 371–423.

Ellis., R. A., Han, F., Pardo, A. (2017). Improving Learning Analytics – Combining Observational and Self-Report Data on Student Learning. *Educational Technology & Society, 20*(3), 158–169.

Eriksson, E., Björklund Boistrup, L. & Thornberg, R. (2022). "You must learn something during a lesson": how primary students construct meaning from teacher feedback. *Educational Studies*, *48*(3), 323-340.

Hadwin, A. F., Nesbit, J. C., Jamieson-Noel, D., Code, J., Winne, P. H. (2007). Examining trace data to explore self-regulated learning. *Metacognition Learning, 2*, 107–124.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye Tracking: A comprehensive guide to methods and measures*. Oxford, Great Britain: Oxford University Press.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological review*, *87*(4), 329.

Kerr, K. (2017). Exploring student perceptions of verbal feedback. *Research Papers in Education, 32*(4), 444-462.

Kia, F. S., Hatala, M., Baker, R. S. & Teasley, S. D. (2021). Measuring Students' Self-Regulatory Phases in LMS with Behavior and Real-Time Self Report. *In LAK21: 11th International Learning Analytics and Knowledge Conference*, Virtual Conference.

Koedinger, K. R., D'Mello, S., McLaughlin, E. A., Pardos, Z. A. & Rosé, C. P. (2015). *WIREs Cogn Sci*, 6:333– 353

Kollmorgen, S., & Holmqvist, K. (2007). *Automatically detecting reading in eye tracking data*. (Working Papers 144, Department of Cognitive Studies, Lund University.)

Lee, Y. J. (2017). Effects of teachable agents on children's noticing and reading feedback in a digital educational game (Document ID: 48039). [Master's thesis, University of Vienna]. E-Theses – Hochschulschriften- Service. http://othes.univie.ac.at/48039/

Paris, S. G. & Newman, R. S. (1990). Developmental aspects of learning. *Educational Psychologist 25*(1), 87-102.

Perry, N. E., Winne, P. H. (2006). Learning from learning kits: gStudy traces of students' self- regulated engagements with computerized content. *Educational Psychology Review, 18*(3), 211–228.

Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the motivated strategies for learning questionnaire*. MI: University of Michigan.

Rovers, S. F. E., Clarebout, G., Savelberg, H C. M., de Bruin, A. B. H & van Merriënboer, J. J. G. (2019). Granularity matters: comparing different ways of measuring self-regulated learning. *Metacognition and Learning 14*(1), 1–19.

Silvervarg, A., Wolf, R., Blair, K. P., Haake, M., & Gulz, A. (2021). How teachable agents influence students' responses to critical constructive feedback. *Journal of Research on Technology in Education*, *53*(1), 67-88.

Simola, J., Salojärvi, J. & Kojo. L. (2008). Using hidden Markov model to uncover processing states from eye movements in information search tasks. *Cognitive Systems Research, 9*(4), 237-251.

Tempelaar D., Rienties B., Giesbers B. (2015). In search for the most informative data for feedback generation: Learning Analytics in a data-rich context. *Computers in Human Behavior*, 47, 157–167.

Tempelaar D., Rienties B., Mittelmeier, J., Nguyen, Q. (2018). Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior*, 78, 408–420.

Tempelaar, D., Rienties, B., Nguyen, Q. (2020). Subjective data, objective data and the role of bias in predictive modelling: Lessons from a dispositional learning analytics application. *PLoS ONE 15*(6): e0233977.

Ternblad, E-M., & Tärning, B. (2020). Far from success – far from feedback acceptance? The influence of game performance on young students' willingness to accept critical constructive feedback during play. In I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), Artificial Intelligence in Education. Lecture notes in computer science: Vol. 12163. (pp. 537-548). Springer, Cham.

Tärning, B., Lee, Y. J., Andersson, R., Månsson, K., Gulz, A., & Haake, M. (2020). Assessing the black box of feedback neglect in a digital educational game for elementary school. *Journal of the Learning Sciences, 29*(4-5), 511-549.

Van Halem, N., van Klaverena, C., Drachslerbc, H., Schmitzd, M., Cornelisza, I. (2020). Tracking Patterns in Self-Regulated Learning Using Students' Self-Reports and Online Trace Data. *Frontline Learning Research 8*(3), 140–163.

Veenman, M.V. J., van Cleef, D. (2019). Measuring metacognitive skills for mathematics: students' self-reports versus on-line assessment methods. *ZDM Mathematics Education 51*(4), 691–701.

Wengelin, Å., Torrance, M., Holmqvist, K., Simpson, S., Galbraith, D., … & Johansson, R. (2009). Combined eyetracking and keystroke-logging methods for studying cognitive processes in text production. *Behavior Research Methods, 41*(2), 337-351.

Winne, P. H., Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 531-566). San Diego, CA: Academic Press.

Zimmerman, B. J. (2001). Theories of self-regulated learning and academic achievement: An overview and analysis. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives,* (pp. 1–37). Mahwah, NJ: Erlbaum.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Exploring Relationships Between Temporal Patterns of Affect and Student Learning

Anthony F. BOTELHO[a], Seth A. ADJEI[b], Vedant BAHEL[c], & Ryan S. BAKER[d]
[a]*University of Florida, USA*
[b]*Northern Kentucky University, USA*
[c]*University of British Columbia, Canada*
[d]*University of Pennsylvania, USA*
*a.botelho@ufl.edu

Abstract: Numerous prior articles have studied the relationships between student affect and various outcomes of learning. Prior research has found these relationships are complex; shifts in student affective experiences and the duration in which students remain in particular states form temporal patterns that are often difficult to interpret. Much of the existing research in this area focuses on the correlation between the overall prevalence of particular affective states and learning outcomes, ignoring the temporal and sequential characteristics of student affect. In this work, we leverage temporal clustering methods to identify emerging patterns of affect while students work within a computer-based learning platform to explore how sequence patterns of student concentration, boredom, confusion, and frustration correlate with student performance on a delayed assessment test. Similar to prior work, we find strong relationships between affect and student learning, even accounting for a measure of prior knowledge. Additionally, we identify that the directionality of these correlations in regard to specific affective states differs across clusters. While some affective states such as boredom are identified to exhibit negative relationships with learning within some patterns, a strong positive relationship between frustration and learning is found within one of the emerging pattern clusters.

Keywords: Affect, Temporal Clustering, Student Knowledge, Affect Dynamics

1. Introduction

There is a complex relationship between students' affective experiences and their learning. Affective transitions are highly complex and often do not seem to follow a specific interpretable sequence (Karumbaiah et al., 2021a), and there is also variation in the duration of individual affective experiences. Despite these known complexities, much of the research on the relationship between affect and learning looks for linear relationships between the prevalence of affective states with learning gains. This line of work has shown some clear relationships (see research synthesis in Karumbaiah et al., 2022). For instance, boredom is associated with poorer learning outcomes in several studies (e.g., Rodrigo et al., 2009; Bosch & D'Mello, 2017; Gong et al., 2019)) and engaged concentration is replicably associated with better learning outcomes in several studies (e.g., Rodrigo et al., 2009; Brom et al., 2014; Fancsali, 2014). However, confusion has inconsistent relationships with learning, varying considerably between studies (Karumbaiah et al., 2022). One empirical study suggested that the key to the inconsistent findings around learning's relationship to confusion may be the duration of these affective experiences. Liu and colleagues (Liu et al., 2013) found that extended confusion and frustration were associated with poor learning outcomes, but that brief confusion and frustration were associated with positive learning outcomes.

While previous works have observed trends within and across affective states, the current work attempts to observe how patterns of affect relate to student learning as measured by a delayed measure of learning (described further in a later section). It is therefore important to further define what these patterns refer to and how this work may be distinguished from other related research on student affective patterns (Andres et al., 2019) as introduced in the previous section. To help illustrate this, consider the example sequence in Figure 1. Each of the lines represents a probabilistic estimate of one of four affective states (engaged concentration, boredom, confusion, and frustration) over the span of the assignment. In describing a sequence such as this one, prior works including D'Mello & Graesser (2011)

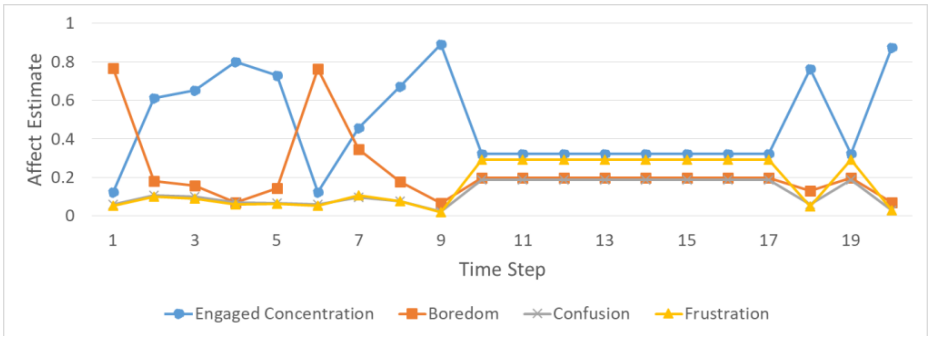


Figure 1. Example sequence of affect estimates for a student working on an assignment.

and Botelho et al. (2018) would have described affective chronometry, or periods of this sequence where the student remained in a single affective state for an extended time period. Alternatively, other works such as D'Mello & Graesser (2012) would have described the sequence's affect dynamics, characterized by the shifting between one affective state and another. Observing this sequence, however, it can be noted that there are few consistent trends; shifts and periods of affect appear to occur in bursts, with subsequent periods of uncertainty where there is arguably no clear singular state exhibited. The patterns exhibited within and across these affective states may provide novel insights into the relationships between affect dynamics and student learning.

Work on affect dynamics has not been extensively connected to the work on affect chronometry. Typically, chronometry is only considered in affect dynamics in terms of comparing brief (20 second or less) and extended periods (60-80 seconds) of the same affective state (Liu et al., 2014; Andres et al., 2019). Another limitation in the affect dynamics literature is the focus on relatively simple one-step models, which only look at single affective transitions (i.e. from state 1 to state 2). Affect appears to manifest in complex patterns over time, and our understanding of these is limited.

Temporal or sequence clustering is now a well-studied practice in research pertaining to pattern recognition (Liao, 2005). While specific implementations of time-series clustering methods vary, many rely on the generation of correlation-, spectral-, and wavelet coefficient-based features (Montero & Vilar, 2014) in combination with "warping" or scaling procedures (Berndt & Clifford, 1994) to describe the sequence holistically, before clustering based on these generated features.

There have been major advances in the detection of affect as well. Early work in affect detection within learning relied upon a complex array of sensors (Conati et al., 2003; D'Mello et al., 2007), utilized in laboratory settings, making it difficult to scale across long time periods and large selections of learners. Although more scalable sensor-free affect detection soon emerged, early models were only moderately better than chance (D'Mello et al., 2008; D'Mello & Graesser, 2012). Recent work leveraging deep learning algorithms has increased the quality of sensor-free affective detectors to the point where they can plausibly be used for fine-grained work (Botelho et al., 2017; Hutt et al., 2019), achieving AUC values approaching or equaling 0.8 for many affective states. Furthermore, even early affect detectors were more likely to obtain statistically significant findings about rare affective states than self-report or observational methods (Karumbaiah et al., 2022), possibly because they provided the opportunity to collect a much larger number of measurements, by 3 or more orders of magnitude.

In this paper, we leverage contemporary affect detectors to study affect dynamics and learning, specifically applying affect detectors developed using deep learning for the ASSISTments platform (Botelho et al., 2017). Specifically, we apply temporal clustering algorithms to investigate the patterns of student affect that emerge within student work on an assignment and examine whether affective states correlate differently with learning outcomes between the emerging patterns.

2. Methodology

The dataset used in this work is composed of interaction logs from students who used the ASSISTments learning platform (Heffernan & Heffernan, 2014) to complete mastery-based "skill builder" assignments during the 2016-17 through 2018-19 academic years. A skill builder is a problem set from which the system continues to present a series of problems to a student until they answer 3 consecutive

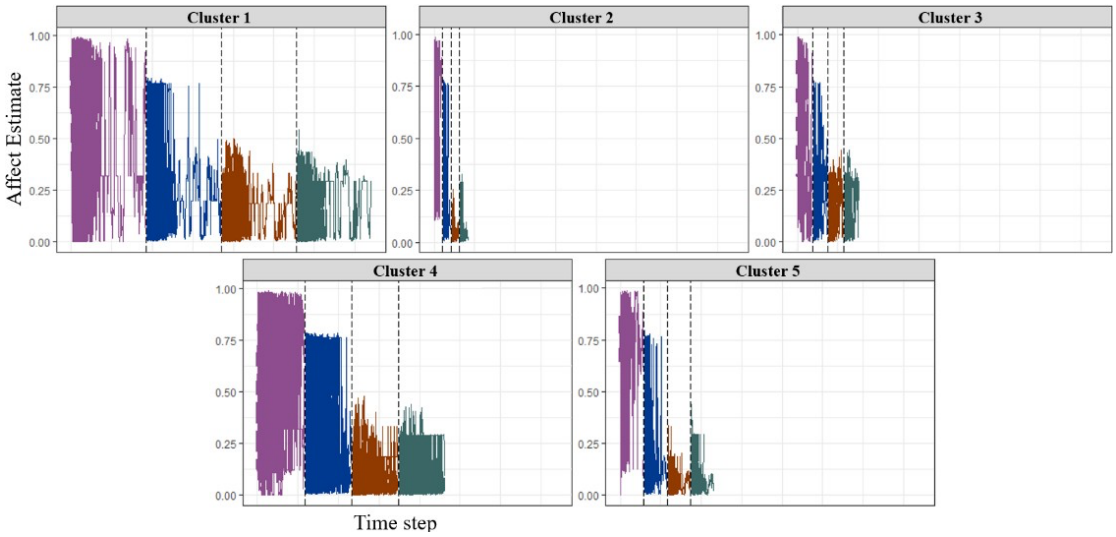


Figure 2. The sequences of affect estimates within each of the five affect pattern clusters.

problems correctly. The dataset used here contains 3065 skill builder assignments from 867 students; this data did not include students from rural settings, following prior works which have identified that the affect detectors used in this work have difficulty extrapolating to this population (Ocumpaugh et al., 2014; Karumbaiah et al., 2021b). We applied a deep-learning-based affect detector, pre-trained on historic ASSISTments data, developed and reported in Botelho et al. (2017), based on the collection of affect labels collected using BROMP classroom observations (Baker et al., 2020). The Botelho et al. (2017) model utilized in this work inherently assumes that students exhibit only one affective state at a time and produces a probability distribution over the 4 affective states.

To examine the relationships between patterns of student affect and learning, we utilize another feature of ASSISTments which provides a delayed learning outcome measure. The Automatic Reassessment and Relearning System (ARRS) is a feature of the ASSISTments ecosystem which can be coupled with skill builder assignments where teachers opt to have students reassessed on the knowledge component they had learned during the initial skill builder assignment (Xiong et al., 2013). This assessment has previously been used as a measure of student learning (Wang & Heffernan, 2015).

2.1 Patterns of Affect in Relation to Learning

To explore the relationship between patterns of affect and this measure of student learning, we apply a two-phase methodology. First, we cluster sequences of affect estimates and then we use those clusters within a regression analysis. Traditionally, clustering methods are used to group similar samples based on their proximity within a defined feature space. However, in this work we are observing sequences of student affect, where each time step in a sequence is defined by four probabilistic affect estimates.

In this work, we utilize the TSClust R package (Montero & Vilar, 2014) to generate the temporal features for our observed sequences of affect estimates before then applying a K-means clustering method based on a dynamic time warping distance measure (Ratanamahatana & Keogh, 2004; Lemire, 2009). This process involves first creating a matrix of samples as described by a vector of features before then using a simple elbow method to examine the ratio of within- and cross-sum of squared distances of samples at increasing numbers of clusters (i.e. the range of 2 to 15 clusters in our analysis). From this, we select K=5, and use a coding package-supplied method to plot these 5 emerging clusters as seen in Figure 2. These clusters represent different temporal patterns of student affect.

We build 5 separate logistic regression models, one for each cluster, to examine how each affective pattern correlates with our learning outcome. For each cluster, we generate a set of features to help describe the dynamics of each affective state within the cluster; these features include the natural log of the length of each sequence in addition to the mean and variance of detector estimates of each of three observed affective states (boredom, confusion, and frustration) composing each sequence; the fourth affective state, engaged concentration, is omitted due to a high inverse correlation and collinearity between these estimates and the other three affective states. Due to the large differences in

Table 1. *Coefficients for each cluster's regression predicting the one-week assessment performance. Boldface indicates statistical significance while marginally-significant values are denoted with* †.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Intercept	-0.214	0.143	-0.320	0.522	3.078
Prior Percent Correct	**2.521**	**3.140**	**3.221**	0.352	1.520
Log Sequence Length	**-0.457**	-0.236	**-0.428**	**-0.355**	-0.870
Mean Boredom	0.189	-0.538	**-0.559**	-0.242	-0.721
Boredom Variance	0.265	$0.824^{\dagger}$	-0.301	-0.310	-1.474
Mean Confusion	0.275	0.286	0.149	-1.002	-0.710
Confusion Variance	0.168	-0.390	0.005	-0.113	-1.117
Mean Frustration	-0.234	-0.837	0.299	**1.827**	0.110
Frustration Variance	0.221	-0.899	0.048	0.687	0.312
High Boredom X Low Boredom Var.	0.336	0.203	0.422	0.583	$2.562^{\dagger}$
High Confusion X Low Confusion Var.	-0.051	0.555	0.154	-0.199	1.307
High Frustration X Low Frustration Var.	-0.261	0.046	-0.430	-0.979	-0.541

scales of estimates across the different affective states, likely caused by the rarity of confusion and frustration, we use a median split to dichotomize the mean and variance values to represent high/low mean and high/low variance.

3. Results

Figure 2 shows the results of the temporal clustering analysis. Within each cluster, the chart displays 4 sections denoted by a vertical dotted line, ordered by engaged concentration, boredom, frustration, and confusion. Within each cell, the x-axis refers to the time step of the contained sequences while the y-axis represents the estimates of each affective state; the lines within each section of each cell represent an overlay of these estimates for each of the affective sequences found within the cluster. Clusters 1 and 4 show the longest sequences; this denotes that the students who spent the most time and attempted the most problems are contained within these clusters, but does not necessarily indicate that students remained in any particular affective state for a longer period compared to other clusters.

The 5 regression analyses are shown in Table 1. Across these regressions, there are several observations worth highlighting. First, the measure of prior knowledge (prior percent correct), reliably predicted assessment performance in Clusters 1, 2, and 3, but not in Clusters 4 and 5. The length of the sequences was a reliable (negative) predictor in Clusters 1, 3, 4, and 5, likely as an inverse measure of both initial knowledge and speed of learning within this mastery-based assignment.

4. Discussion and Future Works

Analyzing the relationship between affect and learning across five distinct temporal clusters, we find that boredom and frustration exhibit reliable relationships with learning, but only in some clusters. Within several of the clusters, boredom was negatively correlated with learning (cf. Karumbaiah et al., 2021a). Curiously, a relatively strong positive relationship between high frustration and student learning was found for Cluster 4, aligning to previous findings by Pardos et al. (2014); from our analysis, students with high frustration were twice as likely to answer the assessment item correctly. Distinctive from Cluster 1 where we see similarly-long sequence lengths, frustration emerges as a positive predictor of learning; in other clusters it is associated negatively or weakly with learning. The findings across clusters align with other prior work (Baker et al., 2010). The temporal clusters we have generated provide means to identify this "productive" frustration previously theorized in (D'Mello & Graesser 2012). It is also worth noting that this positive relationship between frustration and our learning outcome is found in the cluster with the lowest overall performance on the assessment. Cluster 4 appears to characterize students who are struggling to learn the material, where frustration is an indicator of the most successful of these students; perhaps for a struggling student, the only way to succeed is to persist through that struggle and the frustration it generates (e.g. D'Mello & Graesser 2011).

As the clusters were generated using multivariate, continuous-valued estimates of affect over time, it is worth highlighting that we see at most one affective state emerging as a reliable predictor of learning within each cluster. These significant states were rarely the highest estimate among the detector outputs; in other words, if one were to convert the estimates to discrete labels by simply taking the state corresponding with the highest estimate, subtle affective trends would not have been possible to identify. However, the lack of a discrete label makes it more difficult to interpret the patterns that emerged (cf. (Andres et al., 2019). Retaining full probability information capturing more complex relationships within and across affective states, but sacrifices a level of interpretability as a result.

We have, then, a trade-off between characterizing complex affective patterns and interpretability. Future work may overcome this by exploring better ways of describing the patterns within each cluster. We approached this initially by dichotomizing the means and variances of each affective state, but future work could consider additional measures of affective dynamics and chronometry, to better represent the properties being captured by the temporal clustering.

Another potential limitation of the current work is in the use of sensor-free detectors to measure affect. While the detectors have been shown to exhibit high performance in predicting human-recorded observation labels (Botelho et al., 2017), (further) improvements to these detectors could provide more accurate insights into patterns of student affect. It is also important to emphasize that the detector model used in this work was trained on a dataset collected between 2010 and 2012, as reported in Botelho et al. (2017). It is uncertain whether later changes to the system or user population might lead to lower detector reliability (cf. Levin et al., 2022). Future work should examine longitudinal and contextual factors that may affect detector performance, to identify and mitigate potential threats to validity.

Overall the findings of our analyses provide new insights into the relationship between the dynamics and chronometry of affect and student learning. The subtle complexities of affect, in a temporal sense, are not currently well-understood, but the methods presented in this work provide a means to better explore this. This method may be particularly useful for better understanding affective and behavioral patterns exhibited by students and eventually we may be able to use these clusters to identify what recommendations to give students (for instance, who should persist and who should seek help). In this fashion, these analyses may be a step towards building better supports for students that avoid negative interactions and experiences that may lead to poorer learning outcomes.

Acknowledgements

We would like to thank multiple agencies for supporting this work including the National Science Foundation (#1903304) and Schmidt Futures.

References

Andres, J. M. A. L., Ocumpaugh, J., Baker, R. S., Slater, S., Paquette, L., Jiang, Y., ... & Biswas, G. (2019, March). Affect sequences and learning in Betty's Brain. In *Proc. of the 9th Int'l Conf. on Learning Analytics & Knowledge* (pp. 383-390).

Baker, R. S., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *Int'l Journal of Human-Computer Studies*, *68*(4), 223-241.

Baker, R. S., Ocumpaugh, J. L., & Andres, J. M. A. L. (2020). BROMP quantitative field observations: A review. Learning Science: Theory, Research, and Practice. New York, NY: McGraw-Hill.

Berndt, D. J., & Clifford, J. (1994, July). Using dynamic time warping to find patterns in time series. In *KDD workshop* (Vol. 10, No. 16, pp. 359-370).

Bosch, N., & D'Mello, S. (2017). The affective experience of novice computer programmers. *Int'l Journal of Artificial Intelligence in Education*, *27*(1), 181-206.

Botelho, A. F., Baker, R. S., & Heffernan, N. T. (2017, June). Improving sensor-free affect detection using deep learning. In *Proc. of the 2017Int'l Conf. on Artificial Intelligence in Education* (pp. 40-51).

Botelho, A. F., Baker, R. S., Ocumpaugh, J., & Heffernan, N. T. (2018). Studying Affect Dynamics and Chronometry Using Sensor-Free Detectors. *In Proc. of the 11th Int'l Conf. on Educational Data Mining,* 157-166.

Brom, C., Buchtová, M., Šisler, V., Děchtěrenko, F., Palme, R., & Glenk, L. M. (2014). Flow, social interaction anxiety and salivary cortisol responses in serious games: A quasi-experimental study. *Computers & Education*, *79*, 69-100.

Conati, C., Chabbal, R., & Maclaren, H. (2003, June). A study on using biometric sensors for monitoring user

emotions in educational games. In *Workshop on assessing and adapting to user attitudes and affect: Why, when and how*.

Baker, R. S., Gowda, S. M., Wixon, M., Kalka, J., Wagner, A. Z., Salvi, A., ... & Rossi, L. (2012). Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. *In Proc. of the 5th Int'l Conf. on Educational Data Mining,* 126-133.

D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, *25*(7), 1299-1308.

D'Mello, S., Chipman, P., & Graesser, A. (2007). Posture as a predictor of learner's affective engagement. In *Proc. of the Annual Meeting of the Cognitive Science Society* (Vol. 29, No. 29).

D'Mello, S., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145-157.

D'mello, S. K., Craig, S. D., Witherspoon, A., McDaniel, B., & Graesser, A. (2008). Automatic detection of learner's affect from conversational cues. *User modeling and user-adapted interaction*, *18*(1), 45-80.

Fancsali, S. (2014, July). Causal discovery with models: behavior, affect, and learning in cognitive tutor algebra. In *Proc. of the 7th Int'l Conf. on Educational Data Mining,* 28-35.

Gong, L., Liu, Y., & Zhao, W. (2019, March). Dynamics of Emotional States and Their Relationship with Learning Outcomes during Learning Python with MOOC. In *Proc. of the 2019 7th Int'l Conf. on Information and Education Technology* (pp. 71-76).

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. *Int'l Journal of Artificial Intelligence in Education*, *24*(4), 470-497.

Hutt, S., Grafsgaard, J. F., & D'Mello, S. K. (2019, May). Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. In *Proc. of the 2019 CHI Conf. on human factors in computing systems* (pp. 1-14).

Judkins, D. R., & Porter, K. E. (2016). Robustness of ordinary least squares in randomized clinical trials. *Statistics in Medicine,* 35(11), 1763-1773.

Karumbaiah, S., Baker, R., Tao, Y., & Liu, Z. (2022, March). How does Students' Affect in Virtual Learning Relate to Their Outcomes? A Systematic Review Challenging the Positive-Negative Dichotomy. In *LAK22: 12th Int'l Learning Analytics and Knowledge Conf.* (pp. 24-33).

Karumbaiah, S., Baker, R. B., Ocumpaugh, J., & Andres, A. (2021a). A re-analysis and synthesis of data on affect dynamics in learning. *In IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3086118.*

Karumbaiah, S., Lan, A., Nagpal, S., Baker, R. S., Botelho, A., & Heffernan, N. (2021b). Using Past Data to Warm Start Active Machine Learning: Does Context Matter? In *LAK21: 11th Int'l Learning Analytics and Knowledge Conf.*,151-160.

Lemire, D. (2009). Faster retrieval with a two-pass dynamic-time-warping lower bound. *Pattern recognition*, *42*(9), 2169-2180.

Levin, N., Baker, R.S., Nasiar, N., Fancsali, S., Hutt, S. (2022) Evaluating Gaming Detector Models for Robustness Over Time. *In Proceedings of the 15th Int'l Conf. on Educational Data Mining*.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11), 1857-1874.

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R. (2013, July). Sequences of frustration and confusion, and learning. *In Proc. of the 6th Int'l Conf. on Educational Data Mining, 59-66*.

Montero, P., & Vilar, J. A. (2015). TSclust: An R package for time series clustering. *Journal of Statistical Software*, *62*, 1-43.

Ocumpaugh, J., Baker, R., Gowda, S., Heffernan, N., & Heffernan, C. (2014). Population validity for Educational Data Mining models: A case study in affect detection. *British Journal of Ed. Technology*, *45*(3), 487-501.

Pardos, Z.A., Baker, R.S., San Pedro, M.O.C.Z., Gowda, S.M., Gowda, S.M. (2014) Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. *Journal of Learning Analytics, 1* (1), 107-128.

Ratanamahatana, C. A., & Keogh, E. (2004, April). Making time-series classification more accurate using learned constraints. In *Proc. of the 2004 SIAM Int'l Conf. on data mining* (pp. 11-22).

Rodrigo, M. M. T., Baker, R. S., Jadud, M. C., Amarra, A. C. M., Dy, T., Espejo-Lahoz, M. B. V., ... & Tabanao, E. S. (2009, July). Affective and behavioral predictors of novice programmer achievement. In *Proc. of the 14th Conf. on Innovation and technology in computer science education* (pp. 156-160).

Wang, Y., & Heffernan, N. T. (2014, June). The effect of automatic reassessment and relearning on assessing student long-term knowledge in mathematics. In *Int'l Conf. on Intelligent Tutoring Systems* (pp. 490-495).

Xiong, X., Li, S., & Beck, J. E. (2013, May). Will you get it right next week: Predict delayed performance in enhanced ITS mastery cycle. In *The 26th Int'l FLAIRS Conference*

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Engagement Estimation using Time-series Facial and Body Features in an Unstable Dataset

Xianwen ZHENG[1*], Minh-Tuan TRAN[1], Koichi OTA[1], Teruhiko UNOKI[2] & Shinobu HASEGAWA[1*]
[1*] *Japan Advanced Institute of Science and Technology, Japan*
[2] *Kansai Gaidai University, Japan*
*zhengxianwen@*jaist.ac.jp, hasegawa@jaist.ac.jp*

Abstract: The global education sector has been deeply shaken by COVID-19 and forced to shift to an online teaching model. However, the lack of face-to-face communication and interaction in online learning is critical to high-quality teaching and learning. Research on engagement is a crucial part of solving this problem. Because engagement is of time-series data with an ongoing change, research datasets used for engagement analysis need a certain preprocessing method to capture time-series related engagement features. This research proposed a novel deep learning preprocessing method for improving engagement estimation using time-series facial and body information to restore traditional scenes in online learning environments. Such information includes head pose, mouth shape, eye movement, and body distance from the screen. We conducted a preliminary experiment on the DAiSEE dataset for engagement estimation. We applied skipped moving average in data preprocessing to reduce the influence of the extracted noises and oversampled the low engagement level data to balance the engaged/unengaged data. Since engagement is continuous and cannot be captured at a particular instant in time or single images, temporal video classification generally performs better than static classifiers. Therefore, we adopted long short-term memory (LSTM) and Quasi-recurrent neural networks (QRNNs)sequence models to train models and achieved the correct rate of 55.7% (LSTM) and 51.1% (QRNN) using the original key points extracted from OpenPose. Finally, we proposed the optimization structure network achieved the engagement estimation correct rate of 68.5% in proposed LSTM models and 64.2% in QRNN models. The achieved correct rate is 10% higher than the baseline in the DAiSEE dataset.

Keywords: Engagement, online learning, time-series analysis, oversampling, LSTM, QRNN

1. Introduction

Online learning has been transformed in recent years thanks to learning management systems and web resource development (Hurlbut, 2018). Moreover, Covid-19 has deeply shaken the global education sector. This situation has prevented face-to-face teaching and forced schools to shift to an online teaching model. Online learning frees learners from space and time constraints. Therefore, it has become popular and widespread. However, there are some problems with online learning. Sometimes course content is theoretical, and less practice makes the learners who take online courses unable to concentrate on the courses or maintain high learning efficiency (Wong, 2019; Hasegawa, 2020). In addition, learning activities differ in space and time from learner to learner, which causes difficulty in providing individual learners with appropriate support. In this situation, the research on engagement has been brought to the limelight. From an educational point of view, engagement is a mental state that can help learners feel positive and realize high-quality learning (Kage, 2013) and positively influence social, cognitive, and personality. Engagement has two characteristics, continuity and change. Understanding these expressions with the characteristics helps support learners' engagement and opens opportunities to improve learning quality.

Although datasets with various machine learning methods have an essential role in recent studies on engagement estimation, some issues remain in the existing engagement study datasets. In

active learning, learners' learning motivation is usually high, and engaged time is more than not engaged. Hence, the number of engaged/unengaged data is imbalanced (Chang, Zhang, Chen, & Liu 2018), which is harmful to estimation in our preliminary research. The insufficient scale dataset is another vital issue in engagement estimation research (Hasegawa, 2020; Chang, 2018). Based on the characteristics of engagement in education, the research dataset needs to have the elements of time series and changes. At the same time, there are requirements for the race, culture, learning content, and data labeling methods of the experimental participants. It is unrealistic in the short term to collect a large amount of dataset with engagement labels to improve the estimation accuracy. Thus, the existing engagement datasets that fit this study are precious.

This article proposed a novel deep learning preprocessing method to improve estimation accuracy by merging two pre-trained models trained on different time-series facial and body features. We designed the facial and body expression features using the key points extracted from OpenPose. Oversampling and skipped moving average methods were adopted to restructure the data to solve the imbalanced dataset problem. Long Short-Term Memory (LSTM) and Quasi-Recurrent Neural Networks (QRNN) models (Bradbury, Merity, Xiong, & Socher 2017) are applied to estimate the learners' engagement. Finally, we built a deep learning network model that combined the trained two models using different conditions.

2. Related Work

Gupta et al. presented a multi-label video classification dataset for engagement research (Gupta, D'Cunha, Awasthi, & Balasubramanian 2016). There are four affective states of boredom, confusion, engagement, and frustration, with four levels of crowdsourced labels for each video recorded in an e-learning environment. They established benchmark results on this dataset with temporal and static models based on the Convolutional Neural Networks (CNN) model.

Dhall et al. introduced a dataset for learner engagement detection and localization that contains 78 subjects aged 19-27 (Dhall, 2019). The dataset contains videos about the subjects watching MOOC materials, learner videos are then labeled manually into 4 engagement levels. Multiple Instance Learning Approach and Deep Multi Instance Learning were proposed to predict the engagement level of individuals and presented baseline results.

Karimah et al. proposed four different LSTM models to identify the most effective combination of pre-processing methods for an imbalanced distribution dataset (Karimah, Unoki, & Hasegawa, 2021). The applied pre-processed methods are under-sampling, oversampling, normalization, dimensional reduction, and the combination scenarios where the pre-processed methods are ordered.

We also previously proposed transfer learning on time-series face and body features captured by built-in PC cameras to improve the engagement estimation on small and irregularly wild datasets. We composed a time-series dataset with engagement level labels of online tasks. To solve the insufficient data issue, we pre-trained a long short-term memory sequence deep learning model on the DAiSEE dataset and transferred the trained model to share learned features extraction and retrain the composed model on the new dataset.

The existing engagement datasets, DAiSEE and EmotiW, are unbalanced. They have a limited amount of low-level engagement data. Moreover, based on the characteristics of ongoing and changing engagement, the method in time-series is the trend of future work, and the study in the time- series method still needs to be developed. Furthermore, the insufficient dataset is one of the most significant challenges in time-series engagement estimation/detection study. Therefore, it is necessary to make the most of the existing data.

3. Dataset

DAiSEE dataset, a publicly affective states video classification dataset in the e-environment for engagement, boredom, confusion, and frustration state recognition in e-shopping, e-healthcare, and e-learning, was used in our experiment (Gupta, D'Cunha, Awasthi, & Balasubramanian 2016). They captured 9068 video snippets from 112 (80 male, 32 female) participants of Asian race aged 18-30.

Moreover, each snippet has 10sec with a unique identification number for marking it. Each video in the dataset was labeled with one affective state of engagement, and each label was arranged at four levels: (1) very low, (2) low, (3) high, and (4) very high. There is a baseline for the static (Frame classification/prediction) and dynamic (temporal video classification) experiments by the standard video and deep learning models in the DAiSEE dataset. They achieved 48.6% (C3D Training), 56.1% (C3D Fine Tuning), and 57.9% (LRCN) accuracy in temporal video classification

We used the facial and body contour key points extracted from OpenPose and the original four levels of engagement labels in the DAiSEE dataset to train a Long Short-Term Memory (LSTM) sequence deep learning network model. From the result of the preliminary experiments, we found that the imbalance of the very low and low engaged data affected prediction results, especially, of the low-engaged level. Besides, from the DAiSEE dataset, very low and low label videos are very similar (Dewan, Lin, Wen & Uddin, 2018). Thus, the four levels labels are rearranged into three levels: (0-1) low engaged, (2) normally engaged, and (3) high engaged.

4. Features

The recorded videos in the engagement research dataset mainly contain upper body information due to online education characteristics. Therefore, we extracted the upper body key points, shoulders, elbows, wrists, and the whole facial key points using OpenPose. Some studies mapped the body posture and movement into affects to investigate the relation between bodily expressions and specific affective states. They found that posture and motion types of information are two separated pathways in the brain for recognizing biological information. Both are useful and important for perceiving affect from body expressions (Kleinsmith & Bianchi-Berthouze, 2013). Moreover, the leaning direction, the openness of the body, and the head position play an important role in affect perception (James, 1932). Therefore, following suggestions of other research, we design facial and body estimation engagement features. Finally, the designed features include original body key points, head pose, eye information, mouth shape, and body movement with 32 dimensions.

5. Data Pre-processing

5.1 Skipped Moving Average

Our previous experiment (Zheng, Hasegawa, Tran, Ota, & Unoki, 2021) showed that the existing library OpenPose generated noise and lost video frames affecting the final estimated result when extracting external information. In addition, the number of frames per second in the sample videos, ranging from 20 to 30 fps, is too much for our requirements. Due to the characteristic of engagement ongoing and changing, we aim to find the change in learners' engagement and give support when learners are low engaged. Thus, the features per second are enough, not per frame.

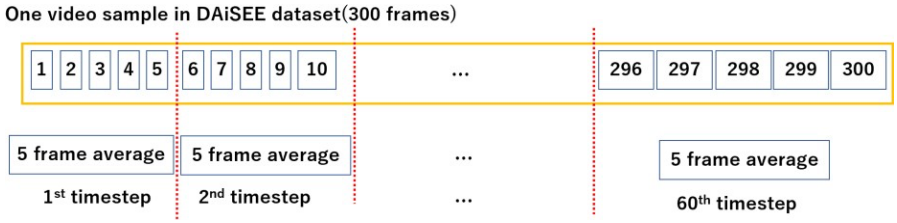


Figure 1. The Skipped Moving Average Method.

We borrowed the idea from the Moving Average method to meet our time-series classification studies and dubbed it Skipped Moving Average (SMA). The skipped moving average approach is illustrated in Figure 1. A video has 300 frames in the DAiSEE dataset. We used the estimated average value of a set period (5 frames) for the one-time step's training data. Then, for deep learning models, one video will have sixty-time steps. To not impact the sample size of the high-level engagement data, we only take the first timestep in every six-time step in the high-level sample data.

5.2 Over Sampling

An imbalance in sample size between the engaged and unengaged data affects the estimation accuracy. Re-arranged the four-level labels into three is one of the solutions in machine learning research but cannot achieve the desired effect. Therefore, we adopt oversampling to resample the low- engaged data.

In the DAiSEE dataset, 30 fps is too much for our requirements. Engagement is a continuously changing mental state and cannot be present as a moment of external expression. Therefore, we set 5 frames as the fixed period and calculated the fixed period average value in all the settings. Besides, we got six times sampling data by shifting one by one average value in the low engagement level data. The low-level engagement data become six times than before.

6. Experimentation

To improve the performance of affective and emotional engagement estimation, we optimized LSTM and QRNN network structures trained on original key points and designed features. In the four experiments, we used the extracted original facial and body key points from OpenPose and the designed features to train LSTMs and QRNNs deep learning models, respectively. Then, we used the proposed optimized structure networks to combine the trained original key points and designed feature models (Figure 2).

The extracted original key points and designed features are different conditions training features. We used the different condition data to examine our proposed methods. First, we used the original Key Point extracted from OpenPose to train LSTMs and QRNNs deep learning models. The calculated moving average (MA) and oversampled (OS) data trained LSTMs and QRNNs deep learning models to compare the performance of the proposed two methods. Finally, we trained LSTM and QRNN models using the combined moving average (MA) value and oversampled (OS) data as the phased results. In data processing, we normalized pre-processed training data range from -1 to 1 and divided it into 8:2 for training and validation set. All experiment setting patterns were trained by a two-layer model containing one hidden LSTM layer and a dense output layer with three cell nodes with a SoftMax function. We tried 16, 32, 64, 128, 256, and 512 cells for the hidden LSTM layer to find our experiment's best deep learning model hyperparameters. Moreover, each cell setting is trained by 100, 200, 300, and 500 epochs, respectively.

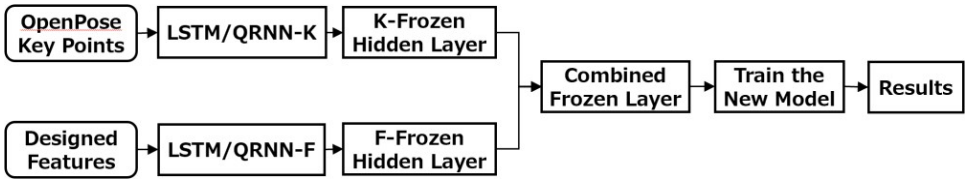


Figure 2. The Optimized Structure Network Process.

The trained LSTM and QRNN deep learning sequence models using original key points and designed features will be reused in this step. We freeze the hidden layers in the pre-trained models. Then, the frozen layers were combined to build new LSTM/QRNN deep learning models. The new model contains the combined frozen layers and a dense output layer with three cell notes, the same as the previous step. Finally, we examine the proposed model on the processed data from the DAiSEE dataset. There are 32 cell notes in the pre-trained hidden layers that are trained by different condition features. The dense output layer has three cell notes and a SoftMax activation function since the combined models must be retrained with the same data. The pre-trained and combined new models are sitting early stopping to avoid overfitting. Table 1 and 2 shows the results that follow the designed experiment. The results proved that our proposed methods are effective and significantly improved in low-level engagement estimation.

Table 1. *LSTM Experiment results Summary*

Engagement Levels	0	1	2	Accuracy
Original Key Points	0.05	0.74	0.41	0.557
Key Points + MA + OS	0.90	0.60	0.50	0.650

Features + MA + OS	0.93	0.61	0.47	0.652
Proposed Model	0.91	0.66	0.52	0.685

Table 2. *QRNN Experiment results Summary*

Engagement Levels	0	1	2	Accuracy
Original Key Points	0.00	0.93	0.07	0.511
Key Points + MA + OS	0.74	0.57	0.46	0.582
Features + MA + OS	0.73	0.56	0.50	0.586
Proposed Model	0.78	0.68	0.47	0.642

From the proposed sequence deep learning model classification error, we found that eye information, like eye gaze, wink, and eye movement, are essential features for our engagement analysis. The body action and movement improve the performance of sequence deep learning models, but they might bring some partially redundant information that affects estimation results. In some sample videos, the learners' engagement is high, whereas some partially redundant body features may be added and incorrectly lead to the facial features. For instance, the learners touch the face, push glasses, and eat. In other cases, the learners' body features usually do not change in low-level engagement data, whereas learners' eyes drift. Thus, the features in low-level engagement are hard to recognize. Therefore, we need to standardize time series body features and refine the designed features to improve our proposed optimization structure models' accuracy. Individual engagement features introduce another topic to be considered in future work. Besides, normal and high engagement are similar, which are difficult to distinguish even for humans.

7. Conclusion

7.1 Conclusion

In conclusion, we introduced the long short-term memory (LSTM) and Quasi-recurrent Neural Networks (QRNNs) deep learning sequence models to estimate learners' engagement using time- series facial and bodily information. We described the engagement from an online learning education point of view and defined the learners' affective/emotional engagement as our estimation object in this experiment. The training data are body key points extracted from OpenPose and the designed-up body features using the body key points. We applied the moving average method to relieve the widespread noise issue generated by OpenPose and oversampled the low engagement data to balance the number of engaged/unengaged data. To let the long short-term memory (LSTM) and Quasi- recurrent Neural Networks (QRNNs) deep learning sequence models work well on the DAiSEE dataset, we combined the sequences models trained by key points and our designed features.

We achieved the engagement estimation correct rate of 0.685 in proposed LSTM models and 0.642 in QRNN models. The achieved correct rate is 10% higher than the baseline in the DAiSEE dataset. The engagement estimation accuracy has been improved effectively and demonstrated the excellent potential of our proposed preprocessing method and optimized structure deep learning sequence model. In addition, the proposed time-series data processing method has been successful in our previous investigations. These findings compare favorably to the state-to-the-art in engagement estimation/prediction studies.

7.2 Further Work

This article described the engagement estimation experiment and achieved approximately 68% accuracy. However, our ultimate purpose is to assist learners in achieving high-quality learning. How

can we help learners become more engaged? Affective/emotional, behavioral, and cognitive engagement are three components of engagement found in many different elements of online learning. Therefore, we proposed an engagement support system that can provide feedback on the learner's engagement at a reasonable time and save learning history for further research. The next step aims to analyze learners' engagement using recorded time-series expressions and body features to create a learners' engagement support system for increasing online learning quality. Affective and emotional engagement encompasses learners' feelings in the learning process. Thus, the support system will immediately give feedback on the engagement analysis results to both teachers and learners to improve engagement. At the same time, the system saves the estimation results for checking learners' engagement results history for asynchronous online learning courses.

Additionally, cognitive engagement plays a highly strategic role in the learning process. The system analyzes the collected engagement records and makes recommendations for increasing learners' cognitive engagement. Based on the learners ' engagement estimation feedback, professors can alter teaching progress and lecture topic difficulty. Also, students can enhance their self-control and engagement. In summary, the system analyzes the learners' affective/emotional and behavioral engagement in each online course to plan to improve support self-regulated learning skills for maintaining their learning motivation and ultimately achieve the learning goals.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20H04294 and Photron limited.

References

Chang, C., Zhang, C. Chen, L., & Liu, Y. (2018). An Ensemble Model Using Face and Body Tracking for Engagement Detection, *The 20th ACM International Conference on Multimodal Interaction* (pp. 616-622).

Cao, Z., Hidalgo, G., Simon, T., Wei, S., & Sheikh, Y. (2017). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *The IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7291-7299).

Dewan, M. A. A., Lin, F., Wen, D., & Uddin, Z. (2018). A deep learning approach to detecting engagement of online learners. *2018 IEEE Smart World, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation* (pp. 1895–1902).

Dhall, A. E. (2019). EmotiW 2019: Automatic Emotion, Engagement and Cohesion Prediction Tasks. *2019 International Conference on Multimodal Interaction (ICMI '19)* (pp. 14-18).

Gupta, A., D'Cunha, A., Awasthi, K., & Balasubramanian, V. (2016). DAiSEE: Towards User Engagement Recognition in the Wild. arXiv preprint arXiv:1609.01885.

Hurlbut, Amanda R. (2018). Online vs. traditional learning in teacher education: a comparison of student progress. *American Journal of Distance Education*, *32*(4), 248-266.

Hasegawa, S., Hirako, A., Zheng, X. W., Karimah, S. N., Ota, K., & Unoki, T. (2020). Learner's Mental State Estimation with PC Built-in Camera. *Proceedings of learning and Collaboration Technologies. Human and Technology Ecosystems* (pp. 165-175).

James, W. T. (1932). A study of the expression of bodily posture. *Journal of General Psychology*, *7*(2), 405- 437.

Kage, M. (2013). *Theory of motivation to learn Motivational educational psychology*. Kaneko Bookstore, Tokyo.

Bradbury, J., Merity, S., Xiong, C., & Socher, R. (2017). Quasirecurrent Neural Networks. *Neural and Evolutionary Computing*.

Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions On Affective Computing*, *4*(1), 15-33.

Karimah, S. N., Unoki, T., & Hasegawa, S. (2021). Implementation of Long Short-Term Memory (LSTM) Models for Engagement Estimation in Online Learning. *2021 IEEE International Conference on Engineering, Technology & Education (TALE)* (pp. 283-289).

Wong, J., Baars, M., Hsi, S., Davis, D. Zee, T. V. D., Houben, G., & Paas, F. (2019). Supporting Self-Regulated Learning in Online Learning Environments and MOOCs: A Systematic Review. *International Journal of Human–Computer Interaction, 35*(4-5), 356-373.

Zheng, X. W., Hasegawa, S., Tran, M. T., Ota, K., & Unoki, T. (2021). Estimation of Learners' Engagement Using Face and Body Features by Transfer Learning. *International Conference on Human-Computer Interaction* (pp. 541-552).

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Investigating How Achievement Goals Influence Student Behavior in Computer Based Learning

J. M. Alexandra L. ANDRES[a*], Stephen HUTT[a], Jaclyn OCUMPAUGH[a], & Ryan S. BAKER[a]
[a]*Graduate School of Education, University of Pennsylvania, USA*
*aandres@upenn.edu

Abstract: This study uses a mixed methods approach to explore relationships between goal achievement orientation and changes in student achievement, behavior, and affect while using Betty's Brain. Qualitative coding was applied to student interviews to identify either performance or mastery goal approaches applied during their interaction with the game. A combination of quantitative methods were used to determine the influence and strength of goal orientations on student outcomes and interviews were thematically examined for patterns across student experiences. Findings from this study imply that students who communicate a goal approach (either performance approach or mastery approach) are more likely to experience confusion and engage in review-related behaviors such as reading informational resources and noting the accuracy of their responses.

Keywords: Education, achievement goals, human-computer interaction, affect and behavior

1. Introduction

Achievement Goal Theory seeks to understand how differences in students' goals influence their learning behaviors (Elliot & McGregor, 2001). Elliott & Dweck, (1988) identified two primary goals in achievement situations: performance goals (which involve attempts to maintain positive perception of one's own abilities) and mastery goals (which involve the pursuit of new knowledge or skill mastery mastering a new task). They also sought to classify these goals in terms of valance, a classification that proved useful in a wide range of research (Pintrich, 2000; Senko, Hulleman, & Harackiewicz, 2011). Negatively-valanced orientations (i.e., avoidance goals) have been shown to negatively impact achievement, educational interest, study habits, help-seeking, and increase anxiety (Senko et al., 2011; Wolters, 2004). However, positively-valanced goal orientations (i.e., approach goals) are associated with positive outcomes, including reduced anxiety, increased effort and persistence, increased performance, and help-seeking behavior (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002; Niepel, Brunner, & Preckel, 2014; Wolters, 2004).

In general, mastery goals may seem preferable to performance goals. Considerable research shows that mastery approach goals have the potential to improve cognitive, motivational, and behavioral processes, including improvements related to anxiety and help-seeking behavior (Wolters, 2004). However, recent findings suggest that mastery avoidance goals are extremely rare in real-world learning (Skaalvik, 2018) and performance-approach goals have also been associated with positive outcomes (such as increased effort, persistence, and performance; Harackiewicz et al., 2002; Niepel et al., 2014). In fact, the behavior and dynamics around these goals are often moderated by contextual factors such as motivation, reception of feedback, autonomy, interest, and self-regulated learning (Wilson & Narayan, 2016). The complexity of the patterns around achievement goals suggest that it may be valuable to investigate their emergence in greater detail.

This study uses a mixed-methods approach to examine goal achievement in interviews conducted with students as they interact with a computer-based science learning environment. Specifically, we use data from Betty's Brain, where a range of both positive and negative affective states have already been studied (Andres et al., 2022; Hutt et al., 2021). In past studies of these interviews, we examined the relationship between affect and anxiety as well as the differences in student perceptions in response to frustration (Andres et al., 2022; Baker et al., 2021). In this article, we examine how students' achievement goals interact with students' affect, metacognitive and problem-solving strategies, and

learning outcomes. We leverage a combination of open-ended interviews, log data, and affect detectors to analyze how achievement goals may be impacting student experiences in this system.

2. Methods

2.1. Betty's Brain

Betty's Brain is an open-ended, computer-based learning system that implements a learning-by-teaching model. Students teach a virtual agent named "Betty" by creating a causal map of scientific processes (e.g., thermoregulation or climate change; Leelawong & Biswas, 2008). Students choose how to navigate a variety of learning resources as they build their maps and quiz Betty. These quizzes are used to demonstrate Betty's "learning" and are graded by a mentor agent, Mr. Davis, who scaffolds students' learning and teaching endeavors (Biswas, Segedy, & Bunchongchit, 2016).

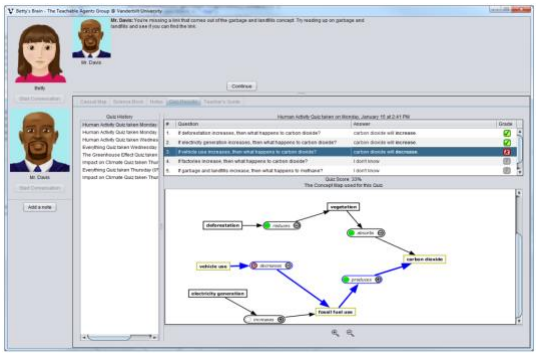


Fig.1. *Screenshot of quiz results and the causal map Betty uses to answer quiz questions*

2.2. Study Design

Data was collected at an urban middle school in the Eastern United States, from 99 sixth-graders (approximately 11-12 years old) who used Betty's Brain in their regular science classes. Their school's population is approximately 60% White, 25% Black, 10% Asian, and 5% Hispanic, with approximately 10% enrolled in the free/reduced lunch program. Individual demographics were not collected. The same students used Betty's Brain to complete two science inquiry scenarios conducted in December 2018 (study 1) and February 2019 (study 2). Data were collected over the course of seven school days in both time periods. Students and their parents were consented/assented prior to the study. On the first day of the study, students completed a 30-45-minute paper-based pre-test. On day 2, students participated in a 30-minute training session that familiarized them with the learning goals and user interface of the software. Following the pre-test and training, students used the Betty's Brain software on days 2 through 6, for approximately 45-50 minutes each session, using concept maps to teach Betty about the causal relationships about science content. On day 7, students completed a post-test that was identical to the pre-test, in order to assess their learning.

Students used Betty's Brain to complete a causal map about climate change (study 1) or thermoregulation (study 2). During both studies, interviews explored how students were working with the system during specific experiences. As students interacted with Betty's Brain, automatic detectors of educationally relevant affective states (Jiang et al., 2018) and behavioral sequences (Munshi et al., 2018), embedded in the software, identified specific affective patterns or theoretically aligned behavioral sequences. Jiang et al.'s, (2018) real-time affect detection—which had been previously integrated into Betty's Brain included BROMP-based detectors (Ocumpaugh, Baker, & Rodrigo, 2015) of boredom, confusion, engaged concentration, delight, and frustration. Each detector generated an affect prediction (a probability between 0 and 1) every 20 seconds based upon the student's interactions.

These detectors were used to select which student to interview at a specific time, with a focus on shifts in affect or meaningful behavioral patterns indicating self-regulated learning (Hutt et al., 2021). At the end of study 2, science anxiety surveys were administered alongside questions about students' perceptions of difficulty and familiarity with the topic and questions about Mr. Davis. Between the two

studies, minor changes were made to Betty's Brain, including small changes to make Mr. Davis seem more polite (Ocumpaugh et al., 2021). All other procedures were identical.

2.3. Student Interviews and Interview Coding

A total of 594 interviews (358 in study 1; 236 in study 2) were obtained. Interviewers sought a helpful but non-authoritative approach when speaking with students, often asking students what their strategies were (if any) for getting through the system. Questions were added to elicit more information as new patterns and information emerged. Overall, students were encouraged to speak about what they wanted and their feedback about their experience with the software. Audio files from the interviews were collated and stored on a secure file management system available only to the research team members. Three members of the research team manually transcribed the interviews in an agreed-upon format. Metadata, including timestamps and recording IDs, were preserved, but data was de-identified (i.e., each student was assigned a unique alphanumeric identifier).

The code development process followed the recursive, iterative process used in previous research (Weston et al., 2001) that includes seven stages: conceptualization of codes, generation of codes, refinement of the first coding system, generation of the first codebook, continued revision and feedback, coding implementation, and continued revision of the codes (Weston et al., 2001). The conceptualization of codes included a review of related literature to capture meaningful experiences relevant to the study's research questions. Using grounded theory (Charmaz, 1983), a method that is appropriate for the kind of open-ended interviews where students are being asked to interpret their own experiences, we worked with the lead interviewer (the third author) to identify categories that were (1) relevant to either affective theory (e.g., D'Mello & Graesser, 2012) or self-regulated learning theory (e.g., Winne & Hadwin, 1998) and (2) likely to saliently emerge in the interviews.

A coding manual was created through the iterative refinement of this coding scheme, and multiple coders achieved an acceptable interrater reliability (Cohen's Kappa > .6) before coding the transcripts. The coders met and clarified any concerns to avoid misinterpretation or miscoding of the data. Although a number of coding schemes were developed in this manner (Baker et al., 2021), the current study focuses on codes that align with Achievement Goal Theory: performance approach or mastery approach. (Note that while the coding manual included performance avoidance, this code was rarely applied and was excluded from analysis.) As these qualitative codes are not mutually exclusive, a single interview may be coded under both categories. However, as Table 2 shows, this kind of overlap was infrequent.

Table 2. *Summary of coded interviews and unique participants associated with the interviews*

	Study 1 - December				Study 2 - February			
	Perf	Mas	Both	Neither	Perf	Mas	Both	Neither
Interviews	38	59	6	232	46	46	10	154
	(11.76%)	(18.27%)	(1.86%)	(71.83%)	(19.49%)	(19.49%)	(4.24%)	(65.25%)
Students	31	42	6	29	33	33	8	38

2.4. Student Interaction Data: Affect and Behavior

In addition to pre-tests, post-tests, and interview data, we also extracted two different types of data from the logs of student interactions with the system. Affect predictions (generated every 20 seconds) were averaged to produce an rate of incidence for each affective state per student which was then used in the quantitative analyses of this study. The second were 26 SRL-related behaviors that could be distilled from the logs. Some behaviors include viewing explanations for questions, taking quizzes, and marking added responses as correct or incorrect.

2.5 Analysis

The analyses conducted in this study consider differences in student achievement (pre-test and post-test scores and gains), behaviors, and affect in response to the presence and kind of achievement goal

approach students communicated. Rates of performance approach and mastery approach were generated for each student by dividing the total number of times each student's interview was coded for mastery approach or performance approach by the total number of times that student had been interviewed (as this varied by student). This calculation, though sufficient for the current analysis, is still limited by the number of interviews a student engages in and may result in lower achievement goal rates overall despite any explicit expression of an achievement orientation or goal within an interview.

We examine the differences between students who communicated a goal approach (combining the performance and mastery approach codes) to students who did not report a goal approach using t-tests and Linear Mixed Effects Models (LMEMs). Benjamini & Hochberg, (1995) post-hoc corrections were applied when a set of multiple tests was run. We also analyzed the interview data qualitatively, taking the interviews of students with at least one instance of a goal approach code, either performance or mastery, and examining these for recurring themes or responses. Pseudonyms were assigned to participants using *http://random-name-generator.info/* without reference to actual student gender.

3. Results

Our quantitative analysis first uses t-tests to explore the differences between those students who expressed either goal orientation to those who had not. In study 1, students with a goal orientation viewed informational resources more often than peers without a goal approach (M=.18, SD=.18, t(90)=-2.087, p=.04), but that effect flipped in study 2, where they viewed these same resources less often than their peers (M=.12, SD=.14, t(74)=1.998, p=.05). In Study 1, students with a goal orientation also exhibited more frustration than their peers (M=.73, SD=.09, t(90)=-2.136, p=.04). However, given the substantial number of behaviors, affective states, and achievement metrics (pre-test scores and learning gains) tested, these findings do not remain statistically significant after post-hoc correction and must be treated as suggestive but inconclusive.

LMEMs were fit to regress the interaction of pre-test scores and goal approach onto each student's total affective state (boredom, frustration, confusion, delight, engaged concentration). In each of these models (5 affective states observed in 2 studies), pre-test scores and goal approach rates were treated as fixed effects and student IDs as random effects. This approach was chosen due to the repeated (multiple sessions per student) and nested structure (sessions nested within students) of the data (Pinheiro & Bates, 2006). Results show that individual rates of goal approach positively predicted confusion in study 2 (β=.16, R^2=.057, p=.035), but did not significantly predict other states (p's = 0.11 to 0.97).

Additional t-tests were conducted to examine the differences among the 27 students who articulated either goal approach, comparing students who had above average rates against those with below average rates (above zero) of either goal approach orientation in their interviews. The results of this analysis indicate that students with below average goal orientation ratios were more likely to mark their causal links as correct in their maps (M=.008, SD=.12, t(27)=3.11, p=.005) and spent more time viewing graded explanations (M=1969.57, SD=2396.70, t(27)=2.21, p=.005). That is, they were more inclined to spend time reading explanations and verifying the accuracy of their responses, but they were also less likely to view explanations for ungraded questions than those with higher ratios (M=.004, SD =.01, t(27)=2.00, p=.05). However, as with the first set of T-tests, these trends do not remain statistically significant after post-hoc correction and must be treated as suggestive though inconclusive.

Our qualitative analyses suggest that students who had articulated a goal approach also discussed using Betty's Brain resources to evaluate the accuracy of their causal maps. Among students who expressed performance goals, Isabel said "I'm trying to figure out which ones are marked correct by asking Betty questions to figure out which ones are correct and which ones I need to mark wrong." This was remarkably similar to the statements by two students who expressed mastery goals during their interviews. For example, Tiffany reports "I'm just trying to go back to the textbook and find out what's happening right there so I can mark things, more things as correct." Likewise, Janie elaborates further on this process: "I just finished reading the science stuff so I'm going to go back to there and give her more quizzes and questions. And then if you see like which ones she (Betty) got wrong. And if she got like some wrong….I read more about that on the map....and [I] mark[ed] what's correct on the map." That is, both students with performance goals and students with mastery goals discuss these strategies in terms of correctness, a terminology that mimics the platform's assessments. Interestingly, students

who had neither goal were more likely to mark themselves correct – but perhaps not as likely to carefully think about their correctness.

Students who had expressed either goal approach in their interviews also referred regularly to their in-game achievement. Among students who expressed a performance approach, Rosemary expressed a conscious effort to improve her scores, explaining that her "best was a 50 [percent] on the cold detection quiz, and now today I have 100's on both and then now I'm working to improve the…the overall score." Likewise, Wendy described "I got like 7 out of 10 right." Students with mastery approaches gave similar information in their interviews. Terrence, who expressed both performance and mastery approach goals, told the interviewer "I was taking a lot of quizzes… I made a new high score… Yeah, so that's a big improvement." Meanwhile, Carl, who exclusively expressed a mastery approach, said, "I started at 0, and then I took [quizzes] just to see what was going on and then I tested and I got two 20's, and then yesterday I got a 38, so it's…a good improvement." These responses suggest that students who express a goal approach also convey progress on their scores that has at times been achieved through their use of the tools within the platform.

These interview responses clarify what is seen in the quantitative analyses, suggesting that students with a goal approach are more diligent in taking note of the accuracy of their responses and of viewing informational resources that would help them to do so. Though neither analysis is fully conclusive, both quantitative and qualitative methods suggest differences in behavior and interaction, such as increased review-like behavior and commitment to score improvement, within the Betty's Brain platform between the students that have communicated a goal approach and those who have not.

4. Conclusions

This study uses a mixed-methods approach to explore relationships between goal achievement orientation and changes in student achievement, behavior, and affect while using Betty's Brain. Qualitative interviews were coded for performance approach or mastery approach orientations and then used for thematic analysis into student experiences. A combination of quantitative methods were used to examine the influence of goal orientations on achievement, affect, and behavior.

While we did not find differences between performance and mastery goals, we did find that a trend where having either of these orientations influenced review-related behavior and experiences of confusion. We also found that the number of times a student's goal orientation emerged in their interviews suggested possible differences in behavior such that students with below average rates of a communicated goal approach were more likely to exhibit review-related behavior and confirm the accuracy of their responses. Similar patterns were also seen in the qualitative analysis where students who had expressed either kind of goal approach discussed using in-game resources to verify the accuracy of their responses and express improvement and progress around their in-game achievement.

A key limitation of this study is the amount of data. While the current data set was sufficient for detecting some changes in relation to goal achievement, it is unable to capture the full range and diversity of influences that relate to these orientations. Future work should seek to collect more interviews and target interviews to more specifically prompt students about achievement goal orientation. However, the fact that goal orientation emerged unprompted suggests that simply having a goal approach causes salient effects; additional data might be sufficient to demonstrate that a students' unprompted expression of these goals is an important indicator, and the degree to which they continue to return to goal-related themes is a measurement that could be useful to both researchers and educators.

Acknowledgements

This work was supported by NSF #DRL-1561567.

References

Andres, J. M. A. L., Hutt, S., Ocumpaugh, J., Baker, R. S., Nasiar, N., & Porter, C. (2022). How Anxiety Affects Affect: A Quantitative Ethnographic Investigation Using Affect Detectors and Data-Targeted Interviews. In B. Wasson & S. Zörgo (Eds.), *Advances in Quantitative Ethnography* (pp. 268–283). Springer.

Baker, R. S., Nasiar, N., Ocumpaugh, J., Hutt, S., Andres, J. M. A. L., Slater, S., … Biswas, G. (2021). Affect-Targeted Interviews for Understanding Student Frustration. *Artificial Intelligence in Education 2021*.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.2307/2346101

Biswas, G., Segedy, J. R., & Bunchongchit, K. (2016). From Design to Implementation to Practice a Learning by Teaching System: Betty's Brain. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-015-0057-9

Charmaz, K. (1983). The grounded theory method: An explication and interpretation. *Contemporary Field Research*, 109–126.

D'Mello, S. K., & Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

Elliot, A. J., & McGregor, H. A. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology*, *80*(3), 501.

Elliott, E. S., & Dweck, C. S. (1988). Goals: an approach to motivation and achievement. *Journal of Personality and Social Psychology*, *54*(1), 5.

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology*, *94*(3), 638–645. https://doi.org/10.1037/0022-0663.94.3.638

Hutt, S., Ocumpaugh, J., Andres, J. M. A. L., Munshi, A., Bosch, N., Baker, R. S., … Biswas, G. (2021). Who's Stopping You? – Using Microanalysis to Explore the Impact of Science Anxiety on Self-Regulated Learning Operations. *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*.

Jiang, Y., Bosch, N., Baker, R. S., Paquette, L., Ocumpaugh, J., Andres, J. M. A. L., … Biswas, G. (2018). Expert Feature-Engineering vs. Deep Neural Networks: Which is Better for Sensor-Free Affect Detection? *Artificial Intelligence in Education*, 198–211. https://doi.org/10.1007/978-3-319-93843-1_15

Leelawong, K., & Biswas, G. (2008). Designing Learning by Teaching Agents: The Betty's Brain System. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208.

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R. S., & Paquette, L. (2018). Modeling learners' cognitive and affective states to scaffold SRL in open-ended learning environments. *UMAP 2018 - Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 131–138. https://doi.org/10.1145/3209219.3209241

Niepel, C., Brunner, M., & Preckel, F. (2014). Achievement goals, academic self-concept, and school grades in mathematics: Longitudinal reciprocal relations in above average ability secondary school students. *Contemporary Educational Psychology*, *39*(4), 301–313.

Ocumpaugh, J., Baker, R. S., & Rodrigo, M. M. T. (2015). *Baker Rodrigo Ocumpaugh monitoring protocol (BROMP) 2.0 technical and training manual*.

Ocumpaugh, J., Hutt, S., Andres, J. M. A. L., Baker, R. S., Biswas, G., Bosch, N., … Munshi, A. (2021). Using Qualitative Data from Targeted Interviews to Inform Rapid AIED Development. *Proceedings of the 29th International Conference on Computers in Education*, 69–74.

Pinheiro, J., & Bates, D. (2006). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Pintrich, P. R. (2000). The Role of Goal Orientation in Self-Regulated Learning. In *Handbook of Self-Regulation* (pp. 451–502). Academic Press. https://doi.org/10.1016/b978-012109890-2/50043-3

Senko, C., Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement Goal Theory at the Crossroads: Old Controversies, Current Challenges, and New Directions. *Educational Psychologist*, *46*(1), 26–47. https://doi.org/10.1080/00461520.2011.538646

Skaalvik, E. M. (2018). Mathematics anxiety and coping strategies among middle school students: relations with students' achievement goal orientations and level of performance. *Social Psychology of Education*, *21*(3), 709-723.

Weston, C., Gandell, T., Beauchamp, J., McAlpine, L., Wiseman, C., & Beauchamp, C. (2001). Analyzing interview data: The development and evolution of a coding system. *Qualitative Sociology*, *24*(3), 381–400. https://doi.org/10.1023/A:1010690908200

Wilson, K., & Narayan, A. (2016). Relationships among individual task self-efficacy, self-regulated learning strategy use and academic performance in a computer-supported collaborative learning environment. *Educational Psychology*, *36*(2), 236–253. https://doi.org/10.1080/01443410.2014.926312

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Hillsdale, NJ: Erlbaum.

Wolters, C. A. (2004). Advancing Achievement Goal Theory: Using Goal Structures and Goal Orientations to Predict Students' Motivation, Cognition, and Achievement. *Journal of Educational Psychology*, *96*(2), 236–250. https://doi.org/10.1037/0022-0663.96.2.236

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

How Much Support is Necessary for Self-Regulated Learning?

Antonija MITROVIC* & Jay HOLLAND
Intelligent Computer Tutoring Group, University of Canterbury, New Zealand
*tanja.mitrovic@canterbury.ac.nz

Abstract: Self-regulated learning is crucial for learning success, and is even of greater importance for online learning as there is less support and feedback available to students. We describe a simple intervention designed to support self-regulated learning in the context of SQL-Tutor, a mature intelligent tutoring system. SQL-Tutor logged data about all interactions students performed, including interactions with the SRL support. Frequency-based analyses did not identify any differences in behaviors of low or high scoring students. However, epistemic network analysis identified significant differences in how students use help available from SQL-Tutor. Students who scored low on the SQL test asked for high-level help (in the form of partial or full solution), copied the provided solutions and submitted them as their own. We conclude that additional support is necessary for students with weak self-regulation skills.

Keywords: self-regulated learning, intelligent tutoring systems, epistemic network analysis

1. Introduction

Self-regulated learning (SRL) is widely acknowledged as being crucial for learning success, and especially so in less structured environments such as higher education and online learning (Davis et al. 2018, Schumacher & Ifenthaler, 2021). In recent years, especially since the COVID-19 pandemic started, much learning is happening in digital learning environments, which provide less structured learning, limited interactions with teachers and other students, and require additional efforts from students for success in learning. High drop-out ratios are found in many studies investigating MOOCs and other self-directed learning environments (Koller et al., 2013; Kizilcec et al., 2015), pointing out the need for support for SRL skills.

Self-regulated learning is defined as an "active, constructive process whereby learners set goals for their learning and then attempt to monitor, regulate, and control their cognition, motivation and behavior guided and constrained by their goals and the contextual features in the environment" (Pintrich, 2000). SRL encompasses not only cognitive and metacognitive skills, but also motivational and affective aspects of learning (Panadero, 2017). Many students have weak SRL skills and do not engage in SRL spontaneously (Sonnenberg & Bannet, 2016). However, research shows that students can develop SRL skills (Kizilcec et al., 2017), leading to new studies on how SRL can be supported.

Digital learning environments log fine-grained information about students' interactions, and allow for new types of interventions in order to improve learning. Several recent papers discuss support for SRL in MOOCs (Davis et al., 2018; Pérez-Álvarez et al., 2020), blended learning and flipped classroom (Moos & Bonde, 2016). There have also been approaches to support SRL skills in open-ended learning environments (Azevedo et al., 2009; Carpenter et al., 2021).

Despite a lot of research on how to support SRL in MOOCs and blended learning, there is limited literature on how to support SRL in Intelligent Tutoring Systems (ITSs). Our paper addresses that gap: we describe a simple intervention to support SRL skills in SQL-Tutor (Mitrovic & Ohlsson, 1999), a mature ITS. Our main research question is to identify how much students would engage with our SRL support. We start by providing a brief literature review on self-regulated learning, and discuss some approaches for supporting SRL skills in online learning. Section 3 presents our intervention and the experiment design. We then present our findings in Sections 4 and 5, followed by the conclusion and limitations of the current work.

2. Related Work

There are several theories of self-regulated learning (Boekaerts, 1999; Zimmerman, 1986, Winne & Hadwin, 1998; Pintrich, 2000), which focus on various processes students use to plan, monitor and manage their learning. These theories agree that the student's ability to regulate learning is of crucial importance for learning effectiveness. Zimmerman (2022) specifies a cyclical SRL model with three phases: forethought, performance and self-reflection. In the forethought phase, the learner engages in goal setting and strategic planning, identifying actions that will enable goal attainment. In the performance phase, while learning, the student uses various learning strategies and monitors progress. In the self-reflection phase, the student reflects upon their learning and regulates the strategies used. Zimmermann (1986) identifies 14 SRL strategies: self-evaluation, organizing and transforming, goal setting and planning, seeking information, keeping records and self-monitoring, environmental structuring, self-consequences, rehearsing and memorizing, seeking peer/teacher/adult assistance, and reviewing. Zimmerman (2002) studied the relationship between SRL and academic achievement, and reported that students who set precise and actionable goals for themselves often reported higher self-awareness and had higher achievements.

Other SRL theories include similar concepts and processes, which students use to select goals and plan learning, monitor and self-reflect on their progress, and regulate their behaviour, motivation and metacognition in order to achieve the goals. Winne and Hadwin's model (1998) consists of four linked phases: task definition, goal setting and planning, enacting study tactics and strategies, and metacognitively adapting studying. Pintrich (2000) model contains four phases: 1) forethought, planning and activation; 2) monitoring; 3) control; and 4) reaction and reflection.

In traditional learning, SRL is supported by actions teachers perform, and also by asking students to write reflective journals and self-assess. The popularity of MOOCs and flipped classroom in recent years has also resulted in a lot of research investigating different learner features and their relationships to learning outcomes. A big problem in MOOCs is disengagement, with typically large percentages of students not engaging with learning materials. Research shows that the lack of SRL skills is a big factor for disengagement in MOOCs (Kizilcec & Halawa, 2015).

Various approaches have been proposed to engage students more. Moos and Bonde (2016) added SRL prompts to videos in the context of flipped classrooms. The planning prompts were questions added at the start of the video, which aimed to support the planning phase of SRL. The monitoring questions were added in the middle of the video, while self-reflection prompts were added at the end of the video. The authors found that students who received the prompts learnt more and also used more SRL strategies in comparison to the control group. Wong et al. (2019) also used prompts to support SRL. In addition to prompts, when learning from videos students may be required to write annotations and answer questions (Mirriahi et al., 2021). In some cases, adaptive nudges are provided to students to make them reflect on their past experiences and plan their future actions (Dimitrova & Mitrovic, 2021; Mohammadhassan et al., 2022). Other approaches include the use of visualizations in order to draw students' attention to important parts of the video, for example in forms of heatmaps (Chatti et al., 2016). Visualizations can be used to present the summary of the student's activities as a learning dashboard (Matcha et al., 2019) or an open learner model (Hooshyar et al., 2020; Bodily et al., 2018), and also to support social comparison by providing a summary of the activities done by the class (Brusilovsky et al., 2015).

Several research projects investigated the effect purpose-built tools added to MOOCs (Pérez-Álvarez et al., 2020). Davis et al. (2018) developed SRLx, a personalized tool for supporting SRL skills for the edX MOOC platform. The tool supports learners in planning their goals weekly, and provides feedback on realization of the plans via a dashboard. The use of the tool was voluntary. The authors found that only 32% of students used SRLx at least once, but those students engaged with various elements of the MOOC significantly more than their peers.

SRL skills have also been supported in hypermedia learning and games. MetaTutor (Azevedo et al., 2009) incorporates four animated agents, which explicitly support the SRL phases. Crystal Island (Carptenter et al., 2021) requires students to write self-reflections while solving open-ended problems.

3. Experimental Design

The study reported in this paper was conducted as a natural field experiment (Dunning, 2012). We observed learning of students in an existing database course at the University of Canterbury. The students had several lectures on relational databases and SQL before they were introduced to SQL-Tutor in a scheduled lab. They were free to interact with the system as much as they wanted over a period of four weeks, in preparation for a course SQL test on October 7, 2021. The post-test was administered within the system on October 6, 2021.



Figure 1. Experimental phases

The overall design of the experiment is shown in Figure 1. At the start of the first session with SQL-Tutor, students received a short pre-test, after which they were asked to specify their overall goal for using SQL-Tutor (Figure 2). There were no restrictions on how to specify the goal. The next step was to set the goal for the current week (Figure 3), which required students to specify how much time they wanted to spend in SQL-Tutor that week, how many problems they planned to solve, and what level they would like to achieve. The student level in SQL-Tutor ranges from 1 (the minimum) to 9, and depends on the student's progress on specifying queries. For more information about how SQL-Tutor models students' knowledge, please see the paper by Mitrovic (2003).

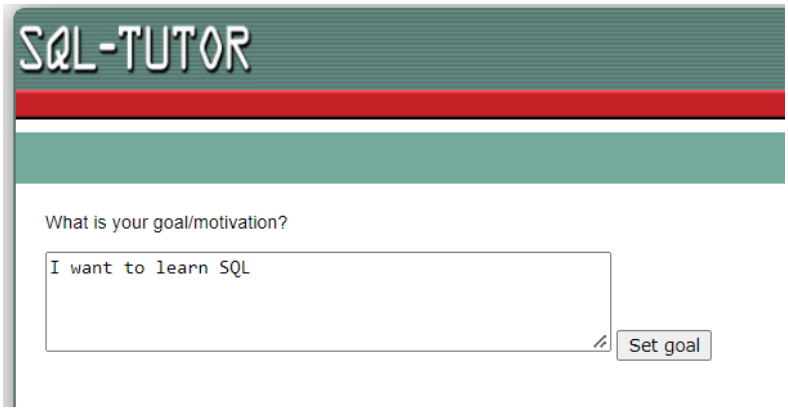


Figure 2. Setting the overall goal

After specifying the weekly goals, the dashboard was shown (Figure 4). The dashboard contains the overall goal, the visualizations showing the progress on the weekly goals, and the open learner model (OLM). From the dashboard, the student could change the overall goal, or go to the workspace and commence with tutored problem solving (Figure 4). At the start of each week, students set the goals for the week and get the dashboard.

Our intervention for supporting SRL skills (Figures 2-4) covers all three phases of Zimmerman's model. The forethought phase is supported by specifying the overall goal as well as the weekly goals. The dashboard supports the performance/reflection phases, as it allows the student to reflect on their progress towards goals, as well as self-assess/reflect on their knowledge using the OLM. Please note that the OLM was a part of the standard version of SQL-Tutor (Mitrovic & Martin, 2007). The OLM is shown to students in SQL-Tutor when they select problems to work on.

The intervention we developed is similar to the one used by Davis and colleagues (2018), with several differences. Davis et al. added support to a MOOC, while in our case the students interacted with an intelligent tutoring system. In addition to the differences based on the type of learning resources (watching videos in the case of MOOC and solving problems in the case of SQL-Tutor), there are differences in how the dashboard is utilized. The dashboard used in (Davis et al.) was only available by request, while in our case, the dashboard was shown to the student at the start of each week (after

specifying the weekly goals), and also at the start of each session. Additionally, students could explicitly request to see the dashboard using the button in the SQL-Tutor workspace. Therefore, our students had more opportunities to see the dashboard. Finally, our dashboard does not only provide feedback on the realization of the goals, but also includes the OLM, thus allowing students to reflect on their knowledge.

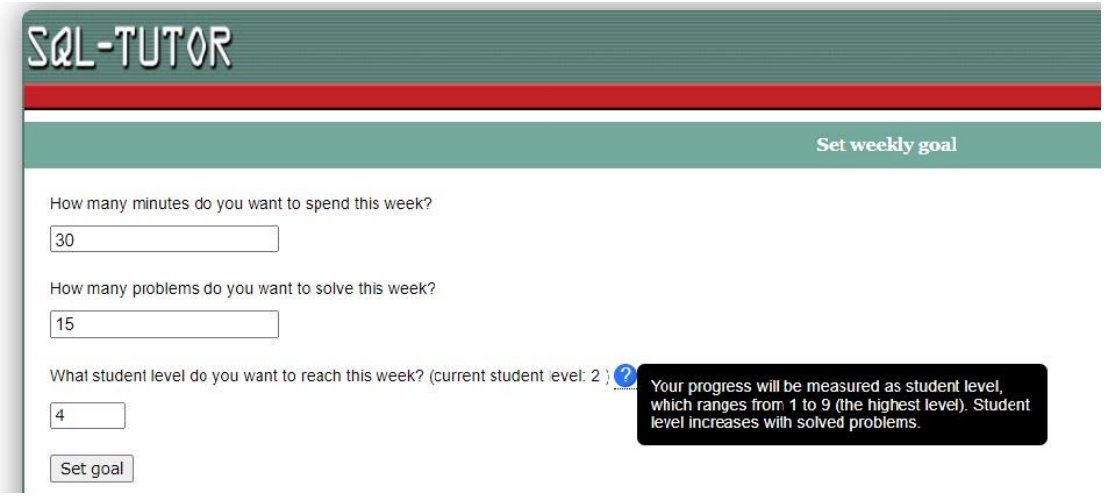


Figure 3. Setting the weekly goal

4. Findings

There were two versions of the test (versions A and B) which were counterbalanced as pre/post-test. Both tests had four questions of similar nature. At the pre-test time, 47 students took test A, and their mean score was 2.17 (sd = .99), while 62 students completed test B with the mean score of 2.16 (sd = 1.05). There was no significant difference (p = .97) between the pre-test scores, showing that they were of similar complexity. The difference in the numbers of students who completed tests A and B is due to the fact that pre-test completion was optional, so some students did not submit their answers.

We collected data from 136 students and eliminated the data about ten students who did not complete the pre-test, two students who did not take the mid-term test, and also data about 16 students who used SQL-Tutor for less than 10 minutes. That leaves data about 108 students. Both pre- and post-tests were optional, so not all students completed them. There were 71 students who completed both tests. Table 1 reports the basic statistics of the system usage. The students interacted a lot with the system, with the average time of over 4 hours. More than half of the students (55%) explicitly requested to see the dashboard (that is in addition to seeing the dashboard automatically at the start of the session/week), which is higher than the rate reported by (Davies et al. 2018).

Table 1. *Summary of interactions with SQL-Tutor*

	Min	Max	Median	Mean (sd)
Pre-test	0	4	2	2.17 (1.02)
Time (min)	13	1,007	188	267.11 (248.80)
Sessions	2	33	7.5	9.23 (6.65)
Solved Problems	0	266	26.5	37.81 (43.97)
Attempts	1	926	114	164.67 (169.40)
Level achieved	1	9	3	3.12 (1.90)
Post-test (71 students)	0	4	2	2.28 (0.92)
Dashboard requests (60 students)	1	28	2	4.52 (6.10)
SQL test (%)	7	100	58	54.76 (21.66)

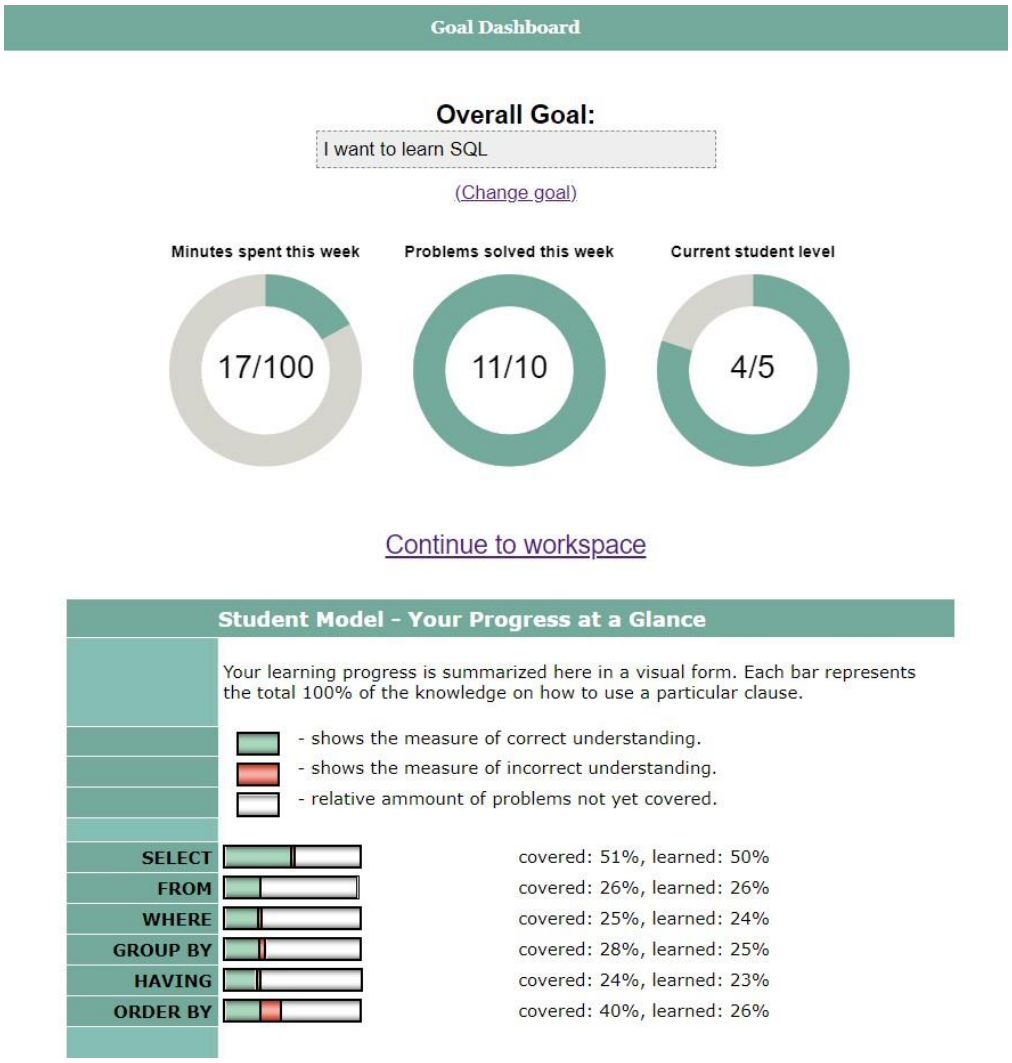


Figure 4. Dashboard

Of 108 students, 92 (85.2%) specified meaningful goals, which ranged in length from a single word to the maximum of 15 words (median = 4, avg = 4.69, sd = 3.16). Table 2 shows several example goal statements. The five most frequently occurring words in the stated goals are: *SQL* (49 times), *Learn* (29), *pass* (14), *get* (12) and *good* (11). Although students had the opportunity to change the overall goal at the start of each week, only two of them did so.

Table 2. *A random sample of short, medium and long submissions for the overall goal*

Overall Goal	Classified as
Learn sql	Intrinsic
To gather a good understanding of SQL	Intrinsic
To become familiar and comfortable with using SQL	Intrinsic
To pass this course with at least an A-	Extrinsic
To pass my course	Extrinsic
Pass tomorrows test	Extrinsic
Achieve high grades	Extrinsic

The overall goal statements were classified by the first author of this paper as intrinsic or extrinsic (as in Table 2). When analyzing the behavior and learning of the students who specified the two types of goals (Table 3), we did not find significant difference on most variables, but there was a significant difference on the SQL test scores and also on dashboard requests. The students who specified intrinsic goals explicitly requested to see the dashboard more often and achieved higher scores on the SQL test.

Table 3. *Comparing students who specified intrinsic/extrinsic goals*

	Intrinsic (59)	Extrinsic (33)	
Pre-test	2.20 (1.08)	2.15 (.91)	
Time (min)	272.90 (265.83)	255.85 (235.26)	
Sessions	9.64 (7.00)	9.27 (6.91)	
Solved Problems	42.64 (54.17)	31.85 (25.46)	
Level achieved	3.08 (2.17)	3.48 (1.58)	
Post-test	N = 39, 2.26 (.94)	N = 21, 2.19 (.81)	
Dashboard requests	**N = 36, 3.81 (4.77)**	**N = 12, 2.92 (2.15)**	t = 2.24, p < .05
SQL test (%)	**57.76 (19.97)**	**47.83 (21.16)**	U = 1202.5, p < .05

As SQL-Tutor is provided to students on a voluntary basis, not all students started interacting with it in week 1. Table 4 specifies the number of students who used the system in a particular week of the study (the *Active* row), as well as the number of students who used it for the first time that week (the *Started* row). Not surprisingly, the number of active students increased in the last week, due to the SQL test.

Table 4. *Numbers of students interacting with SQL-Tutor in different weeks*

	Week 1	Week 2	Week 3	Week 4
Active	33	32	31	98
Started	33	15	17	43

We also analyzed the weekly goals, in order to examine whether students were ambitious in setting their goals and whether they were able to fulfil their plans. We were interested in seeing whether their planning tendencies changed over time. The data is illustrated in Figure 5. Students' goals for the time, problems and student level increased over time. The actual time spent in the system and the number of solved problems were above the goals in the final week, as students were preparing for the test. However, the achieved student levels were constantly below the goals.

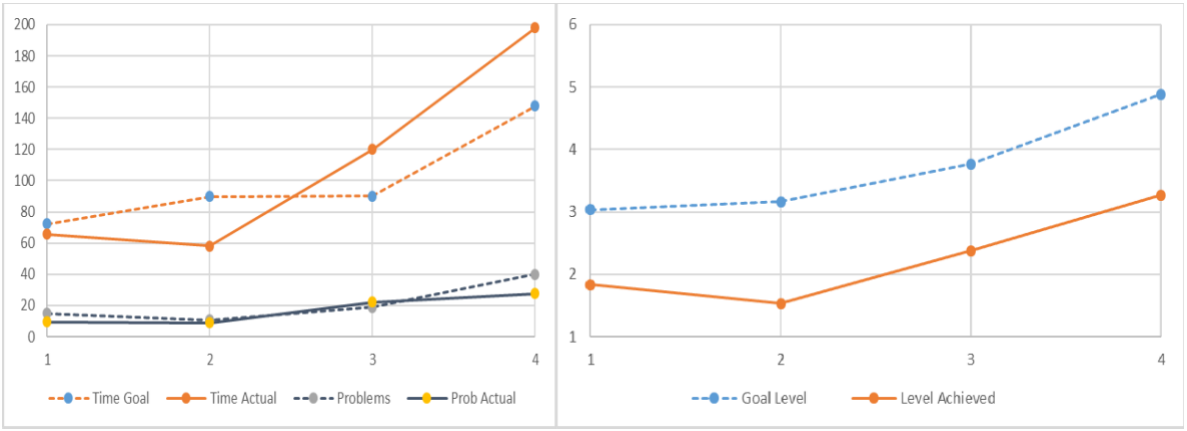


Figure 5. Planned vs achieved time, problems and student level

5. Analyzing Behavioral Differences

Intrinsic or extrinsic motivation is something students came in with and is not due to the SRL support. For that reason, we also analyzed the data collected from the 21 students scoring high on the SQL test (i.e., more than 75%) compared to those who achieved low marks, i.e., less than 35% (Table 5). Apart from the significant difference on the SQL test score, the only other significant difference we observed was on the post-test score (t = 3.49, p < .001).

Table 5. *Comparing Low and High groups*

	Low (22)	High (21)
Pre-test	1.95 (.99)	2.24 (1.09)
Time (min)	246.41 (250.13)	255.48 (238.61)
Sessions	10.41 (8.84)	7.81 (5.09)
Solved Problems	28.27 (27.29)	38.76 (38.51)
Level achieved	2.95 (1.62)	3.33 (1.83)
Post-test	**1.81 (.91)**	**2.88 (.81)**
Dashboard requests	1.73 (2.25)	3.10 (5.42)
SQL test (%)	**24.15 (7.40)**	**85.43 (6.97)**

As the frequency analyses did not discover any differences in students' interactions with SQL-Tutor, we used Epistemic Network Analysis (ENA) to identify behavioral differences. ENA is a quantitative ethnographic method for identifying connections in interactions, which could be conversations between multiple people, but can also be applied to interactions with software systems (Shaffer et al., 2009). ENA provides visualizations for qualitative interpretation along with statistical tests (Shaffer et al., 2016). Epistemic networks model co-occurrence of coded data, where codes represent relevant actions or events. The coded data is analyzed by identifying codes within a moving window which defines the temporal context. The resulting network is aggregated across all lines for each unit of analysis, and visualized as graphs, where codes are represented as nodes, and edges represent relative frequency of co-occurrence between the codes.

We used the ENA1.7.0 Web tool (Marquart et al., 2018) for investigating the interactions. From SQL-Tutor logs, we extracted data about the following actions which are of relevance for SRL, and used them as codes for ENA:

- *Start* (of a session),
- *DashboardRequest* (explicit request by a student),
- *NewDatabase* (when student requests a new database to practise on),
- *StudentProblemChoice* (the student selects the next problem to work on on their own),
- *NewProblem* (request for a new problem),
- *ProblemSolved* (correctly solved problem),
- *ProblemAttempt* (the submitted solution has some errors),
- *ProblemAttemptHint* (incorrect solution, and hint was provided),
- *ProblemAttemptAllErrors* (incorrect solution, and the student requested to see all errors),
- *ProblemAttemptPartialSolution* (incorrect solution; the student requested to see a partial solution),
- *ProblemAttemptFullSolution* (incorrect solution; the student requested to see the full solution).

For the students in the Low/High groups, there was a total of 5,619/6,503 selected events from the logs. When generating epistemic networks, we defined units as the group (Low vs High) subsetted by the student. The unit of conversation was set to include all actions performed by a single student. Figure 6 shows the networks for the Low (left) and High (right) groups. The width of a line connecting two nodes is proportional to the frequency of those two codes co-occuring in interactions.

The two networks look similar, but the ENA tool also produces the network showing the differences between the two groups (Figure 7). The color of a connector between two codes shows which group had more co-occurences of the two codes. For example, the connection between *ProblemAttemptFullSolution* and *ProblemSolved* is colored red, which means that those two codes occured more often for the Low group. The model had co-registration correlations of 0.98 (Pearson) and 0.98 (Spearman) for the x-axis, and co-registration correlations of 0.96 (Pearson) and 0.95 (Spearman) for the y-axis, showing a very good fit. There was a significant difference between the two groups on the x-axis: a two sample t-test assuming unequal differences showed that the Low group (mean = .46, sd = .99) was significantly different from the High group (mean = .49, sd - .58), $t = 3.86$, $p < .001$, Cohen's $d = 1.16$, but there was no significant difference on the y-axis. Further exploration of the difference network shows that the low group more often required additional feedback (at the Full Solution and the Partial Solution level), followed by a correct submission. We suspect that some students from the Low group copied the solution provided by the system (partial or full), and then submitted it as their solution.

Such behavior of course does not result in learning, thus explaining significantly lower scores on the post-test and the SQL test for the students from the Low group.

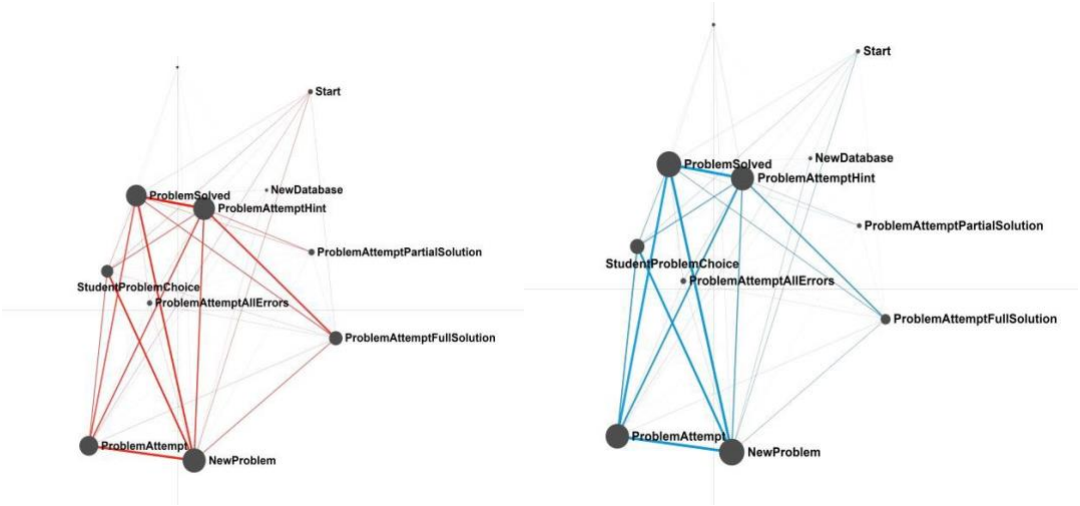


Figure 6. Epistemic networks for the Low (left) and High (right) groups

The percentage of problems solved after seeing the full solution for the Low group is 24.34% (sd = 26.44), while for the High group the average is 16.65% (sd = 13.08). Although the difference between the two groups on just the proportion of problems solved after seeing the complete solution is not statistically significant, ENA provides better insights as it analyzes several measures simultaneously. The locations of codes (i.e., events) in ENAs shows that all codes representing high-level feedback are grouped close to each other (i.e., full solution, partial solution and all errors).

The connection between *NewProblem* and *ProblemAttemptFullSolution* is also more common for the Low group, which shows that some students ask for a full solution immediately, without trying to solve the problem on their own. On the other hand, the connection between *NewProblem* and *StudentProblemChoice* are more frequent for the High group, showing that those students more often select problems for themselves. All these observations indicate stronger SRL skills and learning strategies for the High group.

6. Discussion and Conclusions

We presented a simple support for self-regulated learning added to SQL, and our observations of how students interacted with the support and with the system itself. We extended SQL-Tutor, asking student to i) specify their overall learning goal, ii) specify weekly goals in terms of the time with the system, problems solved, and the level achieved, and iii) a dashboard summarizing their progress on goals, and also providing an open student model. Although we have found that intrinsically motivated students learnt significantly more than their peers and also made significantly more dashboard requests, students' motivation was measured before interacting with SQL-Tutor and could not have been affected by the SRL support. When comparing the interactions of high-scoring students to low-scoring students, frequency analyses did not indicate any significant differences between the numbers of various learning events. However, a comparison of epistemic networks corresponding to the two subgroups of students shows a significant difference. The students who scored low on the SQL tests requested more detailed feedback on their solutions to problems, often asking for a partial or even the full solution. In the case of one student from the Low group, who solved two problems during the total interaction time of 14 minutes, for both of those problems the student first asked for a full solution, then copied and submitted it. Seeing the full solution may result in the illusion of understanding, a phenomenon identified in studies on learning from worked examples (Renal, 2002; Gerjets et al., 2006, Shareghi Najar & Mitrovic, 2013), which may lead to overestimation of competence. Therefore, low-scoring students need to be encouraged to be more active in learning and solve problems on their own. Such students need to improve their SRL and especially help-seeking skills.

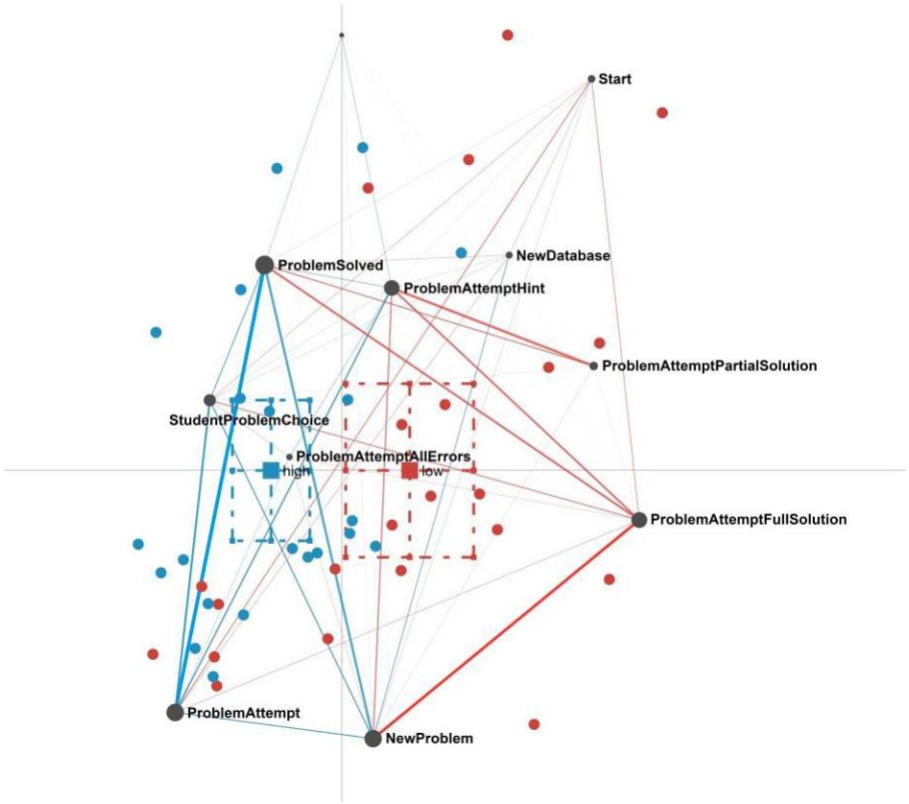


Figure 7. The difference network

There are several limitations of our study. Firstly, our study was only an observational study; we did not compare two groups of students with/without the SRL support. The goal of the reported study was to collect some information about how students interact with the provided SRL support. It is encouraging that more than half of students (55.55%) explicitly requested to see the dashboard, compared to 32% of students engaging with SRLx at least once (Davis et al., 2018). The second limitation is that our SRL support was fixed. In the future, we plan to design adaptive SRL support, and provide different guidance to students with low SRL skills.

References

Azevedo, R., Witherspoon, A., Chauncey, A., Burkett, C., & Fike, A. (2009). MetaTutor: A metacognitive tool for enhancing self-regulated learning. Cognitive and metacognitive educational systems: AAAI Fall Symposium.

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: a systematic review. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 41-50).

Boekaerts, M. (1999). Self-Regulated Learning: Where We Are Today. *International Journal of Educational Research*, 31(6), 445–475.

Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., Zadorozhny, V., & Durlach, P. J. (2015). Open social student modeling for personalized learning. *IEEE Transactions on Emerging Topics in Computing*, *4*(3), 450-461.

Carpenter, D., Cloude, E., Rowe, J., Azevedo, R., & Lester, J. (2021). Investigating student reflection during game-based learning in middle grades science. In *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 280-291).

Chatti, M. A., Marinov, M., Sabov, O., Laksono, R., Sofyan, Z., Yousef, A. M. F., Schroeder, U. (2016) Video annotation and analytics in CourseMapper. *Smart Learning Environments*, *3*(1), 10.

Davis, D., Triglianos, V., Hauff, C., & Houben, G. J. (2018). SRLx: A personalized learner interface for MOOCs. In *European Conference on Technology Enhanced Learning* ECTEL (pp. 122-135). Springer, Cham.

Dimitrova, V., & Mitrovic, A. (2021). Choice Architecture for Nudges to Support Constructive Learning in Active Video Watching. *International Journal of Artificial Intelligence in Education*, 1-39.

Dunning, T. (2012). Natural experiments in the social sciences: A design-based approach. Cambridge: Cambridge University Press.

Gerjets, P., Scheiter, K., & Catrambone, R. (2006). Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations? *Learning and Instruction*, *16*(2), 104–121.

Hooshyar D., Pedaste, M., Saks, K., Leijen, Ä., Bardone, E., & Wang, M. (2020). Open learner models in supporting self-regulated learning in higher education: A systematic literature review. *Computers & Education*, *154*, 103878.

Kizilcec, R.F., Halawa, S. (2015) Attrition and achievement gaps in online learning. In: L@S 2015, pp. 57–66.

Kizilcec, R.F., Perez-Sanagustın, M., Maldonado, J. (2017) Self-regulated learning strategies predict learner behavior and goal attainment in massive open online courses. *Computers & Education*, 104, 18–33.

Koller, D., Ng, A., Do, C., Chen, Z. (2013) Retention and intention in massive open online courses. *Educause Rev.* *48*(3), 62–63.

Matcha, W., Gašević, D., & Pardo, A. (2019). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies*, *13*(2), 226-245.

Marquart, C. L., Hinojosa, C., Swiecki, Z., Eagan, B., & Shaffer, D. W. (2018) Epistemic Network Analysis (Version 1.7.0) [Software].

Mirriahi, N., Jovanović, J., Lim, L. A., & Lodge, J. M. (2021). Two sides of the same coin: video annotations and in-video questions for active learning. *Educational Technology Research and Development*, *69*(5), 2571-2588.

Mitrovic, A. and Ohlsson, S. (1999) Evaluation of a Constraint-Based Tutor for a Database Language. *Int. J. Artificial Intelligence in Education*, 10(3-4), 238-256.

Mohammadhassan, N., Mitrovic, A., Neshatian, K. (2022) Investigating the Effect of Nudges for Improving Comment Quality in Active Video Watching. *Computers & Education*, 176, 2022, 104340.

Moos, D. C., & Bonde, C. (2016). Flipping the classroom: Embedding self-regulated learning prompts in videos. *Technology, Knowledge and Learning*, *21*(2), 225-242.

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in psychology*, *8*, 422.

Pérez-Álvarez, R. A., Maldonado-Mahauad, J., Sharma, K., Sapunar-Opazo, D., & Pérez-Sanagustín, M. (2020). Characterizing Learners' Engagement in MOOCs: An Observational Case Study Using the NoteMyProgress Tool for Supporting Self-Regulation. *IEEE Transactions on Learning Technologies*, *13*(4), 676-688.

Pintrich, P. (2000) The role of goal orientation in self-regulated learning," in *Handbook of Self-Regulations*, eds M. Boekaerts, P. Pintrich, and M. Zeidner (Eds.). San Diego, CA: Academic Press, 451–502.

Renkl, A. (2002). Worked out examples: Instructional explanations support\ learning by self-explanations. *Learning and instruction*, *12*(5), 529-556.

Schumacher, C., & Ifenthaler, D. (2021). Investigating prompts for supporting students' self-regulation–A remaining challenge for learning analytics approaches? *Internet and Higher Education*, *49*, 100791.

Shaffer, D. W., Hatfield, D. L. Svarovsky, G. N., Nash, P., Nulty, A, et al. (2009) Epistemic network analysis: A prototype for the 21st century assessment of learning. Learning and Media, 1(1), 1-21.

Shaffer, D. W., Collier, W., & Ruis, A. R. (2016). A tutorial on epistemic network analysis: Analyzing the structure of connections in cognitive, social, and interaction data. *Journal of Learning Analytics*, *3*(3), 9-45.

Shareghi Najar, A., & Mitrovic, A. (2013, July). Examples and tutored problems: How can self-explanation make a difference to learning? In *International Conference on Artificial Intelligence in Education* (pp. 339-348). Springer, Berlin, Heidelberg.

Sonnenberg, C., & Bannert, M. (2016). Evaluating the impact of instructional support using data mining and process mining: A micro-level analysis of the effectiveness of metacognitive prompts. *Journal of Educational Data Mining*, *8*(2), 51–83.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated engagement in learning. *Metacognition in educational theory and practice*, 277-304.

Wong, J., Baars, M., Davis, D., Van Der Zee, T., Houben, G. J., & Paas, F. (2019). Supporting self-regulated learning in online learning environments and MOOCs: A systematic review. *International Journal of Human–Computer Interaction*, *35*(4-5), 356-373.

Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary educational psychology*, *11*(4), 307-313.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. *Theory into practice*, *41*(2), 64-70.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Development of a Learning Companion Robot with Adaptive Engagement Enhancement

Bowei YAO[a], Koichi OTA[b,a], Akihiro KASHIHARA[c], Teruhiko UNOKI [d] & Shinobu HASEGAWA[b,a*]

[a]*Division of Advanced Science and Technology, JAIST, Japan*
[b]*Center for Innovative Distance Education and Research, JAIST, Japan*
[c]*Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan*
[d]*College of Foreign Studies, Kansai Gaidai University, Japan*
*hasegawa@jaist.ac.jp

Abstract: This research aims to develop a learning partner robot that can adapt its interaction to each learner's individual reactions to enhance learner engagement. Engagement is a mental state that positively influences learning and is an essential element that supports learner independence through immersion in the learning process. However, no established methodology exists to enhance learner engagement with proper support in the individualized and isolated learning process. This research focuses on a learning partner robot as a learning companion in self-directed learning. To realize the partner robot, we implemented the engagement estimation architecture from the learner's facial images during learning, then designed a robot interaction model to enhance learner engagement according to their engagement states, and extended the interaction model to update the robot's strategy according to the learner's response to the interactions. We conducted a comparative experiment with 20 graduate student participants with and without robots. The results indicated that the average engagement during learning was significantly higher in the with-robot condition, and the satisfaction with the robot interaction was highest in the last 1/3 of the learning period.

Keywords: Learner Engagement, Adaptation, Facial Images, Learning Companion Robot

1. Introduction

Self-regulated learning and learner engagement, defined as the involvement of the learners' cognitive and emotional energy to accomplish a learning task (Schunk & Mullen, 2012), is critical to the success of online learning (Schaeffer & Konetes, 2010). However, decreased interaction between instructors and friends and social isolation from the COVID-19 pandemic might easily demotivate and disengage their learning (Akuratiya & Meddage, 2020).

This research aims to develop a learning partner robot that enhances learner engagement to solve these problems. Leite et al. highlighted the increasing popularity of social robots in educational environments (Leite et al., 2013). Such robots have a potential to provide an adaptive learning experience and support learners in a new way that is not possible in today's resource-limited educational environments. However, Belpaeme et al. pointed out a couple of challenges, such as estimating the learners' abilities and progress and coordinating the timing of verbal and nonverbal action to select appropriate robot interactions (Belpaeme et al., 2018). In other words, no established methodology exists to enhance learner engagement with proper support in the individualized and isolated learning process.

This article proposes a learning partner robot that can adapt its interaction to each learner's individual reactions to enhance learner engagement during self-directed learning at online learning. The features of the developed robot are not to support learning in a specific domain but to support self-monitoring of engagement in self-regulated learning by providing feedback to the learners on their engagement status. We also try to realize an adaptive robot interaction for supporting engagement by each learner's preference.

2. Related Work

2.1 Learning Partner Robots

Lu et al. designed a new robot system called Smart Learning Partner, which used natural language processing and emotion recognition to provide learners with an enjoyable learning experience (Lu et al. 2018). Using self-determination theory as a guideline, they designed human-robot interaction in terms of autonomy, which refers to spontaneity in performing tasks, competence, a sense of challenge and effectiveness, and relatedness, a sense of connection with others. However, the effectiveness of this interaction has not yet been discussed.

Lubold et al. introduced a socially responsive speech interface for learning companions of embodied robots with voice adaptation based on acoustic-prosodic entrainment (Lubold et al. 2016). Through evaluation, they showed that social presence was significantly higher for social speech-adaptive speech interfaces than for purely social dialogues. Although these studies are closely related to our target engagement, this research is unique in effectively promoting engagement by integrating engagement estimation and interaction generation.

2.2 Sota as Communication Robots

Recent advances in robotics have led to the usage of various communicative robots in learning support, such as NAO (Softbank Robotics Co., Ltd., 2018), Saya (Hashimoto and Kobayashi, 2005), Robohon (Sharp Corporation, 2016), and PALRO (FUJISOFT Inc., 2010). This research adopts Sota, a table-top communication robot equipped with a camera, microphone, speaker, and network functions (Vstone Co. Ltd., 2010). Sota has advanced functions and an adorable design as a platform that can provide various robot services in conjunction with loT (The Internet of Things) devices and cloud AI. It has a built-in Intel(R) Edison with Linux, and applications using image recognition, speech recognition, and speech synthesis can be easily developed using Java.

3. Proposed Method

3.1 Engagement Estimation

This research implemented the engagement estimation model proposed by Huynh et al. (Huynh et al. 2019). Their work demonstrated the effect of face and gaze related features on engagement estimation in the "Wild" environment. This research extracted the following two feature sets from the learning videos based on Huynh's method.

F1 feature set: 60-dimension data with the average, standard deviation, maximum, and minimum values of gaze direction, eye landmarks, distance from eye to the camera, and head pose at video segments with regular intervals extracted by OpenFace, an open-source toolkit for facial expression analysis (Baltrusaits et al. 2018).

F2 feature set: 128-dimension data extracted by SE-ResNet-50 (Hu et al. 2018), ResNet-50 with a Squeeze-and-Excitation (SE) block, for face regions extracted by OpenFace. SE-ResNet-50 has been pre-trained on MS-Celeb-1M (Guo et al. 2016) and VGGFace2 (Cao et al. 2018).

Learner engagement changes over time. Based on Huynh et al., LSTM-FC and FCLSTM-FC (fully connected right-front LSTM), which combine the LSTM (long short-term memory) and FC (fully connected) layers, train each feature set to capture the change. The ensemble layer receives these results and outputs the averaged learner engagement in each video segment. According to the original model by Huynh et al., the output engagement intensities were the following four levels:

0: Disengaged: 0 <= Engagement-Intensity < 0.4
1: Barely Engaged: 0.4 <= Engagement-Intensity < 0.6
2: Engaged: 0.6 <= Engagement-Intensity < 0.83
3: Highly Engaged: 0.83 <= Engagement-Intensity <= 1.00

3.2 Interaction Network

The components of partner robot interaction depend on verbal and nonverbal robot behavior. Generally speaking, emotions involve verbal and nonverbal domains: verbal expression refers to speeches or languages (Fussell, 2002), and nonverbal expression includes speed, intonation, posture, gesture, body movement, and facial expression (Dael et al., 2012). Based on these references, we designed verbal and nonverbal expressions for robot interactions to enhance learner engagement.

For the verbal expression, we referred to several feedback techniques that instructors use with their learners. "Support," "care," and "praise" positively influence learners' performance and behavior (Brophy 1981, Jiang 2019). In addition, Cutumisu et al. showed that instructors' critical feedback is more effective on learners' memory than confirmatory feedback (Cutumisu et al. 2018). Thus, we prepared three short verbal speeches in each of the four interaction categories shown in Table 1.

Table 1. *Verbal Speeches as Interaction Components*

Types	Short Verbal Speeches		
Support	I'm supporting you.	Please don't lose.	I'll always be there for you.
Criticize	Have you done something?	Are you wasting time?	When will you start learning?
Care	Take it easy.	Please take a rest.	It's tough for you.
Praise	Good job!	I'm proud you.	Keep it up!

For nonverbal expressions, we decided to use gestures and body movement to emphasize the intention of robot speeches and employed the four motion types, "Call," "Presen (presentation)," "Talk," and "Bye," as presets by the MotionAsSotaWish class, Sota's automatic motion generation function. Each motion type defines several motions of a few seconds duration, and Sota acts simultaneously with its speeches. In addition, Sota uses the TextToSpeechSota class, a cloud speech API, which generates .wav files for utterances. As another nonverbal interaction component, we adopted speech rate, and three types were selected: Slow, Medium, and Fast.

To provide the support that reflects each learner's interaction preferences, we propose the interaction network model. The right bottom of Figure 1 shows the initial state of the model, a fully connected network consisting of four layers of interaction components (speech type, motion, and speech rate), with the estimated results of engagement intensity in Section 3.1 as input. We set the initial weights among each path based on previous research (Fried, 2012). The sum of the path weights for each node to the next layer is 1. These weights are used to select the next node when determining interactions probabilistically. It allows different interactions to be generated even in the same situation. When the learner's engagement is "Highly Engaged," the robot does not interact with the learner.

The proposed method uses an engagement estimation model to compare the learner's engagement changes before and after the interaction with Sota. It determines whether the interaction was appropriate for the learner and updates the interaction network. If the learner's engagement remains the same or decreases after the interaction, the weights of all paths where the interaction took place are reduced, and the weights of other paths are increased to the sum of the weights to 1. This makes Sota implement other interaction strategies when a similar situation occurs. The interaction network would slowly find the most suitable weights to achieve good interactive experiences tailored to different learners' personalities. There would be inappropriate interactions while adjusting the weights, but it would improve as the update progresses.

3.3 Implementation

Figure 1 shows the overall architecture of the developed system. The engagement estimation and interaction network modules were implemented on a PC with Intel i7-9900k, 32G RAM, Nvidia GeForce RTX 2080 (6GB), ubuntu 20.04 OS, python 3.8, and TensorFlow 2.0. First, the system takes the learner's face image from a USB camera as input and estimates the engagement intensity at one-minute intervals using a pre-trained model described in section 3.1. Next, the interaction network selects Sota's interaction from the estimated engagement intensity explained in section 3.2. Then, Sota obtains the output of the interaction network in JSON format at one-minute intervals and performs the

selected interaction. Finally, the system updates the interaction network according to the estimated results of the learner's engagement after the interaction discussed in section 3.2.

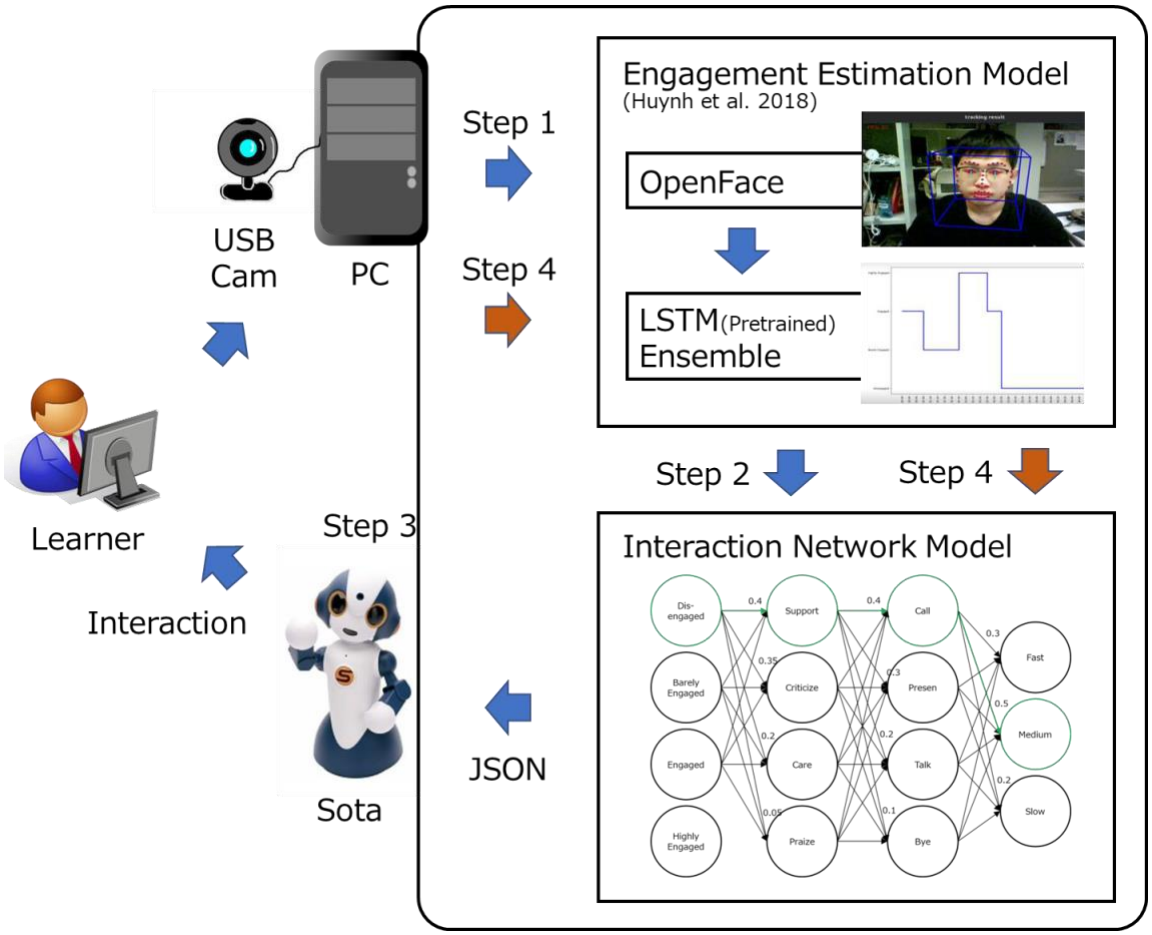


Figure 1. System Architecture

4. Experiment

To evaluate the validity of the proposed method, we conducted a within-subjects design experiment with 20 graduate students. In particular, we evaluated the effect of Sota's interaction on learner engagement and the effect of the interaction network update algorithm. In the experiment, subjects studied 30-minute learning videos, undergraduate level physics and high school level chemistry, with and without Sota conditions. The order of the experiment was counterbalanced concerning the learning videos and the with and without Sota. After learning each video, subjects rated themselves on a 4-point scale (0: Disengaged, 1: Barely Engaged, 2: Engaged, and 3: Highly Engaged). for each 5-minute engagement while watching their own facial videos and on a 5-point scale (1: Very Bad, 2: Bad, 3: Neither, 4: Good, 5: Very Good) for the satisfaction of timing and content with each interaction in the condition with Sota.

4.1 Effect of Sota's Interaction on Learner's Engagement

The average engagement with Sota was 2.07 (S.D. = 0.47), and the average engagement without Sota was 1.71 (S.D. = 0.47). The t-test result $t(19)=2.48$, $p=.022$ showed that the average engagement was significantly higher in the With Sota condition. Although this result is not sufficient due to the short experiment time, it suggests that the interaction by Sota may have a certain effect on enhancing engagement.

Table 2 shows the questionnaire results regarding the number of satisfactions with the timing and content of each interaction by Sota in the With Sota condition. The average satisfaction with the interaction timing was 3.43 (S.D. = 0.608), and the average satisfaction with the interaction content was 3.24 (S.D. = 0.728), indicating that about half of the interactions were positively accepted. The time lag for estimation itself and model update might cause low satisfaction and the accuracy limitations of the engagement estimation model.

Table 2. *Number of Satisfaction of Interaction Timing and Content with Sota*

	1 (Very Bad)	2 (Bad)	3 (Neither)	4 (Good)	5 (Very Good)
Timing	0	22	79	79	13
Content	0	39	72	72	8

4.2 Effect of Interaction Network Update Algorithm

Table 3 compares the averages of the questionnaire results for each subject's satisfaction with the interaction content when divided into 1/3 of the time order. We tested the data with the unpaired ANOVA to compare the three groups, which showed that F=18.9 (p=<.00001). Tukey's HSD was used to compare each group. The results showed Q=3.34 (p=.0498) between the first 1/3 and the middle 1/3, Q=5.22 (p=.0085) between the first 1/3 and the last 1/3, and Q=8.56 (p=.0000) between the middle 1/3 and the last 1/3, respectively. These results show that the subjects' satisfaction with the interaction decreased in the middle 1/3 of the experiment but improved as the weights were adjusted in the last 1/3 experiment.

Table 3. *Average Satisfaction of Interaction Content among Time Order Groups*

	# of Answer	Average	S.D.
First 1/3	63	3.17	0.814
Middle 1/3	61	2.84	0.879
Last 1/3	63	3.70	0.706

5. Conclusion

This research developed a learning partner robot based on Sota that can adapt its interaction to each learner's individual reactions to enhance learner engagement during self-directed learning at online learning. We also conducted a within-subjects design experiment (with/without Sota) with 20 graduate students to evaluate its effectiveness.

We constructed an initial interaction network to generate interactions that promote engagement based on learner engagement. The network is a fully connected network consisting of four layers representing the learner engagement intensity, the robot speech type, and motion and speech rate. The experiment results showed that the average engagement with Sota condition was higher than those without Sota condition. Since each session in this experiment was short (30 minutes) and the subjects interacted with Sota for the first time, we need additional experiments for a long learning period.

We also proposed an interaction updating model in which the robot changes the interaction content by judging the previous interaction based on the estimated learner engagement after the interaction. The experimental results showed that the satisfaction of the last 1/3 of the interaction content was significantly higher than that of the previous interaction content. The current interaction network has only a four-layer structure, which does not satisfy some learners' preferences for Sota because the interaction patterns are small. We should increase the number of layers in the network and add more content and interaction methods. Furthermore, since the weights in the early stages affect the satisfaction of early learner interactions, the speed and rules for updating the weights affect the satisfaction of later interactions.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 20H04294 and Photron limited.

References

Akuratiya, D. A. & Meddage, D. N. R. (2020). Students' Perception of Online Learning during COVID-19 Pandemic: A Survey Study of IT Students. International Journal of Research and Innovation in Social Science. 4(9). 755-758.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L., P. (2018). Openface 2.0: Facial behavior analysis toolkit. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp.59-66.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B. and Tanaka, F. (2018). Social robots for education: A review. Science Robotics. 3(21). DOI: 10.1126/scirobotics.aat595.

Brophy, J. (1981). Teacher praise: A functional analysis. Review of Educational Research, 51(1), pp.5-32. https://doi.org/10.3102/00346543051001005

Cao, Q., Shen, L., Xie, W., Parkhi, O., M., and Zisserman A. (2018). Vggface2: A dataset for recognizing faces across pose and age. In 13th IEEE International Conference on Automatic Face & Gesture Recognition. pp.67-74.

Chiu, T.K.F., Lin, TJ. & Lonka, K. Motivating Online Learning: The Challenges of COVID-19 and Beyond. Asia-Pacific Edu Res 30, 187–190 (2021). https://doi.org/10.1007/s40299-021-00566-w.

Cutumisu, M. and Schwartz, D. L. (2018). The impact of critical feedback choice on students' revision, performance, learning, and memory. Computers in Human Behavior, 78(Jan.). pp.351-367. https://doi.org/10.1016/j.chb.2017.06.029

Dael, N., Mortillaro, M., Scherer, K.R. (2012). The Body Action and Posture coding system (BAP): Development and reliability. Journal of Nonverbal Behavior. 36. pp.97-121.

Fussell, S.R. (Ed.) (2002). The Verbal Communication of Emotion: Interdisciplinary Perspectives; Lawrence Erlbaum Associates: Mahwah, NJ, USA.

Fried, L. (2011). Teaching Teachers about Emotion Regulation in the Classroom. Australian Journal of Teacher Education, 36(3). http://dx.doi.org/10.14221/ajte.2011v36n3.1

FUJISOFT Inc. (2010). PALRO is A robot who cares. Retrieved May 31, 2022. https://palro.jp/en/

Guo, Y., Zhang, L., Hu, Y., He, X., and Gao. J., (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In European Conference on Computer Vision. pp.87-102.

Hashimoto, T. Kobayashi, H. (2005). Development of the receptionist system with an anthropomorphism face, Proceedings of the 5th Asian Symposium on Applied Electromagnetics And Mechanics, pp. 190-196.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp.7132-7141.

Huynh, V. T., Kim S. H., Lee G. S., Yang H. J. (2019) Engagement Intensity Prediction with Facial Behavior Features, 2019 International Conference on Multimodal Interaction, pp.567-571, https://doi.org/10.1145/3340555.3355714

Jiang, J., Vauras, M., Volet, S., and Salo, A. E. (2019). Teacher Beliefs and Emotion Expression in Light of Support for Student Psychological Needs: A Qualitative Study. Education Sciences. 9(68).

Lubold, N., Walker, E., and Pon-Barry, H. (2016) Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 255-262, doi: 10.1109/HRI.2016.7451760.

Leite, I., Martinho, C., Paiva, A. (2013). Social robots for long-term interaction: A survey. International Journal of. Social Robotics. 5, 291–308.

Lu, Y., Chen, C., Chen, P., Chen, X., Zhuang, Z. (2018). Smart Learning Partner: An Interactive Robot for Education. In: , et al. Artificial Intelligence in Education. AIED 2018. Lecture Notes in Computer Science(), vol 10948. Springer, Cham. https://doi.org/10.1007/978-3-319-93846-2_84

Schaeffer CE & Konetes GD (2010). Impact of learner engagement on attrition rates and student success in online learning. International Journal of Instructional Technology & Distance Learning.7(5), pp..3-9.

Schunk, D. H., & Mullen, C. A. (2012). Self-efficacy as an engaged learner. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), Handbook of research on student engagement (pp. 219–235). New York, NY: Springer. https://doi.org/10.1007/978-1-4614-2018-7

Sharp Corporation. (2016). Robohon. Retrieved May 31, 2022, from https://robohon.com/global/.

Softbank Robotics Co. Ltd. (2018). NAO the humanoid and programmable robot. Retrieved May 31, 2022 from https://www.softbankrobotics.com/emea/en/nao/.

Vstone Co. Ltd. (2010). Social communication robot Sota. Retrieved May 31, 2022 from https://www.vstone.co.jp/products/sota/.

Zheng, X., Hasegawa, S., Tran, MT., Ota, K., Unoki, T. (2021). Estimation of Learners' Engagement Using Face and Body Features by Transfer Learning. In: Degen, H., Ntoa, S. (eds) Artificial Intelligence in HCI. HCII 2021. Lecture Notes in Computer Science(), vol 12797. Springer, Cham. https://doi.org/10.1007/978-3-030-77772-2_36

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Practical Use of an Error-based Problem Presentation System in Mechanics

Nonoka AIKAWA[a*], Shintaro MAEDA[a], Tomohiro MOGI[a], Kento KOIKE[a], Takahito TOMOTO[b], Isao IMAI[c], Tomoya HORIGUCHI[d] & Tsukasa HIRASHIMA[e]
[a]*Graduate School of Engineering, Tokyo Polytechnic University, Japan*
[b]*Faculty of Engineering, Tokyo Polytechnic University, Japan*
[c]*Chiba Municipal Satsukigaoka Junior High School, Japan*
[d]*Graduate School of Maritime Sciences, Kobe University, Japan*
[e]*Graduate School of Advanced Science and Engineering, Hiroshima University, Japan*
*n.aikawa@st.t-kougei.ac.jp

Abstract: Learners studying mechanics sometimes fail to learn correct scientific knowledge because of misconceptions they have from their experience with everyday phenomena. Error-based Simulation (EBS), which encourages trial-and-error by presenting a simulation based on the learner's solution, is effective in correcting errors. However, learners that make similar mistakes over a long period of time during the trial-and-error process tend to become stuck and fail to make progress in problem solving. In this study, we developed a system for adaptively providing auxiliary problems that address the areas in which stuck learners repeatedly make errors, helping them to overcome these errors and allowing them to progress. We evaluated the effectiveness of this system by testing it in a junior high school class. The results suggest that it was effective for junior high school students' learning.

Keywords: Elementary mechanics, Learning support system, Learning from mistakes, Error-based Simulation, Classroom practice

1. Introduction

Learners studying mechanics may not learn correct scientific knowledge because of misconceptions they have from their experience with everyday phenomena. Junior high school classes attempt to address this issue by providing learners with opportunities to obtain correct knowledge through experiments. However, experiments are costly and require teachers to provide learners with correct explanations of the phenomena involved.

Error-based Simulation (EBS) is an effective method for correcting errors (Hirashima et al., 1998; Hirashima et al., 2017; Horiguchi et al., 2005; Horiguchi et al., 2014). It is a learning support framework that visualizes learners' errors and encourages trial and error. EBS first presents a problem, and then based on the learner's answer to the problem, presents a simulation of what would happen if the learner's (incorrect) answer was correct. The incorrect answer causes the simulation to behave strangely, thereby making the learner aware that they have made an error.

One drawback of EBS, however, is that if a learner gets stuck on a problem in the middle of the trial-and-error process, then EBS can have difficulty helping the learner to overcome the impasse. Specifically, since Mechanics EBS visualizes the motion of an object in response to a force input to the object, it may not be able to effectively visualize complex tasks where many forces act on an object, thus prohibiting the learner from solving the problem.

In this study, an impasse is defined as a state in which similar failures over a long period of time during the trial-and-error process prohibit the learner from making any progress in solving a problem. It is common practice to give a learner who is stuck a clue pertaining to their error or to outright give them the correct answer (Shute, 2008). However, giving the correct answer interferes with the trial-and-error process that EBS is centered around.

Yamamoto et al. (2021) used a simple electroencephalograph to estimate the learner's stuck state for problem posing in arithmetic, identified the cause of the stuck state based on an analysis of the learning task, and provided the corresponding feedback, thereby helping to resolve the learner's

difficulties. The research indicates the need for domain-specific feedback to assist with learners' impasse.

Providing auxiliary problems that help learners to understand the original one might be an effective method for encouraging students to continue the trial-and-error process without giving them the correct answer. One way to create auxiliary problems is to simplify the original problem (Hirashima et al., 1994). Students can use what they learn from solving these simplified problems to help them solve the original problem.

In our study, we developed an EBS system for mechanics that automatically diagnoses the learner's error points and presents auxiliary problems accordingly. Specifically, we designed a system that implements conventional EBS by designing both auxiliary problems and their problem sequences and that analyzes error points based on the learner's answer history. We evaluated this system by conducting an experiment with university students and were able to confirm its effectiveness (Aikawa et al., 2020).

However, elementary mechanics is a junior high school subject in Japan. To extend the applicability of our developed system to beginners, it is important to investigate whether our system was also effective for junior high school students. Therefore, we conducted a class practice session using the system developed in this study.

2. System Used for Classroom Practice

In this section, we provide more details about the system we used in the classroom practice session.

Learners first work on exercises in a system that is similar to conventional EBS (Figures 1(a) and (b)). If they make a mistake on the same problem a certain number of times, then the system offers them the option to solve an auxiliary problem (Figure 1(c)). And if the learner chooses to solve the auxiliary problem, then the system automatically presents them with one (Figure 1(d)). When selecting which auxiliary problem to present to the learner, the system looks at the learner's answer history and analyzes their error points. It then selects an auxiliary problem that is appropriate for that error.

The learner then works on the given auxiliary problem, and if they get it wrong a certain number of times, the system will use a similar procedure to give them an additional one. If the learner answers a given auxiliary problem correctly, then the system will gradually present more difficult problems from the auxiliary problem list. The goal is for the learner to eventually be able to solve the original problem.

To develop this system, we needed a set of auxiliary problems appropriate for the treatment of each error point, as well as a series of problems organized based on the differences between them. We designed and implemented several auxiliary problems and their problem sequences in EBS.

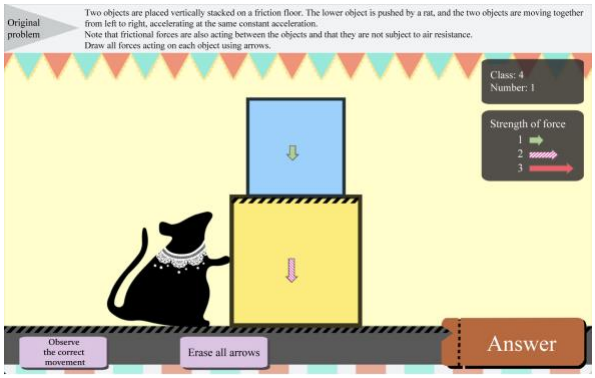
This system was implemented on a tablet. By implementing it in this way, we were able to make adjustments that made it easier for junior high school students to input data, a conclusion based on the opinions of those who had to experience teaching with the system. In this system, learners are first presented with a target problem, which they work on.

2.1 Auxiliary Problems

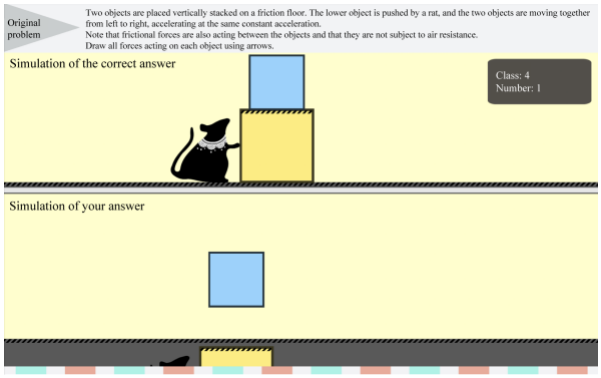
This section describes the auxiliary problems handled by the system. Figure 2 shows all the problems implemented in the system and their problem sequence.

We created auxiliary problems to help learners learn each of the forces in the original that were missing in their answers. The base of the auxiliary problems in this sequence is the "Original problem" shown in Figure 2. In it, two objects are placed on top of each other on a floor, and the lower object is pushed sideways by an external force, causing it to accelerate. There are nine forces at work in this problem that is categorized by the element, such as gravity, vertical drag, and so on. In this example, forces X and P are gravity, forces W and R are vertical drag forces, force S is the force exerted by the upper object on the lower object, force T is the applied external force, and forces V, Q, and H are friction forces.

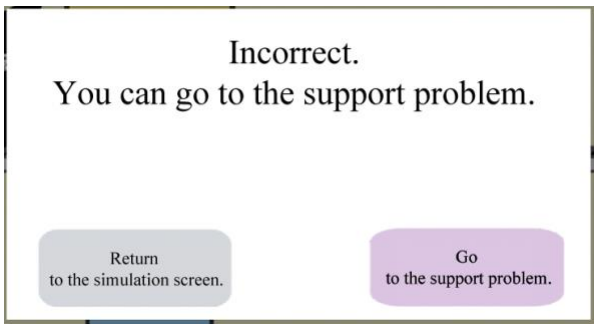
Auxiliary problems were then created for each of these elements so that the learner's attention could be drawn to specific ones by solving the problems related to them. This system implements a series of eight problems, including the seven auxiliary problems created in this way and the original problem.



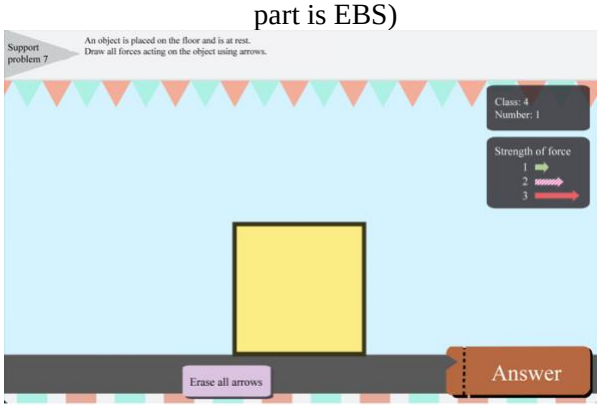
(a) Original problem presentation screen.



(b) The simulation screen presenting the solution in (a). (Upper part is correct simulation, lower part is EBS)



(c) Selection screen for auxiliary problem presentation.



(d) Support problem 7 presentation screen.

Figure 1. System screen.

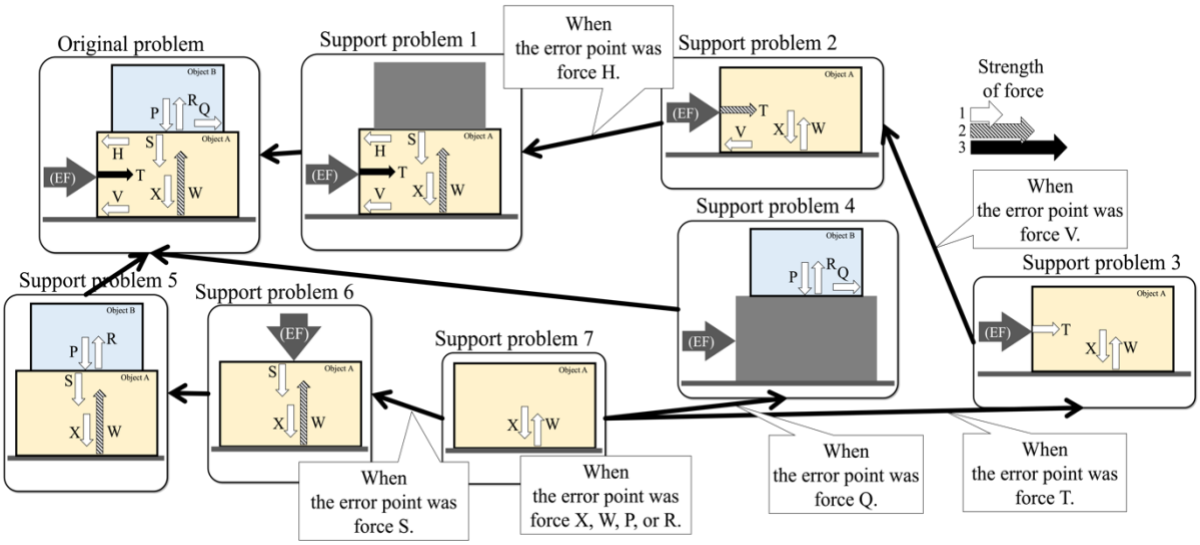


Figure 2. Problem sequence implemented in the system.

2.2 Analysis of Error Points

This section describes the method used by this system to analyze the learner's error point. Since our definition of impasse involves repeatedly making similar mistakes over a long period of time. We have the system look at the learner's answer history and analyze the error point by counting the occurrences of the most frequent errors.

Specifically, the system compares the forces represented in the learner's drawing as an answer with those in the correct answer and counts how many are missing. For example, for the original problem in Figure 2, the acting forces include gravity, vertical drag, a frictional force, the force exerted by the upper object on the lower one, and the applied external force. If the learner's answer omits the force S, diagnosed as "the learner missed the force exerted by the upper object on the lower object."

The system performs this kind of analysis for each answer given by the learner and extracts the forces that are excluded the most often in the learner's answer history. It then uses the most frequently missed as the learner's error point. For example, if a learner gives a solution five times and misses force S four times, force Q three times, and force W once, the system counts them and assumes that the force S missed most frequently is the error point.

When multiple forces are tied for the highest omission count, which is the learner's error point, then each of them is assigned a priority in advance, and the auxiliary problem corresponding to the one with the higher priority is presented. This priority was set in such a way as to eliminate conflicts: gravity has the highest priority, followed by the external forces, vertical drag, the force exerted by the upper object on the lower object, and the frictional forces.

2.3 Auxiliary Problem Presentation

In this section, we describe how the system presents auxiliary problems. The system compares the learner's error point from Section 2.2 with the problem sequence in Figure 2 and then presents the corresponding auxiliary problem. The problem sequence in Figure 2 shows which auxiliary problem is presented when a force is missing, and which auxiliary problem is to be worked on next if the correct answer to the current auxiliary problem is given.

Specifically, the system first presents the original problem. And if the learner answers incorrectly, the system analyzes the error point in the method described in Section 2.2. Next, the system refers to the arrow corresponding to that error point from Figure 2. The system then presents in turn the two auxiliary problems connected by that arrow. If the learner again answers the auxiliary problem incorrectly, the system presents the auxiliary problem corresponding to the error point in the same way. If the learner correctly answers the auxiliary problem, the system returns to the original problem following the arrows in Figure 2.

For example, if the learner continues to make errors about the force S, the system will refer to the arrow "when the error point was force S" in Figure 2. This arrow connects from support problem 7 to support problem 6. Therefore, support problem 7 is presented first, followed by support problem 6. Support Problem 7 is a problem where the object is stationary on the floor. And support problem 6 is a problem in which the top object is removed, and an external force is applied to the bottom object from above. By working on these two auxiliary problems, learners begin to associate the force exerted by the upper object with the one acting downward on the lower object. If the learner correctly answers support problems 6 and 7, the system then presents support problem 5. However, if the learner makes an error on the support problem, the system again analyzes the error point as in Section 2.2 and presents the auxiliary problem, referring to the arrow corresponding to the error point.

The learner works through the aids presented in this way. The system eventually returns to the original problems at the end, at which point the learner is expected to be able to answer them.

3. Classroom Practice and Evaluation

3.1 Overview

Three third-grade classes (29 students per class) at a public junior high school were given a three periods (135 minutes) practice using our system. Classroom practice was conducted in the following order: greetings and explaining classes (41 minutes), a pre-test (10 minutes), learning with the system (64 minutes), and a post-test (10 minutes). At the end of the session, the students completed a questionnaire about the system in 10 minutes. The participants were students aged 14-15 years old; omitting those who were absent, the total number of participants was 81.

Learning effectiveness was evaluated based on the results of the pre-test and the post-test. The test consisted of seven problems, four of which (TP1, TP3, TP4, and TP5) were problems covered by

the system (Learning task) and three of which (TP2, TP6, and TP7) were applied problems that were not covered by the system (Complex task). The content of the pretest and posttest were the same, and the maximum score was 7 (1 point per problem). The students' impressions of the system and the auxiliary problems were also analyzed from the questionnaire responses.

3.2 Test Results

The results of the pre-test and the post-test will be discussed. First, the total results show that most students improved their scores through the use of the system, with a mean score (SD) of 0.64 (0.93) on the pre-test and 3.41 (1.05) on the post-test. No student's score declined from the pre-test to the post-test. The t-test results detected a significant difference between the pre-test and the post-test of $p < .001$. Furthermore, the learning task and the complex task shows the mean score (SD) for the pre-test for all four problems was 0.59 (0.81), while the mean score (SD) for the post-test was 3.11 (SD), indicating a significant improvement in performance. On the other hand, performance on complex tasks also improved, with a pre-test of 0.05 (0.22) and a post-test of 0.30 (0.51).

Some of the test problems are shown in Figure 4, and both the number of correct answers per problem and the percentage of correct answers are shown in Figure 3. A problem that was considered too difficult for junior high school students (Figure 4(c)) was implemented as the original problem, and exercises were conducted. However, more than half of the students answered the problems correctly during the class and 32 students answered them correctly on the test, while none answered them correctly on the pre-test. This suggests that even problems that are initially unsolvable can become solvable by using our system. Additionally, 32 out of 81 students correctly answered the most basic problem (Figure 4(a)) in the pre-test, while all 81 students correctly answered it in the post-test.

Furthermore, 69 students were able to answer test problem 4 (Figure 4(b)) correctly on the post-test, compared to only 2 on the pre-test. This suggests that the system can be providing support in solving difficult problems.

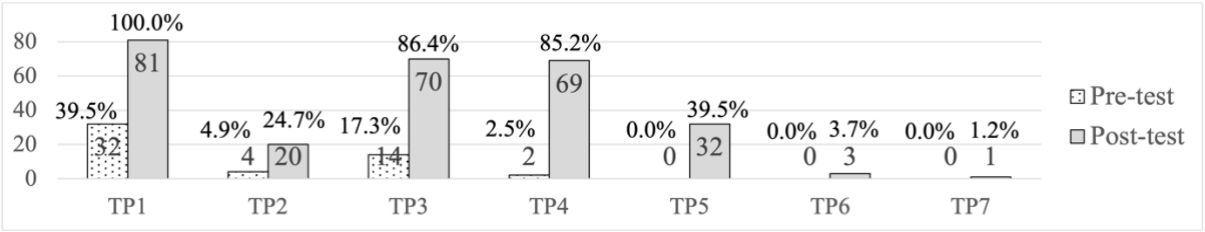
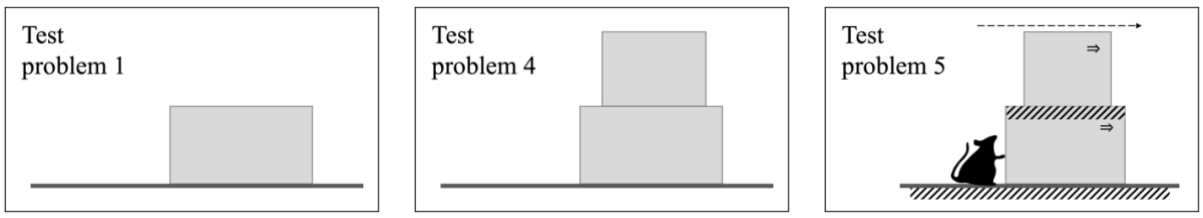


Figure 3. Number of correct answers per test problem (N = 81).



(a) The most basic problems dealt with in this class.
(b) The problem of two objects piled up.
(c) Highly difficult problem.

Figure 4. Problems dealt with in this class.

3.3 Questionnaire Results

The questionnaire was administered using the six-point scale from "1 strongly disagree" to "6 strongly agree," and the average of the responses was calculated. The questions asked were related to the tasks in this study and our system. For the sake of clarity, we refer to this system in this practice session questionnaire as the "app" and to the auxiliary problems as the "support problems."

First, "Do you think learning problems that focus on areas you don't understand will help you to understand them?" was found to receive a high rating of 5.78. Next, "Do you think the support problems presented by the app will help you in comprehending the parts you don't understand?" was found to receive a high rating of 5.46. These results suggest that students felt as though the auxiliary problems

helped them to understand the parts they were unfamiliar with. Then, "Did you feel that you were presented with support problems adaptive to your errors?" was found to receive a high rating of 5.11. This result suggests that students feel the support problems adapt their error. And "If you don't understand a problem, do you think trying to solve a simpler version of it will help you to understand the original problem?" was found to receive a high rating of 5.52. Finally, "Do you think the support problems helped you solve the original problem?" was found to receive a high rating of 5.52. This result suggests that students feel the support problems helped them solve the original problem.

3.4 Discussion

The results of 3.2 and 3.3 suggest that our system has a significant impact on learning. Additionally, materials called "Additional Missions" were handed out to students who completed the exercises during practice time. In total, 61 copies of the additional mission forms were distributed, indicating that 61 of the 81 students had completed the original problem within the time available for the exercise. However, none of the students answered the original problem correctly the first time, suggesting that the system was able to help learners learn to solve the problems they could not solve.

4. Conclusion

In this study, we conducted an evaluation experiment at a junior high school on an EBS system that adaptively presents auxiliary problems to learners who get stuck.

Future work includes analyzing logs and investigating learning activities. In addition, the current problem sequence is manually generated by focusing individually on each force, but in the future, we plan to investigate a method for automatically generating auxiliary problems from the target problems (Aikawa et al., 2021).

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP21H03565 and JP19H04227.

References

Aikawa, N., Koike, K., & Tomoto, T. (2020). Proposal and preliminary evaluation of a system that presents auxiliary problems to break learners' impasse based on tendency of the error in Error-based Simulation. *The IEICE Transactions on Information and Systems (Japanese edition)*, 103(9), 644-647. (*in Japanese*)

Aikawa, N., Koike, K., Tomoto, T., Horiguchi, T., & Hirashima, T. (2021). Characterization of Auxiliary Problems for Automated Generation in Error-Based Simulation. In *International Conference on Human-Computer Interaction* (pp. 3-13). Springer, Cham.

Hirashima, T., Horiguchi, T., Kashihara, A., & Toyoda, J. (1998). Error-based simulation for error-visualization and its management. *International Journal of Artificial Intelligence in Education*, 9(1-2), 17-31.

Hirashima, T., Niitsu, T., Hirose, K., Kashihara, A., & Toyoda, J. I. (1994). An indexing framework for adaptive arrangement of mechanics problems for ITS. *The IEICE Transactions on Information and Systems*, 77(1), 19-26.

Hirashima, T., Shinohara, T., Yamada, A., Hayashi, Y., & Horiguchi, T. (2017). Effects of error-based simulation as a counterexample for correcting MIF misconception. In *International Conference on Artificial Intelligence in Education* (pp. 90-101). Springer, Cham.

Horiguchi, T., Hirashima, T., & Okamoto, M. (2005). Conceptual Changes in Learning Mechanics by Error-based Simulation. In *Proceedings of the 13th International Conference on Computers in Education: ICCE2005* (pp. 138-145).

Horiguchi, T., Imai, I., Toumoto, T., & Hirashima, T. (2014). Error-based simulation for error-awareness in learning mechanics: An evaluation. *Journal of Educational Technology & Society*, 17(3), 1-13.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153-189.

Yamamoto, S., Tobe, Y., Tawatsuji, Y., & Hirashima, T. (2021). In-process Feedback by Detecting Deadlock based on EEG Data in Exercise of Learning by Problem-posing. In *Proceedings of the 29th International Conference on Computers in Education: ICCE2021* (pp. 21-30).

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Prerequisite-driven Q-matrix Refinement for Learner Knowledge Assessment: A Case Study in the Online Learning Context

Wenbin Gan[a*], Yuan Sun[b]
[a]*National Institute of Information and Communications Technology, Tokyo, Japan*
[b] *National Institute of Informatics, Tokyo, Japan*
*wenbingan@nict.go.jp

Abstract: The ever-growing abundance of learning traces in the online learning platforms promises unique insights into the learner knowledge assessment (LKA), a fundamental personalized-tutoring technique for enabling various further adaptive tutoring services in these platforms. Precise assessment of learner knowledge requires the fine-grained Q-matrix, which is generally designed by experts to map the items to skills in the domain. Due to the subjective tendency, some misspecifications may degrade the performance of LKA. Some efforts have been made to refine the small-scale Q-matrix; however, it is difficult to extend the scalability and apply these methods to the large-scale online learning context with numerous items and massive skills. Moreover, the existing LKA models employ flexible deep learning models that excel at this task, but the adequacy of LKA is still challenged by the representation capability of the models on the quite sparse item-skill graph and the learners' exercise data. To overcome these issues, in this paper we propose a prerequisite-driven Q-matrix refinement framework for learner knowledge assessment (PQRLKA) in online context. We infer the prerequisites from learners' response data and use it to refine the expert-defined Q-matrix, which enables the interpretability and the scalability to apply it to the large-scale online learning context. Based on the refined Q-matrix, we propose a Metapath2Vec enhanced convolutional representation method to obtain the comprehensive representations of the items with rich information and feed them to the PQRLKA model to finally assess the learners' knowledge. Experiments conducted on three real-world datasets demonstrate the capability of our model to infer the prerequisites for Q-matrix refinement, and also its superiority for the LKA task.

Keywords: Learner Knowledge Assessment, Prerequisite Inference, Q-matrix Refinement, Graph Representation Learning

1. Introduction

The popularity of online learning has increased in recent years, with increasingly many intelligent tutoring systems (ITSs) becoming available to learners (Hasanov, Laine, &Chung, 2019; Gan, Sun, & Sun, 2020; Gan, Sun, & Ye, 2019a, 2019b). The application of techniques from artificial intelligence and cognitive psychology has brought increasing intelligence to these systems (Hasanov et al., 2019), and has led to a very wide range of advantages, e.g., enabling the knowledge-acquisition through a series of individualized learning activities that accommodate learners with different needs and knowledge proficiencies (Gan, Sun, Peng, & Sun, 2020; Q. Liu, Shen, Huang, Chen, & Zheng, 2021). A key technique underlying these adaptive tutoring services is LKA (Gan, Sun, Peng, & Sun, 2020; Gan, Sun, & Sun, 2022b, 2022a).

LKA assesses learner knowledge in the granularity of a set of specific skills, also called knowledge concepts, based on the learners' exercising logs in the systems, thereby providing the detailed information about their strength and weakness. Detailed assessment of skills requires a fine-grained mapping of items to skills, i.e., the Q-matrix (M. Desmarais, Beheshti, & Xu, 2014). A binary Q-matrix indicates how the skills are combined to correctly answer the items, hence bridging the latent skill mastery patterns of learners and their explicit responses. However, building the Q-matrix for a

domain is a non-trivial task (Desmarais, Xu, & Beheshti, 2015; Desmarais & Naceur, 2013), which requires the involvement of domain experts. It is widely recognized that this subjective process may consist of some misspecifications (Chiu, 2013; Desmarais et al., 2014), which may negatively degrade the estimation of learner knowledge (Chen, 2017; Jiang et al., 2021). To alleviate the subjective bias in manually-defined Q-matrix, existing researches in the data mining and cognitive diagnostics fields have proposed various methods to infer the Q-matrix in a data-driven manner (J. Liu, Xu, & Ying, 2012), further refine the existing Q-matrix (M. Desmarais et al., 2014; Desmarais et al., 2015; Chiu, 2013; Kang, Yang, & Zeng, 2019), and validate the Q-matrix by combining with the cognitive diagnostic models (CMD) (Ma & de la Torre, 2020; de la Torre & Chiu, 2016). They also verified that in most of the cases, the refined Q-matrix has better model fit on the data than the original expert-labeled Q-matrix (M. C. Desmarais & Naceur, 2013; Matsuda, Furukawa, Bier, & Faloutsos, 2015). However, most of the efforts have been devoted to the scenarios of small tests with quite limited number of skills and items. Even for the famous fraction-subtraction dataset with only twenty items and eight skills, the obtained Q-matrix is still controversial (Kang et al., 2019). It is difficult, and sometimes impractical, to extend the scalability and apply these methods to the large-scale online learning context with numerous items and massive skills.

Moreover, most of the existing research refine the Q-matrix by considering a set of Q-matrices in the latent space and determining the one with superior model fitness on the data using some fit strategies (Kang et al., 2019). Typical studies are the maxDiff (De La Torre, 2008), minRSS (Chiu, 2013), and the ALS (M. C. Desmarais & Naceur, 2013). These researches showed good performance on the learner data, but they generally ignore the explainable structural and inherent information among the questions and skills. Actually, the interdependencies between the skills in a domain have long been acknowledged as prerequisites in both cognitive and education science (P. Chen, Lu, Zheng, & Pian, 2018). Correspondingly, a question with a specific skill also requires the prerequisite skills. These prerequisites between pedagogical concepts can be represented as a knowledge graph (KG) (Q. Liu et al., 2019), which contains additional structural information about the learning domain, and provides the potential for further refining the Q-matrix. Nevertheless, the prerequisites specifying the structure of skills has rarely been applied in the refinement of Q-matrix because obtaining the KG of a domain is labor-intensive (Pan, Li, Li, & Tang, 2017; Zhang & King, 2016; Nakagawa, Iwasawa, & Matsuo, 2019). Fortunately, the ever-growing abundance of learning traces in the online learning platforms further enhances our capacity to obtain the prerequisites (KG) directly from the data through data mining techniques. Based on this idea, we infer the prerequisites between skills from the data and use them to refine the Q-matrix, and apply the refined Q-matrix to assess the knowledge proficiency of learners. We try to balance the interpretability of Q-matrix refinement and the performance in applying it to the LKA task.

To assess learner knowledge in the online learning systems, plenty of researches have been conducted leveraging the compelling attribute of deep learning methods, such as the deep knowledge tracing (DKT) (Piech et al., 2015) and Dynamic key-value memory networks for KT (DKVMN) (Zhang, Shi, King, & Yeung, 2017), and excelled at this task. However, most of these models used the underlying skills as input (also termed as skill-level KT models) and did not distinguish questions containing the same skills. The loss of distinctive information related to the questions may lead to imprecise inferences of the learners' knowledge states (Ghosh, Heffernan, & Lan, 2020; Gan et al., 2022b). To enrich the question representation, graph KT methods, such as the graph-based interaction KT (GIKT) (Y. Yang et al., 2020) and graph for KT (GKT) (Nakagawa et al., 2019), have been proposed to learn the embedding of questions from the question-skill graph. However, they ignore the prerequisite information of the skills in the domain, and the sparseness of the original question-skill graph and the learner data greatly limit the capability of the graph learning methods. To overcome this sparseness, in this paper we refine the Q-matrix and propose a graph representation learning model to enrich the embedding learning from the prerequisite-enhanced question–skill graph and integrate it into the LKA.

To summarize, in this paper we refine the Q-matrix for learner knowledge assessment in online learning context and propose a prerequisite-driven Q-matrix refinement framework for LKA (PQRLKA). Specifically, we first explore eight methods to infer the prerequisites from learners' response data in the online learning systems and use it to refine the expert-defined Q-matrix to eliminate the potential subjective tendency of experts in designing the Q-matrix. This Q-matrix refinement method offers two advantages over the existing methods: first, it leverages the additional structural

information about the learning domain to refine the Q-matrix rather than use the model fit indices to brutally change some entries from zeros to ones, thus enabling the interpretability of this process.

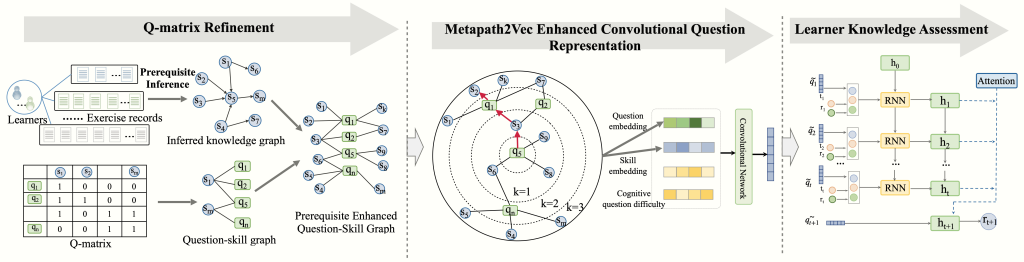


Figure 1: The framework of the proposed method

Second, it refines the Q-matrix by applying the inferred prerequisites in a date-driven manner without relying on the specific cognitive diagnosis models (i.e., model-independent), and hence has good scalability. Based on the refined Q-matrix, we propose a graph representation learning model for LKA to verify its effectiveness in the online learning scenarios. A Metapath2Vec (Dong, Chawla, & Swami, 2017) enhanced convolutional representation method is then proposed to obtain the comprehensive representations of the attempted questions with rich information. These representations for the learners' exercising sequences are fed into the PQRLKA model, which considers the long-term dependencies using an attention mechanism, to finally predict the learners' performance on new problems. The main contributions are listed below.

- We explore eight methods that automatically discover the domain prerequisites from learner response data and leverage it to refine the expert-defined Q-matrix.
- We propose the PQRLKA framework to assess the learner knowledge with a Metapath2Vec enhanced convolutional representation method for comprehensive question representations with rich information.
- We integrate the Q-matrix refinement with the task of learner knowledge assessment, joint these two tasks together and leverage the former to benefit the later.
- We conduct comprehensive experimental evaluations on three real-world datasets. The results demonstrate the superiority and effectiveness of our method in Q-matrix refinement and learner knowledge assessment in the online learning context.

2. Proposed Method

This section introduces the proposed PQRLKA framework for assessing learner knowledge, as shown in Figure 1. The framework proceeds the LKA task using three modules: Q-matrix refinement, Metapath2Vec enhanced convolutional question representation, and learner knowledge state evolution. It first builds a KG for the learning dom ain by inferring the prerequisite relations from the learner response data, and then uses it to refine the expert-defined Q-matrix, thus generating a refined Q-matrix, which can be naturally represented as a prerequisite enhanced question–skill graph. Based on this enhanced graph, it conducts the question representation learning based on a Metapath2Vec enhanced convolutional neural network. To incorporate more distinctive information, A convolutional representation method is then proposed to fuse the question and skill embeddings with the question difficulty, thus obtaining the comprehensive representations of questions (see Figure 3). Compared with the existing methods, the proposed question representation learning method incorporates both the distinctive information of the specific question and the various relations between questions and skills, a factor making it more superior for the later assessment of learner knowledge. These representations are then fed into the attentive LKA network for predicting learner performance. It is worth noting that we conduct the Q-matrix refinement and the learner knowledge assessment together.

2.1 Q-matrix Refinement

2.1.1 Prerequiite Inference from Data

In practical educational scenarios, there always exists a topological order (prerequisites) among the skills in a domain, because skills are taught and learned in sequence. Here we infer the prerequisites

Table 1: *Contingency table for a pair of skills s_i and s_j*

	s_j master	s_j not master	total
s_i master	a	b	a+b
s_i not master	c	d	c+d
total	a+c	b+d	a+b+c+d

Table 2: *Evaluation indices for obtaining KG using the contingency*

Evaluation Index	Formula
Cohen's Kappa	$R_{i,j}^{Kappa} = 2(ad-bc)/\{(a+b)(b+d)+(a+c)(c+d)\}$
Adjusted Kappa	$R_{i,j}^{Kappa'} = 2(ad-bc)/\{(a+c)(c+d)\}$
Phi coefficient	$R_{i,j}^{Phi} = (ad-bc)/\sqrt{(a+b)(b+d)(a+c)(c+d)}$
Yule coefficient	$R_{i,j}^{Yule} = (ad-bc)/(ad+bc)$
Ochiai coefficient	$R_{i,j}^{Ochiai} = a/\sqrt{(a+b)(a+c)}$
Sokal coefficient	$R_{i,j}^{Sokal} = (a+d)/(a+b+c+d)$
Jaccard coefficient	$R_{i,j}^{Jaccard} = a/(a+b+c)$

from the order of the learners' mastery of skills, which is explicitly represented by the exercising performance data. We first build a skill relation matrix $R \in R^{|S|\times|S|}$, in which entry $R_{i,j}$ represents the prerequisite relation $s_i \rightarrowtail s_j$ between skill s_i and s_j. Inspired by the definition of question similarity in previous methods (Tong et al., 2020; Pel´anek, 2019), we explore the underlying KG using eight methods: Skill Transition, Cohen's Kappa, Adjusted Kappa, Phi coefficient, Yule coefficient, Ochiai coefficient, Sokal coefficient, Jaccard coefficient, as described below.

Skill Transition: The skill-transition matrix R contains the transitions of different skills. Its entries are $R_{i,j}^{SK} = \frac{n_{i,j}}{\sum_{k=1}^{|K|} n_{i,k}}$, where $n_{i,j}$ denotes the number of times in which skill s_j is trained immediately after training skill s_i.

To further leverage the impact of the learners' performance of one skill on the performance of another, we summarized the learners' performance on skill pair s_i and s_j in a contingency table (see Table 1). As mentioned above, we interpreted the learners' correct or incorrect responses as mastery indicators of the underlying skills of the given questions. It is worth noting that in Table 1, the question requiring skill s_i occurs before the question requiring s_j in the learning sequence. Based on the contingency table, we discovered the KG using the evaluation indices (Pel´anek, 2019) in Table 2, which measure the agreement of the prerequisite relation between a pair of skills.

As the KG is always a unidirectional graph, we simplified it by a suitable strategy. We also imposed a threshold that controlled the sparsity of the relations in KG. The final skill relation matrix was denoted as R^w, $w \in \{SK, Kappa, Kappa', Phi, Yule, Ochiai, Sokal, Jaccard\}$. The elements along the diagonal of R^w were set to one.

$$\begin{cases} R_{i,j}^w = max(R_{i,j}^w, R_{j,i}^w), R_{j,i}^w = 0, \ if \ R_{i,j}^w \geq R_{j,i}^w \\ R_{j,i}^w = max(R_{i,j}^w, R_{j,i}^w), R_{i,j}^w = 0, \ otherwise \end{cases} \quad (1)$$

$$R_{i,j}^w = \begin{cases} 1, \ if \ R_{i,j}^w \geq threshold \\ 0, \ otherwise \end{cases} \quad (2)$$

2.1.2 Refine the Expert-defined Q-matrix

After obtaining the inferred skill relation matrix $R^w \in R^{|S|\times|S|}$ of the KG that includes all the prerequisites among skills, we refine the original expert-defined q-matrix $O \in R^{|Q|\times|S|}$. The matrix representation $\hat{O} \in R^{|Q|\times|S|}$ of the refined new Q-matrix is obtained as follows:

$$\hat{O} = O(R^w)^T \quad (3)$$

$$\hat{O}_{i,j} = \begin{cases} 1, \ if \ \hat{O}_{i,j} \geq 1 \\ 0, \ otherwise \end{cases} \quad (4)$$

In Eq. (3), the transpose of the skill relation matrix R^w accounts for the skills that are prerequisite to the skills of the current question. In other words, a question requiring a specific skill is also related to the prerequisite skills. The procedure of the refinement can be shown in Figure 2.

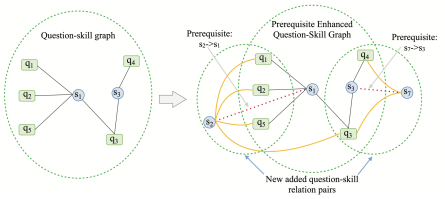


Figure 2: The procedure of Q-matrix refinement.

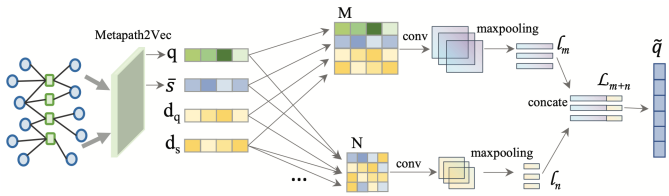


Figure 3: Use of Metapath2Vec enhanced convolutional question representation

2.2 Learner Knowledge Assessment Based on Refined Q-matrix

2.2.1 Metapath2Vec Enhanced Convolutional Question Representation

The matrix $\widehat{\mathrm{O}}$ can be naturally represented as a graph, hereafter denoted as the KG-enhanced question--skill graph. This graph includes not only the multi-hop relations between questions and skills, but also the prerequisite relations among the skills. The embeddings in this graph were obtained using the adapted Metapath2Vec method (Dong et al., 2017).

The adapted Metapath2Vec method proceeds in two main steps: meta-path generation and skip-gram-based embedding learning. A meta-path is a sequence of nodes following the edges in the graph. To assure that all questions and skills appear in the final embeddings, we generate the meta-paths from the KG-enhanced question--skill graph using a question--skill--question (QSQ) pattern, in which every meta-path begins with a question node followed by a skill node and then by a question node; for example, ρ: $\mathrm{q}_1 \xrightarrow{\hat{O}_{1,1}} s_1 \xrightarrow{\hat{O}_{2,1}} q_2 \xrightarrow{\hat{O}_{2,2}} s_2 \xrightarrow{\hat{O}_{3,2}} q_3$. Setting two hyper-parameters---the path length $\wp$ and number of paths $\aleph$---for each question node, we generated all meta-paths on the graph. The heterogeneous skip-gram is leveraged to learn the node embeddings. The interested reader can refer to (Dong et al., 2017) for the original Metapath2Vec method. Finally, we obtain all embeddings (q for questions and s for skills) with the same dimension d of question and skill nodes in the graph, which also enclose the relation information.

To distinctively represent the questions, we fuse the various feature representations through convolutional operations to obtain the comprehensive question embeddings with additional distinctive information, i.e., the skill information and the cognitive difficulty information (Gan, Sun, Ye, Fan, & Sun, 2019). The cognitive difficulty d_q and d_s (represented as vectors) is calculated following prior work in (Gan, Sun, Peng, & Sun, 2020). Following (Y. Liu et al., 2020) and (S. Yang, Zhu, Hou, & Lu, 2020) that learn the high-order latent patterns through feature interactions and convolution operations (rather than directly concatenating the features), we map and fuse the separate features and their interactions using convolution operations, as shown in Figure 3. For questions containing multiple skills, we represent the skill as the average skill embedding $\overline{s} = Avg(s_1, \dots, s_q)$. Fusing the above-obtained features, we generate the linear information M and quadratic information N for question q.

$$M = [q, \overline{s}, d_q, d_s] \in R^{4\times d}, \quad N = [\langle M_i, M_j \rangle] \in R^{4\times 4}, \tag{5}$$

where $\langle . \rangle$ represents the interactions of two vectors obtained by the inner product. We then apply the two-dimensional convolution operation with eight kernels of size 2×2 on both M and N, and maxpooling on each feature map to obtain $l_m \in R^{1\times(d-1)\times 8}$ and $l_n \in R^{1\times 3\times 8}$. These two parts are then concatenated into eight longer vectors $l_{m+n} \in R^{1\times(d+2)\times 8}$ including the convolutions from the separate features and their interactions. Inspired by the multi-head mechanism in the transformer model (Vaswani et al., 2017), we concatenate and linearly transform the eight vectors and hence obtain the final question representation $\tilde{q} \in R^{1\times d'}$.

$$\tilde{q} = Concat(l_{m+n}^1, \dots, l_{m+n}^8) W^O \tag{6}$$

where $W^O \in R^{((d+2)\times 8)\times d'}$ is the parameter that transforms the convolution results into a vector.

2.1.2 Learner Knowledge State Evolution

Since the dynamic evolution of learner knowledge in the online learning systems, the learner exercising sequences are fed into an attentive KT framework that predicts the learner performance, as shown in the right part of Figure 1.

The log data of each interaction in the exercising sequences consists of a tuple representing the question, the correctness, and the elapsed time. Look-up operations are performed on an embedding matrix $E_r \in R^{2 \times d'}$, in which row vector r_t contains the incorrectness or correctness of the responses. The elapsed time et strongly evidences a student's proficiency in knowledge and skills (Shin et al., 2020). This time is converted to seconds and capped at 500 seconds. A d'-dimensional latent embedding vector for et_k is computed as $t_k = et_k W_{et} + b_{et}$, where W_{et} and b_{et} are learnable vectors. The interaction embedding is obtained as $x_t = Concat(\tilde{q}, t_t, r_t)$.

Table 3: *Statistics of the three datasets used in this study*

Dataset	Assist0910	Assist1213	Ednet
# of learners	3,002	22,339	5,000
# of questions	17,705	52,825	12,372
# of skills	123	265	188
# of interaction records	277,540	2,672,532	347,864
# skills per item	1.20	1.00	2.28
# questions per skill	172.33	199.34	149.78
# of attempted items per learner	92.45	119.64	69.57

The sequence data of the learners' exercising process are modeled using LSTM (Piech et al., 2015) within the KT framework, The hidden knowledge state h_t of a learner at step t is updated based on the current input and the previous state as: $h_t = LSTM(x_t, h_{t-1}; \theta)$.

We then employ an attention mechanism that accounts for the impact of previous attempts on the current attempt. A new question will likely be strongly affected by similar questions or questions requiring the same skillset as the new question. To describe these effects, we assume that the learner-knowledge state in the current step is the weighted sum of the aggregated states in the previous steps. The weights are based on the correlations:

$$h_{t+1} = \sum_{i=1}^{t} \alpha_{i,t+1} h_i. \quad (7)$$

The attention $\alpha_{i,t+1}$ was calculated using a combination of the shared skill-based attention and the question similarity-based attention. The learner performance at step $t+1$ can be predicted from the question representation $\widetilde{q_{t+1}}$ and the current knowledge state h_{t+1} as follows:

$$s_{t+1} = tanh(W_s[\widetilde{q_{t+1}}, h_{t+1}]) + b_s, \quad p_{t+1} = \sigma\big(W_p s_{t+1} + b_p\big), \quad (8)$$

where W_* and b_* are parameters in the fully connected layer and the sigmoid activation layer, respectively. Finally, we optimized our model by the cross-entropy loss; specifically, we minimized the following objective function between the true answer l_t and the predicted performance p_{t+1} at each interaction: $\mathcal{L} = -\sum_t \big(r_{t+1} log\, p_{t+1} + (1 - r_{t+1}) log(1 - p_{t+1})\big)$.

3. Experiments

3.1 Datasets

E xperiments were performed on three well-established datasets, namely, Assist0910[1], Assist1213[2], and EdNet[3]. All datasets are public real-world datasets containing the temporal interaction records between learners and real computer-aided tutoring systems. For the EdNet dataset, following (Y. Yang et al., 2020), we randomly selected 5000 students who answered 12,372 questions requiring 188 skills, thus obtaining 347,866 interaction logs. As done in existing studies (Vie & Kashima, 2019; Y. Yang et al.,
2020), we eliminated noise by deleting users with fewer than 10 interaction entries and questions with NaN skills from the three datasets. Table 3 summarizes the statistics of the datasets.

3.2 Compared Model

As part of our model evaluation, we competed the model against several state-of-the-art skill- and question-based KT models: BKT (Corbett & Anderson, 1994), DKT (Piech et al., 2015), DKVMN (Zhang et al., 2017), KTM (Vie & Kashima, 2019), DKT-Q (a variant of DKT that replaces the skills

[1] Assist0910: https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data/skill-builder-data-2009-2010
[2] Assist1213: https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-with-affect
[3] EdNet: http://bit.ly/ednet_kt1

embedding with a one-hot encoding of questions as the input), DKT-Q&S (a variant of DKT that inputs both the questions and skill representations to the DKT), DKT-CQE (a variant of DKT that inputs our convolutional question embeddings to the DKT), GIKT (Y. Yang et al., 2020), DAS3H (Choffin et al., 2019), RKTM (Lai et al., 2021), and AKT (Ghosh et al., 2020).

Among the baseline models, the former three are skill-based models, in which the LKA is based on the skills contained in the questions. The latter eight and the proposed model are question-based models that account for the distinctive question information.

3.3 Setup and Implementation

Before conducting the experiments, we extracted 20% of the sequences in the dataset as the test set and retained the remaining 80% as the training set. To embed the nodes in the graph using Metapath2Vec, we set the length of all meta-paths as $\wp = 7$ and the number of paths as $\aleph = 100$ for each question node in the graph. The embedding dimension d of the skill and question representations was set to 128. The final dimension of the convolutional question representation was $d' = 256$. The size of the hidden layers of the LSTM was set to 256. The other hyperparameters were set through grid searching. The model was optimized using Adam optimization of the learning rate on a case-by-case basis in the three datasets. The norm clipping threshold and batch size were maintained at 10 and 64, respectively. Similar to the existing models, the sequence length of the model input was fixed at 200. The proposed model was implemented using TensorFlow. The other baselines were implemented with their best parameter settings, as specified in the original works. As the evaluation metric, we selected the area under the receiver operating characteristic (ROC) curve (AUC), which is widely used in existing studies.

4. Results and Analysis

4.1 Performance Prediction

The different models were evaluated by their performances in predicting the future learner scores from the estimated knowledge state. Table 4 presents the AUC results of all models on the three datasets. Our model outperformed the other models on all three datasets. Specifically, the AUC scores of the PQRLKA model were 0.8242, 0.7851, and 0.7754 on the Assist0910, Assist1213 and EdNet datasets, respectively, 3.97%, 1.39%, and 2.25%, respectively, above those of the state-of-the-art AKT model. Similarly, to the original DKT model, our model processes the time-series data using a recurrent neural network framework, but achieved 8.3%, 5.95%, and 8.65% higher AUCs than the DKT model on the Assist0910, Assist1213 and EdNet datasets, respectively.

The skill-based BKT model was the worst performer among the models because it tracks the mastery of each skill separately, without considering a contextual trial sequence of all skills. KTM framework, which incorporates several traditional models, typically obtained similar AUC scores to those of DKT and DKVMN. DKT extended with various input-question embeddings (DKT-Q, DKT-Q&S, and DKT-CQE) demonstrated noticeable performance differences. DKT-Q using the one-hot encoding of question representations performed much worse than the original DKT model, owing to the sparsity of question interactions in these datasets. DKT-Q&S and DKT-CQE decidedly outperformed the original DKT and DKT-Q models, consistent with our intuition that each question contains distinctive information even when it requires the same skills as one or more other questions in the dataset. Therefore, incorporating the distinctive question and skill information into the question representations can improve the model performance. The comparison between DKT and DKT-CQE also shows the effectiveness of the proposed convolutional question representation. DAS3H and RKTM are built on the KTM model by considering additional rich information and show better results than the original KTM model. Here we do not present the results of RKTM on the EdNet as the code for implementing RKTM is not published by the author. Our PQRLKA model outperforms GIKT and AKT, which also take the well-designed question embedding as input. This result validates the effectiveness of the Q-matrix refinement for the task of LKA.

	Model	ASSIST0910	ASSIST1213	EdNet
Skill-based Model	BKT	0.6571	0.6204	0.6027
	DKT	0.7412	0.7256	0.6889
	DKVMN	0.7559	0.7247	0.6921
Question-based Model	KTM	0.7582	0.7212	0.6899
	DKT-Q	0.7306	0.7158	0.6812
	DKT-Q&S	0.7616	0.7389	0.7235
	DKT-CQE	0.7998	0.7686	0.7523
	GIKT	0.7845	0.7712	0.7529
	DAS3H	0.789	0.741	0.731
	RKTM	0.7650	0.7690	–
	AKT	0.8152	0.7720	0. 7658
	PQRLKA	**0.8242***	**0.7851***	**0.7754***

Table 4: *Comparisons of the AUC results of different models on the three datasets*

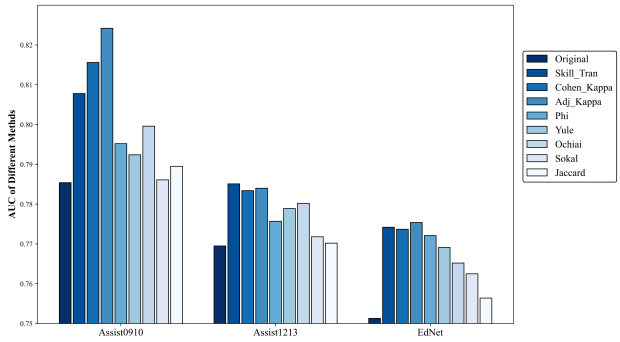


Figure 4: KS inferred from the learner response data

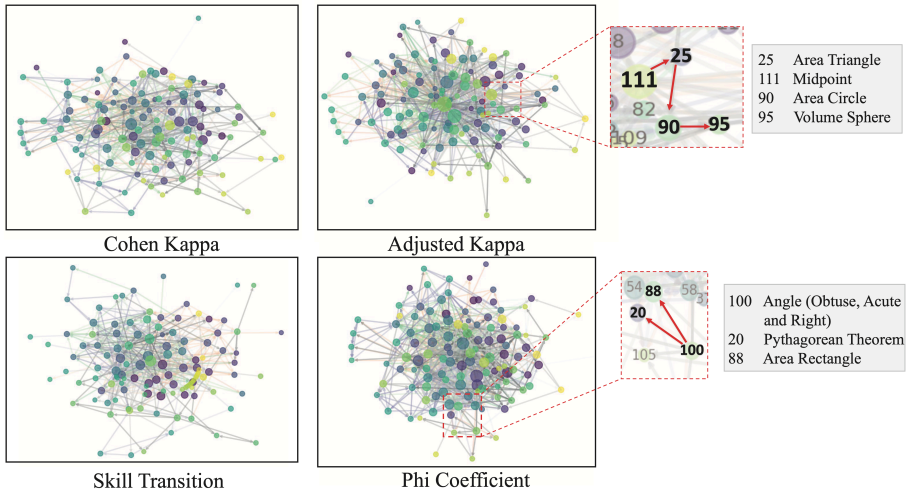


Figure 5: Visualization of the directed KG graphs generated by four methods on Assist0910

4.2 Model Analysis

4.2.1 Comparison of Prerequisite Inference by Different Methods

Figure 4 compares the AUC results of the predictions of nine methods on the three datasets. Here the "original" method represents the embedding learning on the original question--skill relation graph and the other eight methods are based on the prerequisite-enhanced graph. As evidenced in the figure, the eight KG enhanced methods generally outperformed the "original" method on all three datasets, validating the effectiveness of the KG enhanced graph in KT tasks. The adjusted Kappa yielded the best performance on both Assist0910 and EdNet, whereas the skill transaction method performed best on Assist1213.

4.2.2 Visualization of Inferred Prerequisites

Figure 5 illustrates the KGs generated from the prerequisites using four methods, inferred from the learner response data in Assist0910. The right-hand side of this figure enlarges a part of the graphs to show their local connections. The nodes and edges in the KG form dense graphs with similar structures, showing several interconnected nodes. These graphs also show some interesting properties. In the adjust-Kappa graph, four nodes (25, 111, 90, and 95) were locally interconnected and revealed a perfect ordering of the skills (prerequisite and post-requisite relations) in the geometry. The local connections in the Phi coefficient graph also presented reasonable relations among the three skills. These results confirm that our KG discovery methods can infer prerequisite skill pairs from the ordering of learners' mastery of skills.

5. Conclusions

In this paper we proposed a prerequisite-driven Q-matrix refinement framework for learner knowledge assessment in online learning context. We first explored eight methods to infer the prerequisites from learners' response data in the online learning systems and used it to refine the expert-defined Q-matrix to eliminate the potential subjective tendency of experts in designing the Q-matrix. This refinement leveraged the additional structural information (prerequisites) to allow the interpretability of the optimized Q-matrix while enabling the scalability to apply to the large-scale online learning context. Based on the refined Q-matrix, we proposed a Metapath2Vec enhanced convolutional representation method to obtain the comprehensive representations of the attempted items with rich information. These representations for the learners' exercising sequences are fed into the PQRLKA model, which considers the long-term dependencies using an attention mechanism, to finally predict the learners' performance on new items. Extensive experiments conducted on three real-world datasets demonstrated the capability and interpretability of our model to infer the prerequisites from the learning data, and the better performance of the embedding representation on the prerequisite enhanced graph, thus contributing to the superiority of the proposed model for learner knowledge assessment and validating its potential applicability to real online learning environments.

In future work, we intend to embed KG learning into our model in an end-to-end manner. The KG will then be automatically learned in the training process rather than computed from the learner performance data. Such automation can potentially expand the KG to new skills in the domain, for example, enlarging the KG of mathematics from the primary school to high school level.

Acknowledgements

This work is supported by JSPS KAKENHI with Grant Number 18K11597, 20H01720 and the grant-in-Aid for JSPS Fellows 20J15339.

REFERENCES

Baker, F. B. (2001). The basics of item response theory. ERIC.

Cen, H., Koedinger, K., & Junker, B. (2006). Learning factors analysis–a general method for cognitive model evaluation and improvement. In Its (pp. 164–175).

Chen, C. (2017). Q-matrix optimization for cognitive diagnostic assessment (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Chen, P., Lu, Y., Zheng, V. W., & Pian, Y. (2018). Prerequisite-driven deep knowledge tracing. In 2018 ieee international conference on data mining (ICDM) (pp. 39–48).

Chiu, C.-Y. (2013). Statistical refinement of the q-matrix in cognitive diagnosis. Applied Psychological Measurement, 37 (8), 598–618.

Choffin, B., Popineau, F., Bourda, Y., & Vie, J.-J. (2019). Das3h: Modeling student learning and forgetting for optimally scheduling distributed practice of skills. Proceedings of the 12th EDM (pp. 29–38).

Choi, Y., Lee, Y., Shin, D., Cho, J., Park, S., Lee, S., Heo, J. (2020). Ednet: A large-scale hierarchical dataset in education. In International conference on artificial intelligence in education (pp. 69–73).

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. User modeling and user-adapted interaction, 4 (4), 253–278.

De La Torre, J. (2008). An empirically based method of q-matrix validation for the dina model: Development and applications. Journal of educational measurement, 45 (4), 343–362.
De La Torre, J. (2009). Dina model and parameter estimation: A didactic. Journal of educational and behavioral statistics, 34 (1), 115–130.
Desmarais, M., Beheshti, B., & Xu, P. (2014). The refinement of a q-matrix: Assessing methods to validate tasks to skills mapping. In Educational data mining 2014.
Desmarais, M. C. (2012). Mapping question items to skills with non-negative matrix factorization. ACM SIGKDD Explorations Newsletter, 13 (2), 30–36.
Desmarais, M. C., & Naceur, R. (2013). A matrix factorization method for mapping items to skills and for enhancing expert-based q-matrices. In AIED (pp. 441–450).
Desmarais, M. C., Xu, P., & Beheshti, B. (2015). Combining techniques to refine item to skills q-matrices with a partition tree. International Educational Data Mining Society.
Dong, Y., Chawla, N. V., & Swami, A. (2017). metapath2vec: Scalable representation learning for heterogeneous networks. In Proceedings of the 23rd ACM SIGKDD (pp. 135–144).
Gan, W., Sun, Y., Peng, X., & Sun, Y. (2020). Modeling learner's dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing. Applied Intelligence, 50 (11), 3894–3912.
Gan, W., Sun, Y., & Sun, Y. (2020). Knowledge interaction enhanced knowledge tracing for learner performance prediction. In 2020 7th international conference on behavioural and social computing (BESC) (pp. 1–6).
Gan, W., Sun, Y., & Sun, Y. (2022a). Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent education systems. Neurocomputing, 488, 36-53.
Gan, W., Sun, Y., & Sun, Y. (2022b). Knowledge structure enhanced graph representation learning model for attentive knowledge tracing. International Journal of Intelligent Systems, 37 (3), 2012–2045.
Gan, W., Sun, Y., Ye, S., Fan, Y., & Sun, Y. (2019a). Field-aware knowledge tracing machine by modelling students' dynamic learning procedure and item difficulty. In 2019 ICDMW (pp. 1045–1046).
Gan W., Sun Y., Ye S., et al. (2019b). AI-tutor: Generating tailored remedial questions and answers based on cognitive diagnostic assessment, In 6th international conference on behavioral, economic and socio-cultural computing (BESC). IEEE, 2019: 1-6.
Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. In Proceedings of the 26th ACM SIGKDD (pp. 2330–2339).
Hasanov, A., Laine, T. H., & Chung, T.-S. (2019). A survey of adaptive context-aware learning environments. Journal of Ambient Intelligence and Smart Environments, 11 (5), 403–428.
Jiang, C., Gan, W., Su, G., et al. (2021). Improving Knowledge Tracing through Embedding based on Metapath. In Proceedings of the 29th ICCE (pp. 11–20).
Lai, Z., Wang, L., & Ling, Q. (2021). Recurrent knowledge tracing machine based on the knowledge state of students. Expert Systems, e12782.
Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). A survey of knowledge tracing. arXiv preprint arXiv:2105.15106.
Liu, Q., Tong, S., Liu, C., Zhao, H., Chen, E., Ma, H., & Wang, S. (2019). Exploiting cognitive structure for adaptive learning. In Proceedings of the 25th ACM SIGKDD (pp. 627–635).
Liu, Y., Yang, Y., Chen, X., Shen, J., Zhang, H., & Yu, Y. (2020). Improving knowledge tracing via pre-training question embeddings. arXiv preprint arXiv:2012.05031.
Ma, W., & de la Torre, J. (2020). An empirical q-matrix validation method for the sequential generalized DINA model. British Journal of Mathematical and Statistical Psychology, 73 (1), 142–163.
Matsuda, N., Furukawa, T., Bier, N., & Faloutsos, C. (2015). Machine beats experts: Automatic discovery of skill models for data-driven online course refinement. International Educational Data Mining Society.
Minn, S., Desmarais, M. C., & Fu, S. (2016). Refinement of a q-matrix with an ensemble technique based on multi-label classification algorithms. In EC-TEL (pp. 165–178).
Nakagawa, H., Iwasawa, Y., & Matsuo, Y. (2019). Graph-based knowledge tracing: modeling student proficiency using graph neural network. In 2019 IEEE/WIC/ACM WI (pp. 156–163).
Pan, L., Li, C., Li, J., & Tang, J. (2017). Prerequisite relation learning for concepts in MOOCs. In Proceedings of the 55th ACL (pp. 1447–1456).
Pandey, S., & Srivastava, J. (2020). RKT: Relation-aware self-attention for knowledge tracing. In Proceedings of the 29th ACM international conference on information & knowledge management (pp. 1205–1214).
Pavlik Jr, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis–a new alternative to knowledge tracing. Online Submission.
Pel´anek, R. (2019). Measuring similarity of educational items: An overview. IEEE Transactions on Learning Technologies, 13 (2), 354–366.
Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., & Sohl-Dickstein, J. (2015). Deep knowledge tracing. In Advances in neural information processing systems (pp. 505–513).

Shin, D., Shim, Y., Yu, H., Lee, S., Kim, B., & Choi, Y. (2020). Saint+: Integrating temporal features for ednet correctness prediction. arXiv preprint arXiv:2010.12042.
Tong, H., Zhou, Y., & Wang, Z. (2020). HGKT: Introducing problem schema with hierarchical exercise graph for knowledge tracing. arXiv preprint arXiv:2006.16915.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. (2017). Attention is all you need. arXiv preprint arXiv:1706.03762.
Vie, J.-J., & Kashima, H. (2019). Knowledge tracing machines: Factorization machines for knowledge tracing. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, pp. 750–757).
Yang, S., Zhu, M., Hou, J., & Lu, X. (2020). Deep knowledge tracing with convolutions. arXiv preprint arXiv:2008.01169.
Yang, Y., Shen, J., Qu, Y., Liu, Y., Wang, K., Zhu, Y., Yu, Y. (2020). Gikt: A graph-based interaction model for knowledge tracing. arXiv preprint arXiv:2009.05991.
Zhang, J., Shi, X., King, I., & Yeung, D.-Y. (2017). Dynamic key-value memory networks for knowledge tracing. In Proceedings of the 26th www (pp. 765–774).

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Transforming Brainwave Signals into Symbolic Strings Towards Academic Emotion Recognition

Judith AZCARRAGA[a] & Juan Francesco SALCEDA[a]*
[a]*De La Salle University, Manila, Philippines*
*juan_francesco_salceda@dlsu.edu.ph

Abstract: Students tend to experience varying academic emotions while engaged in learning activities. It is important for affective systems to predict certain emotions, particularly negative emotions, to be able to provide proper remediation and improve the learning experience of students. In this study, patterns of brainwave signals or electroencephalogram (EEG) of academic achiever high school students, such as those formed before the onset of an academic emotion, are analyzed by transforming this data into symbolic strings using a modified Shapelet Transform and SAX and from these strings, determining the emotional state of the students. These strings are shortened to further ease analysis. Results have shown that strings transformed from the same EEG feature and same emotion are significantly different to strings from different features and different emotions.

Keywords: Emotion recognition, Time series analysis, EEG, Affective learning environment, academic emotions, symbolic strings

1. Introduction

The emotions that a student experiences have an impact on the receptiveness of that student to the learning process. Students experiencing positive emotions would most likely have higher engagement and motivation to learn while students who are experiencing negative emotions would likely have lower engagement and motivation and be more susceptible to giving up on assigned tasks (Azcarraga, 2014). These emotions, which are involved in the entirety of the learning process, are referred to as academic emotions (Pekrun et al., 2002). There is an incentive then to identify and understand the emotions of students. This is to better understand the current disposition of the student and be able to adjust the learning program accordingly.

Identifying patterns of EEG readings that correspond to a certain academic emotion could prove useful in quickly identifying the emotions of a student by observing the readings produced. This could give the instructor an idea on the current experience of a student in real time. A student would typically experience a sequence of varying emotions throughout their learning process directly associated to which phase of learning they are in (Kort et al., 2001). Looking into the EEG readings as time series data would allow historical information of the emotions of the student to be considered as well in the classification process. This paper will look into the viability of transforming EEG readings into symbolic strings and from those strings attempt to extract patterns which may help monitor the emotional state of students while engaged in some learning activities.

2. Related Works

2.1 Alternative Representations of EEG data

Transformation of data into a form that can be analyzed more easily or to make it compatible with existing techniques is one of the primary steps of pre-processing. Different approaches of representing EEG and other time series data have been explored with one of these alternative approaches being symbolic representation, which transforms data into strings of symbols. Many researchers have found considerable success with using symbolic transformation for continuous data (Daw, Finney, & Tracy, 2003).

Use of symbolic representation has found relative success with using EEG data. One of these is a research by Muthu et al. (2019) which classifies eye movements through the similarity of strings transformed from windowed EEG data through Symbolic Aggregate Approximation (SAX). Another is a research by Raza and Kramer (2020) where SAX is once again used to transform the EEG data directly and from these strings, discriminating subsequences are extracted and used as features for transformation.

2.2 Emotion Recognition using EEG

Brainwave readings from EEG have been found to be an effective way to determine human emotions and has even been applied in the specific context of learning. The study of Heraz et al. (2007) attempted to train a learning agent which focused on predicting changes in what they call the emomental state of a learner, which they define as the combined mental and emotional state. The research done by Mampusti et al. (2011) attempted to classify a subject's emotions as they perform a learning task, Berg's card sorting task, through EEG data prepared using Butterworth bandpass filter. The research done by Azcarraga (2014) used EEG readings from students as they used computerized learning systems to classify the academic emotions of these students and to also visualize these emotions using structured self-organizing maps (SOM).

Expanding upon the existing literature and particularly the precursor study of this paper, a new method of transforming EEG data which can be used to identify and visualize academic emotions is explored. Although the findings of this research focus on the recognition of academic emotions of students, the techniques explored in this research could potentially br used in other applications which make use of EEG readings or even other types of time series data.

3. Methodology

3.1 Dataset Description

The original EEG dataset used in this study was collected from 52 participants, (only 8 of which were used in this study, 3 are female and 5 are male) consisting mostly of academically gifted science high school students who belong to the top 0.01% of the entire high school students in the Philippines. The brainwave signals of students are captured using the Emotiv EPOC sensor while they are answering mathematical problems using learning systems such as Scooter, a scatterplot tutoring system, and Aplusix, an algebra learning software.

The Emotiv EPOC headset sensor, which is a commercial product typically used for gaming purposes is equipped with 14 channels based on the International standard 10-20 locations. These readings are collected at around 2048 samples per second but the final dataset will contain 128 samples per second after being reduced by the device by performing a moving average.

The experiment involving one student typically lasts for an hour. The student is first briefed about the experiment and given a self-assessment then afterwards, the data collection equipment and the device to be used for the learning system are set up and worn by the student. The EEG of a few minutes of neutral activity, wherein the student simply stares at a blank screen with minimal activity, were collected first. The students were asked to report their level of confusion, boredom, frustration and interest every 2 mins or after each problem. These 4 emotions are some of the academic emotions in the study of Pekrun (2002).

For each of the 14 channels, the frequency bands of EEG signals are extracted using the selected data transformation technique. From the readings recorded from each of these channels, the frequencies and features are extracted. Only the frequency bands alpha, beta, and gamma are considered for this research, as guided by the findings established in Azcarraga (2014). The idea behind the use of these frequencies is that these frequencies are typically observed and associated to states of mind where emotions are expressed.

The class label of each observed instance is defined by the self-assessment of the student's affective state during the time frame that the observation was recorded. The labels consist of a rating of the level of intensity felt by the student for each academic emotion at a specific assessment period and an interpretation of this rating on whether the intensity is considered ”high” or ”low”. The interpretation is considered ”high” if it falls above a certain value and ”low” if it falls below

3.2 Pre-processing

Following what has been established in the basis study of this research (Azcarraga, 2014) and a similar study on EEG and emotion detection done by Dhanani et al. (2018), windowing of the data was guided by the findings established by Scherer and Ekman (1984) wherein emotions are found to usually last for 0.5 to 4 seconds. As such, in this study, the data is segmented into 2-second windows with 1-second overlap between windows.

The data was then transformed into a usable form with time bound information Discrete Wavelet Transform. This algorithm uses wavelets, which are defined us mathematical functions which analyze data as different frequency components with a resolution matching its scale (Graps, 1995). A key distinguishing feature of the wavelet transform is that it can give both the time and frequency information of the input signal. This is an important feature since for non-stationary signals, such as EEG, as certain frequencies do not occur all throughout the signal but rather, only at certain points in time. A transformation algorithm such as wavelet transform will be able to capture this information.

For the analyzing wavelet, Symlet was chosen since it was found to be suitable for decomposing EEG signals into frequency sub-bands. The Symlet wavelet tends to have a comparatively lower mean square error and thus is better for extracting various frequency bands from EEG signals (Shete & Shriram, 2014).

Transformation to the frequency domain by DWT was done for each 2-second window, and from this transformed data, the features that were extracted are peak magnitude and mean spectral power. The peak magnitude is the highest amplitude of the sample data while the mean spectral power is the average power spectrum of the signals in the sample data (Azcarraga, 2014). These two statistical features were computed for each frequency band and sensory location channel resulting in a total of 84 features.

Further pre-processing was done on the dataset after windowing and feature selection. EEG readings collected while the student is in a resting phase served as the baseline of the EEG readings of that student and were subtracted from the extracted features of the main dataset. Finally, the dataset was standardized and all values above 3 and below -3 were clipped to 3 and -3, respectively to manage the outliers.

3.3 Transformation to String Patterns

For the creation of string patterns, the standardized dataset is transformed using two data transformation algorithms: a modified Shapelet Transform and Symbolic Aggregate Approximation (SAX). These two techniques were used to transform windows of the EEG data into symbols which when taken as a dataset, form strings.

3.4 String Pattern Analysis

The resulting transformed strings are evaluated by how representative the strings are to the feature and emotion of the original data which was done by calculating similarity and difference among transformed strings. Difference to strings of different emotion and feature and similarity to strings of the same emotion and feature is indicative of the similar strings having a pattern among themselves. This difference and similarity was computed using Levenshtein edit distance where higher distance means a greater difference and lower distance, lesser difference. Then from the distance, the ratio is calculated where two strings of the exact match will give a ratio of 1.0 while two completely different strings will have a ratio 0.

4. Results and Analysis

A resulting string from the transformation represents only a single feature of one user but is equivalent to the entire data collection session. These strings are comprised of five different symbols, represented in this study by the first five letters of the Latin alphabet in uppercase form (A, B, C, D, and E). Using a different number of symbols retains the general information of the original data but increasing the number increases noises while decreasing the number results in lose of nuance.

The analysis of the transformed strings aim to verify if these strings are representative of the

emotion they are labeled as and distinguishable from strings transformed from other features. This would suggest that important information related to the feature was retained even after transformation. This was quantified by calculating the similarity of the transformed strings to other strings of the same emotion and feature using edit distance and statistical significance measures and then inversely, calculating the difference of the same strings to strings of the other emotion labels and features.

4.1 Shapelet Transform Strings

The transformation of Shapelets uses an implementation which follows the original Shapelet Transform paper by Lines et al. (2012). However, in addition. the five most distinguishing shapelets were first selected from all subsequences then for each of these subsequences, the distance to the five shapelets is calculated. The shapelet with the nearest distance is identified and the equivalent symbol of that shapelet is assigned to the subsequence. All subsequences are assigned a symbol which will comprise the final transformed string. The strings transformed in this manner can be seen in Table 1.

Table 1: *Randomly selected sample strings transformed using Shapelet Transform from all emotions.*

Feature	Emotion	Emotion Label	Transformed String
AF3 THETA AP	Bored	L	BEBBABBBBEEABBBABBBBEBBBBEABAB BEBBABBBBEBBBEEBBBBAEABBBB AEBBBBABABABBBBABBBBAA
		H	BBBBAEBABBBBEBBE
		L	BBBB
AF4_BETA_AP	Confused	H	DDDDEDDDDEDDE
		L	DDDDCDDDDBDDDD
		H	DEDDDDEDDDDE
F8 THETA PM	Frustrated	H	ECECCCCEEECCCCEECCEECCECCEECCEC CEEEECCEEEBCCECB
		L	CEEEBCECEECEECBEBECBECEEEBCCEEEE CCCEECCBCCCCBECCEECCECCCEEB
O1 BETA AP	Interested	H	BBEEEBBBBEBEBEBBEEEBEBEEEABBBBEE BBEBEDDAEEBBBBEEBBBEBBEBEBBBEEE E
		L	EEEEAEBEBBBBEDEBBBEEDBE
		H	BDEDDEEEEBBEEAB
		L	ABEBEEEC

4.2 SAX Strings

Among the symbolic representation techniques, SAX is among the more well-known techniques. This technique was used as a baseline of comparison to demonstrate the viability of the novel technique which used Shapelets. The implementation used for this research follows the original paper by J. Lin et al. (2007). The resulting strings that were produced using SAX have a similar visual appearance to the strings produced by the shapelet technique. These transformed strings can be seen below in Table 2.

Table 2: *Randomly selected sample strings transformed using SAX from all emotions.*

Feature	Emotion	Emotion Label	Transformed String
F4 THETA PM	Bored	H	DCDCCCCDDDDCCCDCDCCCDCCCCDCCCCDD CCCCDCDDCDDCDDCCCCD
		L	CCCCDCDDCCCCDCDCCCDDDCCCCDCCDCCCC DCCDCCCCDCCCDCCCCDCDCCCDCCC
		H	CCCC
F8 ALPHA PM	Confused	L	AAAACAAAA
		H	AAAA
		L	AAAAC

AF4 THETA AP	Frustrated	H	CCCCDCCCCDCCCCECC
		L	CCCCDCCCCDCCCCDCCCC
O2 BETA AP	Interested	L	DBBBBCCBBBDBBCCBBBBCBCCBBBB
		H	BBBBCBCCCBBBBCCCCBBBBCDCDDDDBDECCCDEDDB
		L	BBBCC

4.3 Overall comparison

The comparison of edit distances and ratios of the transformed strings can be seen in Table 3. It can be noted that there is a consistent trend of similar strings having higher ratios and different strings having lower ratios for both the shapelet and SAX algorithms. For the distances, there is a similar trend on the difference between the distance of similar strings and different strings: the difference is much higher for shapelet transformed strings compared to SAX transformed strings. Indeed, if measured with these metrics, it can be said that the proposed algorithm which uses shapelet transform can produce an output comparable to or of even better quality than SAX.

Table 3: *Average ratios and distances of the strings transformed using Shapelet compared to the strings transformed using SAX.*

		Compared to Same Feature, Same Emotion		Compared to Same Feature, Diff. Emotion		Compared to Diff. Feature, Same Emotion		Compared to Diff. Feature, Diff. Emotion	
		Shapelet	SAX	Shapelet	SAX	Shapelet	SAX	Shapelet	SAX
Average Ratio	Bored	0.4710	0.4541	0.4102	0.4046	0.2305	0.2706	0.2230	0.2693
	Confused	0.4921	0.4757	0.4222	0.3978	0.2480	0.2787	0.2414	0.2682
	Frustrated	0.4910	0.4772	0.3755	0.3893	0.2390	0.2905	0.2125	0.2750
	Interested	0.6058	0.4875	0.4800	0.4101	0.3636	0.2984	0.3081	0.2799
Average Distance	Bored	35.38	23.80	39.16	25.91	45.75	29.48	46.08	29.22
	Confused	37.63	24.49	43.23	28.08	50.58	31.94	52.12	32.61
	Frustrated	35.42	23.85	46.33	28.31	49.41	31.49	53.39	31.85
	Interested	21.07	20.32	29.30	24.32	30.33	26.27	34.97	27.31

To further support this claim, paired t-tests were conducted on the average ratio when compared to other strings of the same feature and emotion against ratios compared to strings of the same feature but different emotion and strings of different feature but the same emotion. The tests were setup wherein the null hypothesis is that ratios of strings of different feature or emotion have no difference or are greater than the ratios of strings of the same feature and emotion, while the alternative hypothesis is that the ratio of strings of different feature and emotion are less than ratios of strings of the same feature and emotion. These were done on a sample of 30 strings for each emotion and technique.

Table 4: *Resulting p-values of paired t-tests conducted on the average ratios of same emotion and feature against average ratios of other emotion and other feature*

		Against Other Emotion	Against Other Feature
SAX	Bored	0.014800075	1.92E-05
	Confused	4.95E-05	3.21E-08
	Frustrated	0.000407246	0.000150666
	Interested	9.87E-05	3.09E-07
Shapelets	Bored	0.009914672	1.32E-10
	Confused	0.000279542	2.11E-08
	Frustrated	0.000712499	9.66E-10
	Interested	7.32E-05	1.38E-10

From the results shown in Table 4, the alternative hypothesis can be accepted, with very low p-values across all emotions. For the purposes of this research, a confidence level of 0.95 and an alpha of 0.05 is acceptable but almost all the resulting p-values are less than even 0.01 which all the more supports the claim that there is significant difference between strings of the same feature and emotion to strings of a different feature and emotion.

5. Conclusions and Recommendations

In this paper, we explored using different techniques for representing EEG data for identification academic emotions. For transformation to the frequency domain, DWT was used and for transformation to string based representation, SAX and a modified Shapelet Transform was explored. By monitoring the generated emotion strings, affective systems may be able to recognize an academic emotion and provide proper remediation at the right time. The noticeable similarity of strings transformed from similar data shows that such information is retained when EEG data is transformed to strings. This also implies that strings of the same feature and emotions potentially have a common pattern among them. More advanced techniques to verify and analyze the resulting strings could be employed however to possibly provide more information and possibly also allow better recognition of emotions, such those used in natural language processing. Other techniques may be also explored to derive more readable versions of the strings such as Finite State Machines, Regular Expressions, or Hidden Markov Modelling generation which may also be used to extract more informative patterns from the strings.

References

Azcarraga, J. J. (2014). *Analysis and visualization of eeg data towards academic emotion recognition* (Unpublished doctoral dissertation). De La Salle University.

Daw, C. S., Finney, C. E. A., & Tracy, E. R. (2003). A review of symbolic analysis of experimental data. *Review of Scientific instruments*, *74* (2), 915–930.

Dhanani, D., Malonzo, J. L., Santos, K.-A. F. A., & Azcarraga, J. J. (2018). *Predicting stress and emotion levels based on eeg signals.*

Graps, A. (1995, Summer). An introduction to wavelets. *IEEE Computational Science and Engineering*, *2*(2), 50-61. doi: 10.1109/99.388960

Heraz, A., Razaki, R., & Frasson, C. (2007, July). Using machine learning to predict learner emotional state from brainwaves. In Seventh *ieee international conference on advanced learning technologies* (*icalt 2007*) (p. 853-857). doi: 10.1109/ICALT.2007.277

Kort, B., Reilly, R., & Picard, R. W. (2001). An affective model of interplay between emotions and learning: reengineering educational pedagogy-building a learning companion. In *Proceedings ieee international conference on advanced learning technologies* (p. 43-46).

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing sax: a novel symbolic representation of time series. Data Mining and knowledge discovery, 15(2), 107–144.

Lines, J., Davis, L. M., Hills, J., & Bagnall, A. (2012). A shapelet transform for time series classification. In *Proceedings of the 18th acm sigkdd international conference on knowledge discovery and data mining* (pp. 289–297). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/2339530.2339579 doi: 10.1145/2339530.2339579

Mampusti, E. T., Ng, J. S., Quinto, J. J. I., Teng, G. L., Suarez, M. T. C., & Trogo, R. S. (2011, Oct). Measuring academic affective states of students via brainwave signals. In *2011 third international conference on knowledge and systems engineering* (p. 226-231). doi: 10.1109/KSE.2011.43

Muthu, S. P. M., Lau, S. L., & Jou, C. (2019). Recognition of eye movements based on eeg signals and the sax algorithm. In *Intelligent and interactive computing* (pp. 237–247). Springer.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. (2002, 06). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist - EDUC PSYCHOL*, *37*, 91-105. doi: 10.1207/S15326985EP3702 4

Raza, A., & Kramer, S. (2020). Accelerating pattern-based time series classification: a linear time and space string mining approach. *Knowledge and Information Systems*, *62*(3), 1113–1141.

Scherer, K. R., & Ekman, P. (1984). *Approaches to emotion*. L. Erlbaum Asso- ciates.

Shete, S., & Shriram, R. (2014). Comparison of sub-band decomposition and reconstruction of eeg signal by daubechies9 and symlet9 wavelet. *In 2014 fourth international conference on communication systems and network technologies* (pp. 856–861).

Ye, L., & Keogh, E. (2011, January). Time series shapelets: A novel technique that allows accurate, interpretable and fast classification. *Data Min. Knowl. Discov.*, *22*(1-2), 149–182. Retrieved from http://dx.doi.org/10.1007/ s10618-010-0179-5 doi: 10.1007/s10618-010-0179-5

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Considerate, Unfair, or Just Fatigued? Examining Factors that Impact Teacher

Ashish GURUNG[a*], Anthony F. BOTELHO[b], Russell THOMPSON[a], Adam C. SALES[a], Sami BARAL[a] & Neil T. HEFFERNAN[a]
[a]*Worcester Polytechnic Institute, USA*
[b] *University of Florida, USA*
*agurung@wpi.edu

Abstract: It is particularly important to identify and address issues of fairness and equity in educational contexts as academic performance can have large impacts on the types of opportunities that are made available to students. While it is always the hope that educators approach student assessment with these issues in mind, there are a number of factors that likely impact how a teacher approaches the scoring of student work. Particularly in cases where the assessment of student work requires subjective judgment, as in the case of open-ended answers and essays, contextual information such as how the student has performed in the past, general perceptions of the student, and even other external factors such as fatigue may all influence how a teacher approaches assessment. While such factors exist, however, it is not always clear how these may introduce bias, nor is it clear whether such bias poses measurable risks to fairness and equity. In this paper, we examine these factors in the context of the assessment of student answers to open response questions from middle school mathematics learners. We observe how several factors such as context and fatigue correlate with teacher-assigned grades and discuss how learning systems may support fair assessment.

Keywords: halo effect, grading biases, fairness, subjective assessment

1. Introduction

In the context of education, a significant amount of research has been devoted to identifying, examining, and mitigating risks that particular policies, interventions, and instructional strategies may introduce as to the types of opportunities made available to students. Particularly with the introduction of computer-based learning platforms (CBLP), researchers are able to explore issues of fairness and bias through data-driven methods.

Traditional assessments are conducted in two formats: subjectively and objectively. The rise in the integration of technology in classrooms has facilitated the growth of CBLPs. Various CBLPs have been developed with the goal of alleviating difficult or tedious tasks faced by teachers. The most notable of the functionalities is the automation of objective assessments. Common in numerous contexts, the use of close-ended questions such as multiple choice and fill-in problems can be easily automated by computers where there are traditionally a small finite number of acceptable correct answers. The use of such questions has allowed developers to expand upon assessment processes to enrich the learning experience by offering additional feedback and on-demand help (Aleven et al., 2003; 2004; Ostrow et al.,2015; Patikorn et al., 2020). It is challenging to extend the same type of support to open-ended problems such as short answers and essay problems. While recent advancements in natural language processing (NLP) have made strides in automating assessment, this remains predominantly a manual task.

Writing is a critically important skill that helps teachers understand their students' thought processes and the ability to formulate arguments and justifications for their work (Biancarosa et al., 2004; Walton, 1992). In the domain of mathematics, on which the analyses of this work focus, teachers commonly use open-ended problems to gauge student knowledge when close-ended problems can often be solved by shallow learning and applying procedural rules (Livne et al., 2008; Silver, 1995). While subjective assessments are highly valuable, automating the process is not without risk; they are more dynamic compared to objective assessments. The dynamism is due to the variance in responses and the incongruity in teacher grades; the incongruity is caused by various intrinsic and extrinsic factors. Prior

work has found biases in grading behavior attributed to the "Halo Effect" (Cooper, 1981), characterized by a judgment made based on an attribute or characteristic of an individual; commonly, such attributes include gender (Spear, 1984; Roen, 1992), ethnicity (Wen, 1979; Fajardo, 1985), and name or surname (Lebuda et al., 2013). Furthermore, researchers exploring the Halo Effect found the initial favorable impression of students influenced their later evaluation (Malouff et al., 2013; 2014). While the use of rubrics or other standardized procedures provide a common set of metrics, the ultimate grade is typically based on the teachers' judgement. It is important to emphasize that measured biases do not necessarily equate to unfair assessment or evaluation as a teacher's knowledge of their students can also be very positive in terms of providing individual support through feedback and other communication (Hill et al., 2004; 2008; Jacob et al., 2017). In approaching assessment, teachers may consider a number of contextual factors when evaluating student work. In light of this, several questions emerge in terms of how assessment should be conducted to ensure fairness among students, particularly as researchers are moving to develop automated methods that attempt to mimic teacher grading practices.

Our goal in this work is to explore teacher assessment pertaining to open-ended questions. Using data collected from students working in a learning system in pre-COVID-19 classroom settings, we report on a pilot study to examine and explore teachers' approach to assessing open-ended work and how student identity may be considered in the grading process. We build on the study by conducting an exploratory analysis examining whether student-level attributes are predictive of teacher-provided grades when controlling for answer-level descriptors. Finally, we explore whether teacher grading fatigue poses risks to the fairness of student assessment.

In consideration of exploring factors that may affect fair student assessment, this paper addresses the following research questions:

1. Do teachers grade students differently when the students are anonymized?
2. When controlling for answer-level features, are factors of prior student performance and effort a reliable predictor of teacher-provided assessment scores?
3. Does the order in which teachers assess students appear to impact their grading?

To address these questions, we conduct 3 studies exploring various factors that are likely to affect or potentially bias teacher assessment. In the first study, we examine whether the anonymization of student identifiers affects how teachers score their own students. In the second study, we conduct a regression analysis to examine how measures of knowledge and effort correlate with teacher-provided assessment scores for student open-ended work. Finally, we explore the potential effects of grading fatigue on teacher-provided scores.

2. Background

Growth and innovation in Educational Technology (Ed-Tech) have broadly influenced the adaptation and regular usage of CBLPs in classrooms. Researchers and developers approached the design of learning platforms to consider students' various learning needs (Graham et al., 2013; Koedinger et al., 1997; Rafferty et al., 2016) to developing generic platforms (Arroyo et al., 2014; Burstein et al., 2013; Calvo et al., 2010; Heffernan et al., 2014). Systems developed to improve writing (Allen et al., 2016; Burstein et al., 2013; Calvo et al., 2010; Roscoe et al., 2019), mathematics (Arroyo et al., 2014; Heffernan et al., 2014), and programming (Price et al., 2017; Wiggins et al., 2015) are among the many examples of learning systems. Within these systems, a variety of features and supports are developed to support different aspects of learning, including the crowdsourcing of problems and solutions (Bhatnagar et al., 2016; Denny et al., 2008), and the availability of on-demand hints (Cambre et al., 2018; Williams et al., 2016).

Perhaps the most prominent feature of CBLPs is the ability to offer immediate feedback to students. Traditionally assessments are administered in objective and subjective forms. In many domains such as mathematics, students typically work through close-ended problems that can be assessed by simply comparing student answers with a finite set of acceptable correct responses (often with a simple "exact-match" approach), but across domains, the use of open-ended questions, allowing students to utilize language to explain their reasoning, have made it more difficult for CBLPs to immediately score. While examples of automated scoring tools exist (e.g. Baral et al., 2021; Erickson et al., 2020, Allen et al., 2016; Burstein et al., 2013), many CBLPs still rely on teachers/graders for manual assessment. A primary challenge with subjective assessment associated with this manual grading process is its susceptibility to bias.

As introduced in the previous section, the Halo Effect has been the focus of prior research within the context of subjective assessment of student work (Nisbett et al., 1977; Nieva et al., 1980). Prior research has found stereotyped biases interacting with gender (Spear, 1984; Wen, 1979 Martin, 1972; Roen, 1992), ethnicity (Piche et al., 1977, Wen, 1979; Fajardo, 1985), "likeability" and attractiveness (Landy et al., 1974, Cardy et al., 1986), student names (Lebuda et al., 2013), and perceived ability (Babad et al., 1975; Babad, 1980). Other studies exploring the Halo Effect found effects persisting across multiple assignments from the same student (Dennis, 2007), and that this effect was specifically identified in cases where teachers were assessing student writing samples (Forgas, 2011). One possible solution proposed to mitigate teacher bias is the anonymization of student identity during subjective grading (Malouff et al., 2013; 2014), motivating the current work.

2.1 Open-Ended Problems in ASSISTments

In this paper, we analyze the teachers grading open-ended mathematics problems in ASSISTments. ASSISTments is a CBLP that allows teachers to assign content (primarily in middle-school mathematics) and monitor student progress. The system provides students with immediate correctness feedback on close-ended problems and offers computer-provided help in the form of on-demand help and scaffolding. Open-ended problems in ASSISTments are available in two formats: (a) sub-problem of a multipart problem or (b) primary problem. Figure 1(a) shows an open-ended problem presented as a subpart of a multipart problem, and figure 1(b) shows an open-ended problem presented as the main problem. A multipart problem can have more than one open-ended sub-problem as well.

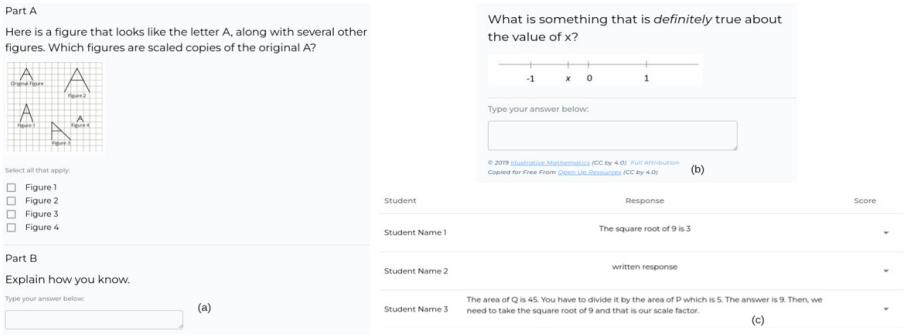


Figure 1. The different types of open-ended problems in ASSISTments.

Once students have completed their assignments, teachers can grade student responses. Figure 1 (c) shows the interface teachers can use to grade their students' responses. Teachers have the option to anonymize their students during the grading process where their identity is hidden, and the responses are shuffled, but the default behavior is to show identifying information of the students.

3. Study 1: Examining Grading Differences When the Student is Anonymized

Our first analysis explores whether teachers assess their students differently when they know the student's identity compared to when students are anonymous. In a purely unbiased, impartial scenario, teacher grading behavior would be solely dictated by the quality of the response as determined by how well the student was able to articulate their thoughts and demonstrate their knowledge of the concepts; this may include aspects such as grammar, use of mathematical terms (in mathematics contexts), and overall completeness. However, we posit there may be other factors that teachers consider including effort and prior performance. Even from a motivational perspective, a teacher may be inclined to bias their grade in a positive direction for students who normally under-perform but applied notable effort as a way to encourage similar behavior in the future (i.e. an arguably positive example of bias). The danger, however, is in whether these perceptions, implicit or explicitly applied during the grading process, impact the types of opportunities that may be available to a student.

We conducted a study involving 9 teachers who commonly assigned and graded open response problems in ASSISTments (14 teachers were part of a larger study, but only 9 participated in this portion). We selected 3 problems containing an open-response sub-part that was assigned within the month prior to the study. Each teacher had already assigned and graded their student work for at least

one of the three problems using the default scoring paradigm (i.e. teachers scored while knowing the identity of each student). In the month after this scoring was completed, we collected all student responses from across all teachers, anonymized them, and then randomly selected approximately 25 student responses to present to each teacher, ensuring that at least 10 of the responses were from each respective teacher's own students (if fewer than 10 were randomly selected, the difference was sampled from that teacher's student responses and added to the 25 given to the teacher); duplicate responses (e.g. empty responses or answers of "I don't know") were removed so that each teacher had a set of unique answers (resulting in some teachers having fewer than 10 responses from their own students if one was a duplicate). Each teacher was then asked to grade the given set of student answers, such that, for their own students, they would be anonymously grading the same set of answers as they had non-anonymously in the past (the additional responses and amount of time between grading reduced the likelihood that teachers would recognize their own students' responses). This presented the opportunity to measure each teacher's intra-rater reliability (i.e., how well they agreed with their past selves) and whether their grading was biased in a particular direction when they knew the student compared to not.

As the grading was done on a 5 points scale of 0-4; we applied Weighted Cohen's Kappa with linear weights to measure the variation in teachers grading behavior per response and found that the agreement coefficient as low as k=0.2293 and as high as k=0.7368, indicating that there was a large degree of disagreement between the two-time points, as shown in Table 1. These resulting scores were notably lower than we had initially hypothesized, suggesting that there were large differences in how teachers approached the grading of these when students were anonymized.

Table 1. *Exploring the grading behavior of teachers when they had access to students' identity vs. when students were anonymized using Linear Weighted Cohen's Kappa.*

Teacher	N	Intra-rater reliability (Weighted Cohen's Kappa)	Intra-rater reliability (Relaxed Cohen's Kappa)	Avg. grade diff (Initial - anonymized)
Teacher1	10	0.3878	0.8077	-0.2
Teacher2	10	0.7368	1	0.2
Teacher3	10	0.3939	0.6269	-0.2
Teacher4	10	0.4737	0.7368	0.3
Teacher5	11	0.2293	0.443	0.27
Teacher6	19	0.3596	0.5662	0.57
Teacher7	9	0.3571	0.3571	0.44
Teacher8	10	0.4531	0.5312	0.3
Teacher9	9	0.6539	0.761	-0.66

In addition to the Kappa measure, we also observed a relaxed calculation of Weighted Kappa. Given that grades are given on a 5-point scale and a teacher's assessment may reasonably vary by a small degree, we observe intra-rater agreement with an off-by-one adjustment (e.g. a scoring difference of one in either direction is treated as the same score when calculating Kappa). This adjustment resulted in notably higher Kappas, suggesting that the overall difference of scores was not as large as the first Kappa value suggested; while there was low precise agreement, teachers were relatively consistent within a grade point of themselves when the student was anonymized.

Pairing these Kappa scores with the average grade difference (the right-most column of Table1), we see that there is apparent bias in a particular direction. The positive difference exhibited by the majority of teachers suggests that teachers were more likely to grade anonymous students lower, on average, than when they knew the students. We conducted a permutation test across teachers to estimate the average difference in teacher grading behavior when students were anonymous versus not. We observed that teachers on average were more likely to give higher scores, 0.163 with 95% CI= [-0.1367, 0.4627] when students were not anonymized; while not statistically significant, likely due to our small sample size, this result is suggestive that there was some bias observed in the study.

We followed this empirical analysis with a set of semi-structured qualitative interviews first with the teachers as a group, and then with 2 teachers individually for an extended session to gain better context as to how they approach grading and why they believed their grades changed during the study. Overall, the teachers unanimously described considering contextual information of the student when

approaching the grading process. Several teachers mentioned that they consider motivational aspects (e.g., trying to encourage students to apply effort) when determining grades, as we had initially hypothesized. Several teachers also mentioned attempting to consider student effort when grading but did not use action-level reports in the system to do so (suggesting that it would be too time-consuming). The interviewed teachers similarly acknowledged potential risks to fairness highlighted by the observed differences, with one teacher expressing the intention to always grade anonymously following the study; others disagreed with this course of action citing perceived benefits of understanding the context of student work in order to provide better feedback to students in conjunction with the grade. While arguably anecdotal due to the limited sample size of teachers involved in this study, these conversations highlight the presence of bias in the form of a Halo Effect where teachers have an inherent motivation to give higher grades to perceived lower-performing students (again, not necessarily in a manner that affects fairness, but is still a form of bias).

4. Study 2: Exploring Related Factors of Student Assessment

Building on our findings from the pilot study we conduct a quantitative analysis investigating the role of student identity in the grading behavior of teachers. In this section, we attempt to explore the relationship between various answer- and student-level features and teacher-provided grades. Specifically, we seek to address the second and third research questions by exploring 1) whether prior student performance is a strong predictor of grade after accounting for concept-knowledge and other answer-level features (such as, for example, the number of words in the response), and 2) whether a measure of student effort correlates with the grades they ultimately receive while controlling for other measures of knowledge and ability. These analyses are meant to collectively provide insights into what a teacher may consider when assessing student work.

4.1 Description of the Dataset

We collected a dataset of authentic student responses to open-ended problems from the ASSISTments platform. The dataset consists of action logs of students interacting with open-ended problems for the academic years of 2018 through the beginning of 2020 (i.e., up to but excluding the period of remote learning in response to the COVID-19 pandemic). A notable portion of open-ended problems in the data are from open educational resources (OER).

The dataset contains action logs for open-ended problems assigned in the system. Overall, the dataset includes 344,847 action logs from 7,535 students working on 2,268 problems within 2,636 assignments. It is important to highlight that it is additionally the case that problems can contain multiple parts; particularly in the OER content, it is common for open-ended questions to exist as a sub-part of a multi-part problem (e.g., asking students to explain their reasoning after solving a closed-ended question of the same concept). Students worked on 3,404 distinct open-ended problems, reflecting that many problems contained open-response questions for multiple sub-parts. Grades for the student responses observed in this work followed a 5-point integer scale ranging from 0-4.

Following other work that observed response time as a measure of effort (see Gurung et al., 2021), we use student action logs to calculate how much time on task formulating their open response answers as a measure of effort. ASSISTments record three types of actions that are of interest to our analyses: starting a problem, leaving without answering to resume later, and submitting a response. We combine these actions into action pairs to compute the amount of time a student spent formulating their response to the problems, accounting for cases of students leaving and resuming work on the problem. These action pairs can be described using the notation of "(first action, second action)" where they represent two consecutive actions of a student taken within a session. The time for the action pairs represents the amount of time a student took between the two recorded actions. Our dataset observed two primary action pairs: "Problem Started-Submitted Response" and "Problem Resumed-Submitted Response" distinguished by the action observed prior to students submitting their response.

With this measure of student time-on-task, we apply a log transformation to create a pseudo-normal distribution and remove samples with a z-scores outside the range [-3, 3]; this filtering step attempts to remove very large outliers that may impact or bias our results. We also examined the open response problems and found that teachers graded only 19,446(~20%) of the 97,105 problems. The resulting number of action pairs used in our analyses for graded open responses are in table 2.

Table 2. *Filtered action pairs for students with open responses*

Action pairs	Graded responses	Ungraded Responses
(Problem Started, Submitted a response)	18295	73176
(Problem Resumed, Submitted a response)	1151	4483

There are several features of student answers that are likely to correlate with teacher-provided grades as identified in Baral et al. (2021). These features include the "Response Category," a categorical variable that indicates that the student response contains only words (positive class) as opposed to a mixture of linguistic and mathematical terms and expressions (negative class), as well as the "Number of Words," a simple count of the number of words (as denoted by spaces) in the student's answer.

In addition to the answer-level features, we calculate several student-level features to describe recent and historic performance measures. These measures include "Prior Percent Correct," the average correctness for all problems attempted by the student prior to beginning the open response problem (representing a long-term measure of general mathematics ability), and "Prior Sub-Part Performance," the average correctness across all prior sub-parts of the problem containing the open response question (representing content-specific knowledge).

Additional features were calculated including problem difficulty, the number of prior problem sub-parts, the number of prior problems started by the student, and the number of help requests made on earlier sub-parts, but these were found to either be highly correlated with other measures or not correlated with our outcomes of interest and were therefore omitted from our analyses. All pairwise Spearman correlations of features were calculated, omitting features introducing risks of collinearity.

The inclusion of both long- and short-term performance measures helps to distinguish and control for knowledge of the given mathematical concept as compared to general student ability. While initially hypothesized to be highly correlated, these measures ultimately exhibited a low correlation ($r = 0.014$) and could be used in our analyses without introducing risks of collinearity. Presumably, performance on the prior subparts should be a meaningful predictor of student performance on the open response problem given that they both pertain to the same mathematical concept. If while controlling for this it is found that the student's prior percent correct is similarly predictive, while not alone causal, such a result would suggest that a teacher may take prior student ability into account when grading the open response.

4.2 Factors Related to Student Grade

We use regression analysis to investigate the relationship between the described measure of effort (as measured by time-on-task), teacher-provided grades, and answer- and student-level features using a linear regression model (LM). We additionally include an interaction term of prior sub-part performance $\times$ the number of words.

From the regression results reported in Table 3, we found that the model ($R^2 = 0.133$) showed both prior sub-part performance and prior percent correct were reliable and meaningful predictors of student grade. This suggests that both student ability and content knowledge predict student grades. We also found that students with linguistic (word response category) responses correlate with lower scores as compared to responses that contained mathematical terms and expressions. The observed interaction term is found to be statistically reliable, but the low coefficient suggests that the relationship of this term is not very meaningful in comparison to the other more impactful features.

Table 3. *Linear Regression and Mixed-Effect model coefficients observing assessment score.*

-	**LM (R^2 = 0.133)**		**MLM**	
teacher			Variance 0.4523	Std.Dev. 0.6725
	coefficient	std. errs	coefficient	std. errs
intercept	0.2803	0.157	1.1648***	0.2226
response category (Words)	- 0.3927***	0.033	- 0.3924***	0.0322
prior sub-part perf.	1.3684***	0.088	1.3054***	0.0827

words per answer	0.0332***	0.006	0.0393***	0.0053
prior sub-part perf. × words per answer	- 0.0185*	0.006	- 0.0189**	0.0057
log transformed time on task	0.0662*	0.02	0.0323	0.0202
prior percent correct	1.275***	0.172	- 0.1497	0.2127

LM: Linear Model, MLM: Mixed Linear Model, Significance: *** < 0.000, **<0.001, * < 0.05

Additionally, we extended the LM by introducing the teacher as a random effect in a mixed-effect linear model (MLM), also reported in Table 3. The grading process, as reported by the teachers in our pilot study, accounts for students' perceived ability. In this model, the prior percent correct measure was no longer a significant predictor of student grade in the MLM suggesting that longer-term performance is not a prominent factor considered by teachers when assessing students. Prior sub-part performance, as a measure of concept knowledge, however, was still a reliable and meaningful predictor of student score.

5. Study 3: Potential Impacts of Fatigue on Grading

To address the final research question, we conduct one last analysis to observe how fatigue may affect teacher grades. In other words, we explore whether the ordering in which teachers grade student work leads to any potential risk of unfairness or bias due to implicit sequence or temporal effects. Particularly when a teacher has a large number of students to grade, it may be difficult to grade consistently for all students even when using a rubric. Particularly with the amount of time and attention needed to assess student open response answers, a teacher may find it difficult to give the same amount of attention to the 50th student as they do to the 1st student on a given assignment. Similarly, teachers may grade more strictly or leniently due to unconscious comparisons with previous students (e.g. a mid-grade student response may look better when assessed after a low-grade student response).

To explore this, we use an expanded dataset collected from 2018 to January 2020 to observe the mean and variance of grades over the course of observed grading sessions. This dataset contains 219,189 graded student open responses across 5562 problems from 3847 assignments.

Understanding that teachers may not grade all students for a given problem or assignment in one sitting, we find the order in which teachers graded student answers for each problem on a given day and plot the distribution of grades over this session ordering. Teachers who graded more than 50 students within a single span of time were omitted due to the sparsity of data. We visualize this trend in Figure 5 to observe whether fatigue appears to exhibit any temporal effects. If fatigue did affect how teachers grade we would expect to see trends that either affect the mean of teacher scores (rising or falling, on average, as teachers grade more students over time) or the standard deviation of scores.

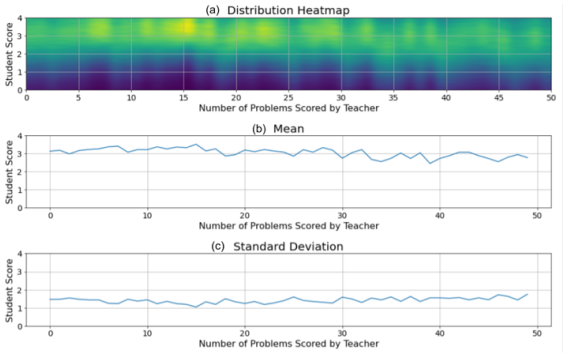


Figure 5. Visualizing teacher grading patterns as they grade sequences of student responses

From figure 5, there is little evidence that teachers' grading pattern changes significantly over time both in terms of mean as well as variance (i.e. the width of the distribution does not appear to change significantly, apart from the notable decrease in sample size as the number of problems scored extends above 25-30). While this does not speak to fatigue being an issue for individual teachers or scenarios, it does suggest that fatigue exhibits a low risk in terms of fairness and bias on average.

6. Discussion

Across the three studies presented in this work, we have attempted to gain a better understanding of how teachers approach the assessment of student open-ended work and identify factors that may impact given scores. As it pertains to addressing issues of fairness, the collective results of our analyses provide evidence that, somewhat unsurprisingly, teachers do consider contextual information beyond that pertinent to a given answer when assessing student work; this consideration does seem to bias grades in a positive direction, as suggested by the study observing anonymous grading. Qualitative data from our interviews support this, as well as the desire to utilize effort-based measures to even further inform student assessment practices.

From our regression analyses in Study 2, we were able to establish that teachers do not, on average, consider students' general mathematics ability as much as demonstrated knowledge on the given skill (i.e. in that the measure of concept knowledge was still found to be statistically significant after controlling for teacher effects, while the longer-term outcome was not. This is a little surprising particularly because several teachers from our first study described considering historic student performance when approaching scoring, largely from a motivational perspective for students who typically under-perform compared to their peers.

Finally, though we had initially hypothesized that grading fatigue may be a factor that affects how teachers grade, we found little evidence of this. We found that the mean and standard deviations of teacher scores remained relatively consistent while grading multiple students consecutively.

As we do identify some amount of bias in how teachers grade students, it is important to reiterate that this does not necessarily equate to teachers following unfair or inequitable assessment practices. At the current stage of this work, we are ill-equipped to say whether the measured bias is likely to contribute to impacts on the types of opportunities that students may receive; we can say from these studies that teachers, on average within the context observed in this paper, seem to be considerate of a range of factors that describe student performance, and are seemingly not largely impacted by obvious negative factors such as fatigue.

7. Limitation & Future work

As we identified certain biases exhibited in teacher assessment, it is important to explore whether these biases are more prevalent among certain student populations than others to identify deeper risks to student fairness. These populations may refer to protected demographic labels such as race, gender, ethnicity, or other geographic descriptors, but also latent groups of students. Similarly, while no meaningful effect of grading fatigue was observed across all teachers, this may still be an issue in more individualized cases.

As certain unmeasured recency or other timing effects may impact how teachers score (e.g. as the teachers re-scored student work from a month prior, and had undoubtedly moved on to new content areas), future work could also replicate this study to randomize when anonymization occurs within the study design.

Among the largest limitations of the current work is the correlational nature of the analyses presented. While we have attempted to pair some of our results with insights from experienced teachers, the causal mechanisms impacting student scores, particularly in reference to those factors external to the student's response itself, could be further explored through additional future studies. It is important to understand what is considered when assessing students so that teachers, education researchers, policymakers, and learning system developers can start to address the questions as to what should be considered and what factors may lead to unnecessary risks to fair student assessment.

8. Conclusion

While we are able to identify bias and student-level factors that seemingly correlate with student grades, it is uncertain as to whether such bias is truly negative in regard to fair student assessment. While we did not find any obvious risks to fairness within this presented set of analyses, this work represents a step toward identifying and mitigating such risks in educational assessment. This work further attempts to emphasize the distinction between bias and fairness, recognizing that the consideration of contextual information can have many positive benefits in terms of student learning, motivation, and engagement.

However, it is equally important to pursue further study of these issues to determine whether such benefits implicitly lead to inequitable opportunities.

Acknowledgements

We would like to thank the NSF (e.g., 2118725, 2118904, 1950683, 1917808, 1931523, 1940236, 1917713, 1903304, 1822830, 1759229, 1724889, 1636782, & 1535428), IES (e.g., R305N210049, R305D210031, R305A170137, R305A170243, R305A180401, & R305A120125), GAANN (e.g., P200A180088 & P200A150306), EIR (U411B190024 & S411B210024), ONR (N00014-18-1-2768), and Schmidt Futures.

References

Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2004). Toward tutoring help seeking. In Intl. Conf. on ITS Aleven, V., Stahl, E., Schworm, S., Fischer, F., & Wallace, R. (2003). Help seeking and help design in interactive

learning envs. Review of educational research, 73(3), 277-320.

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-Based Writing Instruction. Grantee Submission.

Arroyo, I., Woolf, B. P., Burelson, W., Muldner, K., Rai, D., & Tai, M. (2014). A multimedia adaptive tutoring system for mathematics that addresses cognition, metacognition and affect. Intl. Journal of AI in Education, 24(4), 387-426.

Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., & Heffernan, N. (2016, April). Axis: Generating explanations at scale with learnersourcing and machine learning. In Proceedings of the Third (2016) ACM Conference on Learning@ Scale (pp. 379-388).

Babad, E. Y. (1980). Expectancy bias in scoring as a function of ability and ethnic labels. Psychological Reports, 46(2), 625-626.

Babad, E. Y., Mann, M., & Mar-Hayim, M. (1975). Bias in scoring the WISC subtests. Journal of Consulting and Clinical Psychology, 43(2), 268.

Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). Improving Automated Scoring of Student Open Responses in Mathematics. Intl. EDM Society.

Bhatnagar, S., Lasry, N., Desmarais, M., & Charles, E. (2016, September). Dalite: Asynchronous peer instruction for moocs. In European Conf. on Technology Enhanced Learning (pp. 505-508). Springer, Cham.

Biancarosa, G., & Snow, C. E. (2004). Reading next: A vision for action and research in middle and high school literacy: A report from Carnegie Corporation of New York. Alliance for Excellent Edu.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated essay scoring system. Handbook of automated essay evaluation: Current applications and new directions, 55-67.

Calvo, R. A., O'Rourke, S. T., Jones, J., Yacef, K., & Reimann, P. (2010). Collaborative writing support tools on the cloud. IEEE Transactions on Learning Technologies, 4(1), 88-97.

Cambre, J., Klemmer, S., & Kulkarni, C. (2018, April). Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection. In Proceedings of the 2018 CHI Conf. on Human Factors in Computing Systems (pp. 1-13).

Cardy, R. L., & Dobbins, G. H. (1986). Affect and appraisal accuracy: Liking as an integral dimension in evaluating performance. Journal of applied psychology, 71(4), 672.

Cooper, W. H. (1981). Ubiquitous halo. Psychological bulletin, 90(2), 218.

Dennis, I. (2007). Halo effects in grading student projects. Journal of Applied Psychology, 92(4), 1169.

Denny, P., Hamer, J., Luxton-Reilly, A., & Purchase, H. (2008, September). PeerWise: students sharing their multiple-choice questions. In Proceedings of the fourth intl. workshop on computing education research (pp. 51-58).

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Douglas N Walton. 1992. Plausible argument in everyday conversation. SUNY Press.

Erickson, J. A., Botelho, A. F., McAteer, S., Varatharaj, A., & Heffernan, N. T. (2020, March). The automated grading of student open responses in mathematics. In Proceedings of the Tenth Intl. Conf. LAK

Fajardo, D. M. (1985). Author race, essay quality, and reverse discrimination. Journal of Applied Social Psychology, 15(3), 255-268.

Forgas, J. P. (2011). She just doesn't look like a philosopher…? Affective influences on the halo effect in impression formation. *European Journal of Social Psychology*, 41(7), 812-817

Graham S., MacArthur C. A., & Fitzgerald J. (Eds.). (2013). Best practices in writing instruction. Guilford press.

Gurung, A., Botelho, A. F., & Heffernan, N. T. (2021, April). Examining Student Effort on Help through Response Time Decomposition. In LAK21: 11th Intl. Learning Analytics and Knowledge Conf. (pp.292-301).

Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. Intl.Journal of AI in Education, 24(4), 470-497.
Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. Journal for research in mathematics education, 39(4), 372-400.
Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. The elementary school journal, 105(1), 11-30.
Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. Intl. Journal of AI in Education, 8(1), 30-43.
Landy, D., & Sigall, H. (1974). Beauty is talent: Task evaluation as a function of the performer's physical attractiveness. Journal of Personality and Social Psychology, 29(3), 299.
Lebuda, I., & Karwowski, M. (2013). Tell me your name and I'll tell you how creative your work is: Author's name and gender as factors influencing assessment of products' creativity in four different domains. Creativity Research Journal, 25(1), 137-142.
Livne, N. L., Livne, O. E., & Wight, C. A. (2008). Enhanching Mathematical Creativity through Multiple Solution to Open-Ended Problems Online.
Malouff, J. M., Stein, S. J., Bothma, L. N., Coulter, K., & Emmerton, A. J. (2014). Preventing halo bias in grading the work of university students. Cogent Psychology, 1(1), 988937.
Malouff, J. M., Emmerton, A. J., & Schutte, N. S. (2013). The risk of a halo bias as a reason to keep students anonymous during grading. Teaching of Psychology, 40(3), 233-237.
Martin, W. D. (1972). The sex factor in grading composition. Research in the Teaching of English, 6(1), 36-47. Nieva, V. F., & Gutek, B. A. (1980). Sex effects on evaluation. Academy of management Review, 5(2), 267-276.
Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. Journal of personality and social psychology, 35(4), 250.
Ostrow, K. S., & Heffernan, N. T. (2015, June). The role of student choice within adaptive tutoring. In Intl. Conf.on AI in Edu. Springer, 752–755.
Patikorn, T., & Heffernan, N. T. (2020, August). Effectiveness of crowd-sourcing on-demand assistance from teachers in online learning platforms. In Proceedings of the Seventh ACM conf. on Learning@ Scale.
Piché, G. L., Michlin, M., Rubin, D., & Sullivan, A. (1977). Effects of dialect-ethnicity, social class and quality of written compositions on teachers' subjective evaluations of children. Comms. Monographs, 44(1), 60-72.
Price, T., Zhi, R., & Barnes, T. (2017). Evaluation of a Data-Driven Feedback Algorithm for Open-Ended Programming. Intl. Educational Data Mining Society.
Rafferty, A. N., Brunskill, E., Griffiths, T. L., & Shafto, P. (2016). Faster teaching via pomdp planning. Cognitive science, 40(6), 1290-1332.
Roen, D. (1992). Gender and teacher response to student writing. In Gender issues in the teaching of English (pp. 126-141). Heinemann.
Roscoe, R. D., Allen, L. K., & McNamara, D. S. (2019). Contrasting writing practice formats in a writing strategy tutoring system. Journal of Educational Computing Research, 57(3), 723-754.
Silver, E. A. (1995). The Nature and Use of Open Problems in Mathematics Edu.: Mathematical and Pedagogical Perspectives. Zentralblatt fur Didaktik der Mathematik/Intl. Reviews on Mathematical Edu., 27(2), 67-72.
Spear, M. G. (1984). The biasing influence of pupil sex in a science marking exercise. Research in Science & Technological Edu., 2(1), 55-60.
Wen, S. S. (1979). Racial halo on evaluative rating: General or differential? Contemporary Educational Psychology, 4(1), 15-19.
Wiggins, J. B., Boyer, K. E., Baikadi, A., Ezen-Can, A., Grafsgaard, J. F., Ha, E. Y., ... & Wiebe, E. N. (2015, February). JavaTutor: an intelligent tutoring system that adapts to cognitive and affective states during computer programming. In Proceedings of the 46th acm technical symposium on computer science education (pp. 599-599).
Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., & Heffernan, N. (2016, April).

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Adaptive Scaffolding to Support Strategic Learning in an Open-Ended Learning Environment

Anabil Munshi*, Gautam BISWAS, Eduardo DAVALOS, Olivia LOGAN, Gayathri NARASIMHAM & Marian RUSHDY
Institute for Software Integrated Systems, Vanderbilt University, USA
*anabil.munshi@vanderbilt.edu

Abstract: An important goal for intelligent learning environments is to provide adaptive support for learners. This paper presents an adaptive scaffolding framework designed to support strategic learning in Betty's Brain, an open-ended learning-by-teaching environment. The framework includes an online strategy detector that identifies students' sub-optimal use of cognitive-metacognitive strategies, and a conversational adaptive feedback mechanism to help students overcome their difficulties. We ran a pilot study with undergraduate students to determine how well they applied the adaptive scaffolds provided by our framework. Our qualitative case study analysis used students' activity logs, one-on-one interactions with researchers, and eye-gaze patterns on the Betty's Brain screens, to infer their learning behaviors and performance. Findings suggest that some scaffolds helped students develop more effective and strategic behaviors to overcome their difficulties. The results also suggest a roadmap to improve scaffold design and better support long-term learning strategies in Betty's Brain.

Keywords: Adaptive scaffolding, learning strategies, open-ended learning environments

1. Introduction

External scaffolding adapted to students' learning difficulties can help them achieve their learning goals while working on open-ended learning tasks, such as constructing models of science phenomena (Basu et al., 2017). Adaptive scaffolding is especially needed in open-ended learning environments or OELEs (Land, 2000), where students can develop and refine their problem-solving approaches. In this paper, we build and evaluate an adaptive scaffolding framework in Betty's Brain (Leelawong & Biswas, 2008), an OELE where students learn about scientific processes by modeling them as cause-and-effect relations to teach a virtual agent Betty.

The goals of this paper are to present our adaptive scaffolding framework and study its usefulness. Learners who apply effective learning strategies to build, monitor, debug and refine their models achieve higher success in their learning outcomes in Betty's Brain (Munshi et al., 2018b; Segedy et al., 2013). We have previously designed scaffolds in the Betty's Brain environment that tracked 'inflection points' in students' behaviors during learning, and offered timely feedback to help them regulate their behaviors (Munshi et al., 2022). While evaluating these scaffolds, we noted that high and low performing students reacted differently to the scaffolds they received. Low performers often lacked the ability to interpret the feedback, causing them to continue to exhibit ineffective behaviors and fail to successfully complete their tasks. This suggests a need to develop more systematic strategy-oriented feedback that students can interpret and adopt in the context of their learning tasks. We use such findings from our prior work to inform the scaffolding approach presented in this paper.

The rest of this paper is organized as follows. Section 2 provides a background for our scaffolding design framework. Section 3 presents an overview of the Betty's Brain environment. Section 4 discusses the design and implementation of our adaptive scaffolding framework in Betty's Brain. Section 5 outlines the pilot study and data collection scheme (involving students' activity logs, in-the-moment interviews, and eye-tracking data) to evaluate the scaffolding framework. Section 6 reports the findings from data analysis, and the lessons learned, that provide an opportunity to further improve the feedback design to support the needs of a more novice middle school population in our future studies. Section 7 presents our conclusions and directions for future work.

2. A Framework for Adaptive Scaffolding

Scientific model-building is a multi-faceted process achieved by developing, debugging, and refining the model in stages. Typically, K-12 students have little experience with such cognitively demanding practices, and have difficulties engaging in the model-building process (Gilbert et al., 2016). To address such challenges, we develop an adaptive scaffolding framework to support learners as they build causal models of a scientific process in Betty's Brain. *Scaffolding* describes pedagogical support calibrated to the learner's current level of understanding. It helps them accomplish tasks they could not accomplish alone (Wood et al., 1976). This difference between what a learner can accomplish by themselves versus with external assistance is characterized by the *zone of proximal development* (ZPD) (Vygotsky et al., 1978), that informs how students' range of understanding and problem-solving can be extended in a way that provides sufficient challenge but prevents frustration or boredom. Puntambekar & Hubscher (2005) note some key features of scaffolding: a shared understanding of the activity and goals; ongoing diagnosis of the task, subtasks, and the student's current level of understanding; tailored assistance. The goal for scaffolding is to improve students' self-directed learning skills and strategies (Kim et al., 2018).

The challenges for adaptive scaffolding are to ensure that the intervention does not happen too soon or too often, that students do not become too reliant on feedback for making progress, and to avoid feedback that is perceived negatively and evokes frustration. Identifying individual students' ZPD thus becomes an important aspect to trigger adaptive scaffolds. In this paper, we seek to adhere to these principles in constructing the adaptive scaffolds. Our scaffolds are real-time conversational feedback delivered by virtual agents within Betty's Brain. They seek to directly influence students' cognitive strategies, thereby also impacting their metacognitive processes. The feedback is provided when students face difficulties in their model-building, debugging, and problem-solving tasks while learning.

3. The Betty's Brain Learning Environment

Betty's Brain (Leelawong & Biswas, 2008) is an OELE that uses learning-by-teaching to help K-12 students learn about scientific processes like climate change. Students teach a virtual agent Betty about *causal links* (cause-and-effect relations) between concepts, to build a map of the scientific process.

The learning environment provides students with a set of tools to learn and teach Betty. The ***science book***, a set of resource pages, helps students access information about cause-and-effect relations in the science domain (e.g., climate change) to teach Betty a causal model of a scientific process (e.g., the greenhouse effect). The ***causal map*** provides a drag-and-drop interface to build causal models (by adding, deleting, or modifying concepts and causal links). The **quiz** tool allows students to probe Betty's understanding by asking her to take a quiz, where the quiz results help them to evaluate the correctness of the current causal map. A mentor agent Mr. Davis scores these quizzes. *Overall*, these tools provide students with the resources needed to complete their learning goals. But novice learners, who fail to invoke cognitive and metacognitive strategies to aid their model-building and debugging processes, often find it difficult to use these tools effectively. Adaptive scaffolding, provided at opportune moments during learning, can help these students become better learners as they teach Betty.

4. Designing Adaptive Scaffolds in Betty's Brain

Triggering scaffolds in students' ZPD is important to support strategic learning. Weinstein et al. (2000) define *strategic learning* in terms of students' knowledge of learning strategies, ability to apply strategies, and their self-regulation, including cognitive and metacognitive processes. Applying the ZPD framework to support strategic learning requires first identifying *students' current abilities* in terms of their cognitive and metacognitive processes and their *difficulties* in the *context of their current tasks*. Identifying the trigger conditions and the context for scaffolding is especially challenging in OELEs like Betty's Brain where students can pursue their own strategies. This requires a good understanding of how students apply cognitive and metacognitive strategies during learning, and how to interpret their activities in the context of their current task. Selecting meaningful scaffolds and providing them 'as-needed', to help students engage in more effective behaviors, is also a challenge.

The Betty's Brain task model (Kinnebrew et al., 2017) provides a framework for identifying students' cognitive and metacognitive behaviors from their activity patterns, by mapping tasks and activities (viz., **Read**ing science book pages; **Build**ing causal models by adding, deleting, or modifying links; checking the models by taking a **Quiz**) to higher-level cognitive processes like information

acquisition (**IA**), solution construction (**SC**), and solution assessment (**SA**). Cognitive strategies relate to IA, SC, and SA activities. Metacognitive strategies result from meaningful combinations of the IA, SC, and SA processes. When combined sequentially, these actions can be expressed as binary relations (Basu et al., 2017) and further interpreted as "effective" behaviors (e.g., *Read→Add Correct Link*). Prior Betty's Brain research has used sequential pattern mining on activity sequence data to detect behavioral patterns that suggest the use of productive or unproductive strategies (Munshi et al., 2018b). In this paper, we track such strategies *online* to trigger appropriate scaffolds.

Figure 1 presents the implementation of our scaffolding framework in Betty's Brain. Students' mouse clicks, interpreted as *events* by the system, are recorded as their current *action* (e.g., adding a *"A increases B"* link) and the active *view,* or context, of the current action (e.g., the student is engaged in the Build view). The 'learner model controller' uses this information to construct sequential relations, or behaviors. To ensure that we scaffold *only* when a student has difficulties, our framework specifically tracks behaviors that suggest ineffective strategy use. A *pattern counter* ensures that a scaffold is triggered only when an ineffective strategy is used multiple times, thereby giving the student some agency the first time to monitor their situation and refine their strategy without external intervention. An ineffective strategy used more than a set number of times (say, 2) is considered an active 'pattern' and placed on a *priority queue* (Figure 1). The scaffolds, triggered by the highest-priority active pattern at a time, are delivered by the mentor agent Mr. Davis or Betty. The agent keeps track of the time that has passed since the student was last scaffolded, to avoid intervening too often. If this time exceeds a threshold (say, 5 minutes), the agent fetches the current highest-priority pattern, which becomes the **triggering condition** to find relevant scaffolding. Each trigger condition is mapped to a **conversation tree** (Segedy et al., 2013) that allows feedback delivery in a back-and-forth conversational manner tailored to address students' gaps in strategy use at that trigger condition.

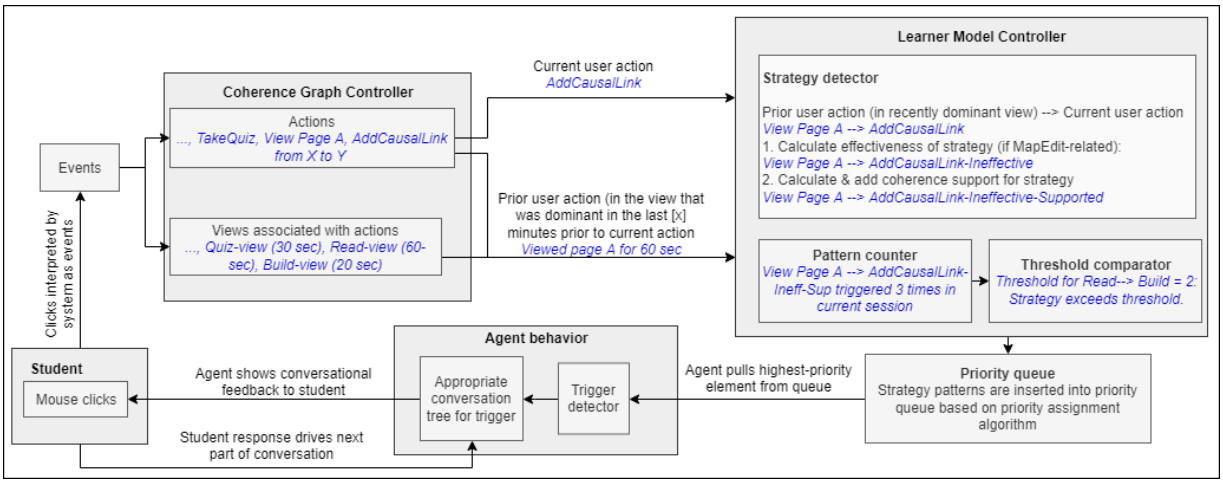


Figure 1. Implementing the adaptive scaffolding framework in Betty's Brain *(examples in blue)*

A "scaffold level" parameter further determines if a student receives *general feedback* (level 0) or more *contextual feedback* (level 1). A student is initially provided a Level-0 scaffold but if the same type of scaffold is triggered again, suggesting that the student has been unsuccessful in developing the associated strategy so far, they would now receive a Level-1 (more direct) scaffold. In the current design iteration (Table 1), we did not assign multiple levels to some scaffolds, because their trigger conditions required contextual feedback directly at Level-0 for the scaffold to be useful and actionable. Table 1 lists the seven types of scaffolds included in the current iteration of our scaffolding framework. In the first column, the three broad types of behavior patterns appear in **bold**: *Read→Build* (IA→SC), *Quiz →Build* (SA→SC) and *Quiz→Read* (SA→IA), while the additional task contexts tracked by the pattern detection framework for triggering scaffolds are indicated in *italics*. In the second column, the levels of specificity of the scaffold, where applicable, are indicated in **bold**.

5. Methods

To evaluate the usefulness of scaffolds, we conducted a pilot study with six undergraduate students (median age: 20 years; 2 males, 4 females) at Vanderbilt University, USA. We ran this study with undergraduates to

see how this more mature group interpreted and used the feedback, so that we may derive actionable insights to refine the feedback before testing it out with middle schoolers. Since the objective of the pilot study was to observe each student's learning process, interact one-on-one, and perform a qualitative analysis of feedback usefulness, the low number of participants was sufficient.

At the beginning of the study, participants were provided an overview of Betty's Brain. Then they worked on an introductory unit for 15 minutes, to get hands-on experience using the learning tools. They then switched to the "climate change" unit and worked on building their causal models for 2 hours. An experimenter observed each student and interacted with them when needed.

Table 1. *Types of strategic adaptive scaffolds in current iteration of scaffolding framework*

Scaffold Type	Excerpt from Conversation Tree
Read→Build Correctness: Shortcut Link **(the Build is a shortcut i.e. A→C link, instead of A→B→C)**	**Level 0:** Mr. Davis: "... Your map may have shortcut link(s) from … part of the science book. Do you want me to tell you more about shortcut links? ..."
	Level 1: Mr. Davis: "… you still have a shortcut link coming out of [Concept A]. ... Review pages [X, Y] to find out more about the missing link(s) ...."
Read→Build Correctness: Incorrect Link **(Build incorrect but not shortcut)**	**Level 0:** Mr. Davis: "You may have incorrect links ... There are three ways a link may be incorrect: … (a) linking two unrelated concepts; (b) sign (increase/decrease) is wrong; (c) direction is wrong. Review Pages [X, Y] ..."
	Level 1: Mr. Davis: "… incorrect link out of [Concept A]. Review pages …"
Read→Build Coherence **(Build is not coherent to prior Read)**	**Level 0:** Betty: "This … is not related to what we read." Mr. Davis: "... A good strategy is to work on one topic at a time … helps .."
	Level 1: Mr. Davis: "... you just added a link [A→B] from Page [Y] but you were reading Page [X]. Try to add all links from Page [X] first …"
Quiz→Build Correct Link Annotation **(unmarked correct links)**	Mr. Davis: "... There are correct links .. that you have not marked on your map yet. ... Select a quiz question graded 'correct' (green checkmark). …"
Quiz→Build Incorrect Link Annotation **(unmarked incorrect links)**	Mr. Davis: "... Select a quiz question graded incorrect (red X). … This means that at least one of the links… is wrong. Mark these as 'maybe wrong'..."
Quiz→Build Coherence **(when Build is not coherent to last Quiz)**	Mr. Davis: "... After Betty takes a quiz, … first teach her the concepts she did not answer correctly. This way, she can get these right the next time …"
Quiz→Read Coherence **(when Read is not coherent to last Quiz)**	Mr. Davis: "... I can help you find information from the science book to correct Betty's wrong answers. Type … in the *search bar* of the page …"

We collected the following data: **(a) Activity logs:** Students' system interactions logged as activities with timestamps. **(b) Screen recordings:** System interface visible on laptop screens recorded as video files using OBS Studio. **(c) Interviews:** An experimenter interviewed each student one-on-one at different points while learning, especially after they received scaffolds, to know their thoughts on the feedback and how they planned to use it. Interviews were recorded using OBS, transcribed using Otter.ai software, and manually checked for transcription errors. **(d) Eye-gaze logs:** Tobii 5C eye-trackers recorded students' eye-gaze movements which were superimposed on OBS screen videos.

6. Findings

6.1 Count of Scaffolds Received by Participants

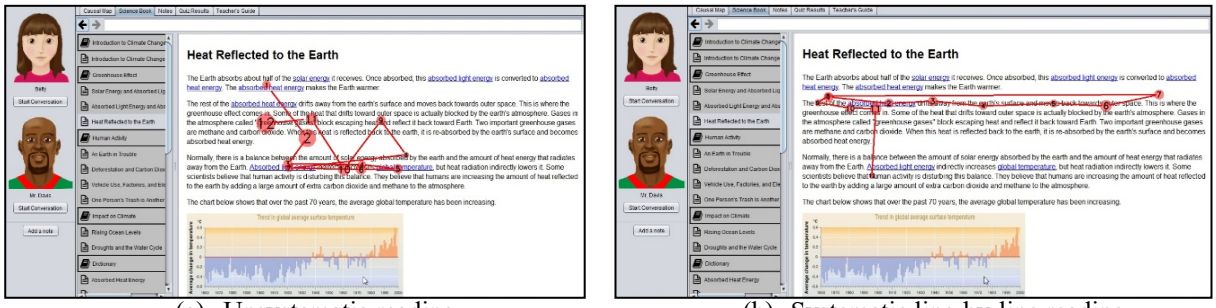
Table 2. *Count of different types of adaptive scaffolds received by study participants*

Participant ID	Count by Scaffold Type				Total Count
	Read→Build Shortcut Link	Read→Build Incorrect Link	Read→Build Coherence	Quiz→Build Correct Link Annotation	
P01	**2**	**4**	**1**	-	**7**
P02	-	**2**	-	-	**2**
P03	**1**	**2**	-	**1**	**4**
P04	**3**	**2**	-	**4**	**9**
P05	-	**2**	-	-	**2**
P06	**2**	**2**	**2**	**4**	**10**
Total (n=6)	**8**	**14**	**3**	**9**	**34**

From Table 2, 34 scaffolds were triggered across the 6 participants. Only 4 of the 7 adaptive scaffolds from Table 1were triggered during the study and are included in Table 2. Of these, *Read→Build: Incorrect Link* was triggered most frequently, and received at least twice by every student. The dominance of this scaffold across students suggests that they frequently attempted a *Read→Build* strategy to teach Betty but were often unsuccessful with applying it on their maps.

6.2 Case Study Results

We study usefulness of the scaffolds received by students in Table 2 by doing a qualitative analysis that combines information from the different data streams (Section 5) to assess how each type of adaptive feedback affected students' subsequent strategic learning behaviors and performance. Due to space constraints, we discuss the results for two randomly selected representative students - P03 and P06.



(a) Unsystematic reading (b) Systematic line-by-line reading

Figure 2. Contrasting gaze patterns of student P03 while reading a Science Book page

Read→Build Shortcut Link Feedback: This scaffold was triggered by a *Read→AddShortcutLink* pattern where, after reading, the student added "shortcut links", i.e., a direct link between two concepts, e.g., A→C, instead of building the more complete relations A→B→C. The scaffold provided pointers on how to read the Science Book more strategically so as not to miss important causal chains. Student P03 received this feedback once. She followed Mr. Davis' suggestion and started reading the Science Book. Eye-gaze movements from this period (Figure 2) show that her gaze initially moved between unrelated sections of the page (Figure 2a), suggesting an inability to find the information needed to fix the shortcut link. Then she studied the links on her map more closely and came back to reading. Then she started more systematic line-by-line reading (Figure 2b), that was followed by fixing the shortcut link. This suggests that the feedback helped the student develop a better map-debugging strategy. Student P06 received this feedback twice. After the first feedback at t=13 min, she corrected the shortcut links, although she did not review the Science Book as suggested. At t=22 min, P06 received the second feedback. Eye-gaze fixations show that she ignored a part of the feedback that suggested her to "read Science Book pages". After the scaffold, P06 did not read again, instead adding links she assumed would fix the shortcut link. But she was unable to fix the errors. We notice a reliance on intuition versus carefully reading the feedback here. *Overall*, the *Shortcut Link Feedback* was useful as a map-debugging-by-reading strategy at times when students read and followed the feedback carefully.

Read→Build Incorrect Link Feedback: This scaffold was triggered by a *Read→IncorrectLinkEdit* pattern. Mr. Davis explained *why* links may be incorrect, and pointed to certain pages to identify incorrect links. Student P03 received this feedback 2 times. At t=36 min, after a Level-0 one, she fixed one of two incorrect links. In an interview, she surmised that she only fixed her last link since Mr. Davis showed up right after she had added it, making her guess it could be wrong. After getting the more direct Level-1 feedback, P03 fixed her other errors. But she continued to add incorrect links, suggesting that she had not internalized the suggested strategy. Student P06 also received this feedback 2 times. After the first feedback at t=29 min she went to the suggested Science Book page but could not extract the correct causal information, so her map errors remained. At t=38 min, P06 received a Level 1 feedback that she quickly scanned (gaze movements) and treated more as a 'corrective' hint versus a 'strategic' one. Instead of the suggested reading, she deleted a link which did not fully solve the issue. *Overall*, this feedback had partial success in helping students address their model-building errors. We need to make the feedback logic more transparent so that students avoid acting on their assumptions of the feedback.

Read→Build Coherence Feedback: This scaffold was triggered by a *Read→IncoherentLinkEdit* pattern, i.e., adding links unrelated to recent reading. Mr. Davis suggested a coherent *Read→Build* strategy to better organize the causal model and prevent potential errors. This feedback was not triggered for P03, suggesting that she showed coherent *Read→Build* behaviors. Student P06 received the feedback twice. Eye-gaze movements after the first feedback show the student pay more attention to a piece of text in the Science Book, which she then used to refine her understanding and correct map errors. At t=33 min, P06 received the next feedback but did not follow the suggestion this time.

Quiz→Build Correct Link Annotation: This feedback was triggered by a *Quiz→Build* pattern, where a student did not annotate ("mark as correct") the links proven correct in a quiz. Mr. Davis suggested that the student use the 'Checkmark' feature to annotate the link(s), explaining how this could help spot incorrect links. Student P03 received this feedback once. After the feedback, she looked at the quiz questions graded as correct, used the Quiz Explanations feature to look at associated links (verified from gaze movements) and marked the unmarked correct links on her map. Student P06 received this 4 times. Gaze movements from the first time suggest that P06 followed Mr. Davis' suggestion by marking correct links. She continued marking links associated with other correct quiz answers, suggesting that she found the strategy useful and internalized it. However, from an interview, the feedback was not as useful to her later, because she was already aware of this strategy by that time. This shows a need to better monitor students' behaviors and look for evidence of strategy internalization. *Overall*, this feedback was followed by students and helped them to adopt the intended strategy in their work.

7. Conclusions

This paper discusses our strategy-oriented adaptive scaffolding framework in Betty's Brain. Scaffolds were triggered when students displayed continued ineffective use of strategies as they worked on their model-building tasks. Four out of seven scaffolds were triggered in the pilot study, but that may be attributed to the more mature undergraduate participants. The combination of log traces, screen recordings, interview responses, and eye-tracking data for qualitative case studies allowed us to make deep inferences about student behaviors and scaffold usefulness. Findings from Section 6 show that some scaffolds were useful while others require further contextualization and elaboration. As next steps, we want to use these results to refine our scaffolds. For instance, we will revise the agent conversations to make the purpose of the strategies clearer to students and help them understand the contexts in which they may apply. Such additional clarification is especially important to support more novice middle schoolers, who often find it hard to understand and apply productive learning strategies (Munshi et al., 2022). We would then like to a quantitative study in a middle school classroom for scaffold evaluation. We will extend our framework to explicitly study relations between students' cognition and affect in self-regulation (Munshi et al., 2018a), and how these influence their strategies in Betty's Brain. Our goal will be to combine cognitive-metacognitive strategy triggers with affective triggers, to generate more engaging and productive scaffolds for learning.

References

Basu,S., Biswas,G. and Kinnebrew,J.S. (2017) Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Modeling and User-Adapted Interaction*, 27, 5–53.

Gilbert, J. K., & Justi, R. (2016). Models of modelling. *Modelling-based teaching in science education,* pp. 17-40.

Kim, N. J., Belland, B. R., & Axelrod, D. (2018). Scaffolding for optimal challenge in K–12 problem-based learning. *The Interdisciplinary Journal of Problem-based Learning*, 13(1), Article 3.

Kinnebrew, J., Segedy, J., and Biswas, G. (2017). Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. *IEEE TLT*., 10(2):140–153.

Land, S. (2000). Cognitive requirements for learning with open-ended learning environments. *Educational Technol Research and Development*, 48, 61-78.

Leelawong,K. and Biswas,G. (2008) Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, 18(3), 181–208.

Munshi, A., Rajendran, R., Ocumpaugh, J., Biswas, G., Baker, R.S., & Paquette, L. (2018a). Modeling Learners' Cognitive and Affective States to Scaffold SRL in Open-Ended Learning Environments. ACM UMAP 2018.

Munshi, A., Rajendran, R., & Biswas, G. (2018b). Studying the Interactions between Components of Self-Regulated Learning in Open Ended Learning Environments. ICLS Vol 3, pp 1691-1692. London.

Munshi, A., Biswas, G. (2022) Analyzing Adaptive Scaffolds that Help Students Develop Self-Regulated Learning Behaviors. In review. *Preprint on arXiv*.

Puntambekar, S. and Hubscher, R. (2005). Tools for scaffolding students in a complex learning environment: what have we gained and what have we missed? *Educational Psychology*, 40(1), 1–12.

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes. Harvard U Press.

Weinstein, C., Husman, J., & Dierking, R. (2000). Self-regulation interventions with a focus on learning strategies. In M. Boekarts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 728–744).

Wood,D., Bruner,J. and Ross,G. (1976) The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 17, 89–100.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Unveiling the Process of Collaborative Learning Through the Use of Digital Whiteboard Historical Action Logs

Hua Leong FWA
Singapore Management University
*hlfwa@smu.edu.sg

Abstract: The Covid-19 pandemic has forced many educational institutions to transit from face-to-face to online delivery of lessons. Precluded from meeting face-to-face in a classroom due to pandemic curbs, a digital whiteboard affords both teacher and learners to be collaborating and completing a task from home instead. In this paper, we first used a survey to assess learner's acceptance and use of digital whiteboard. We next web-scraped and visualized historical action logs recorded by the digital whiteboarding tool to uncover the process of problem-solving between and within the teams. This is significant as social-loafing and free-riding are typically difficult to detect and prevent in collaborative learning. In addition, instructors traditionally evaluate only the product but not the process of collaborative learning as most collaborative work of learners occur outside scheduled teaching hours. From the results of the survey, we can conclude that the digital whiteboard is accepted by the learners. It also encourages collaboration and leads to higher learning motivation of learners. From the visualizations of the historical action logs, we were able to discover differing interaction and collaboration practices among the teams as well as the contributions of team members within each team (e.g. some teams exhibiting uneven effort distribution). We believe that our findings are significant as to the best of our knowledge, this is the first study that uses historical action logs from digital whiteboard to gain insights into the process and evidence of collaborative learning through visualization techniques.

Keywords: digital whiteboard, collaborative learning, visualization, historical action logs

1. Introduction

Collaboration is an essential skill for the 21st century workforce. The affordance of modern technology has enabled businesses to build global teams possibly located in different geographical regions and time zones. It is thus imperative that modern and future workers learn to collaborate in distributed and diverse settings for the attainment of common business goals. In the modern classroom, collaborative learning, an educational approach which involves learners working together in small teams towards a common goal, is posited to lead to enhanced learning outcomes as compared to didactic instruction. In contrast to didactic instruction where learners receive knowledge passively from the teacher, collaborative learning is an active learning approach where learners participate actively with each other and co-construct their knowledge in the process. Collaborative learning (Smith & MacGregor, 1992), conditioned on enhanced interactivity, individual accountability and teamwork is underpinned by the theories of constructivism (Piaget, 1952) and the sociocultural theory of learning (Vygotsky, 1980). It is also recognized that adoption of collaborative learning fosters social interaction (Tinto, 1997), develops high order thinking skills (Webb, 1982), embraces diversity (Loes et al., 2018) and leads to higher motivational levels of learning (Astin, 1977).

Traditionally, before the proliferation of technological tools into classrooms, teachers rely on blackboard, whiteboard or even pen and paper for groups of learners to work on collaborative tasks. The use of whiteboards or blackboards, however, constrain a single team of learner to be working on the board at any one time. A team of learners huddling at the board in turn occluded other teams from observing what the team is doing. In addition, the short-lived nature of the content on the board poses further obstacles to learning (Price et al., 2008). The advent of interactive whiteboard (IWB) eases the collaboration effort (Armstrong et al., 2005) by allowing other teams to be viewing and even

contributing by annotating using their tablets or laptops while the main team is working on the main interactive whiteboard.

In recent years, development of digital whiteboards eliminated the need to invest in a physical interactive whiteboard hardware. A digital whiteboard offers almost equivalent features as a physical interactive whiteboard with a pure software offering. The Covid-19 pandemic has forced many educational institutions to transit from face-to-face to online delivery of lessons. Precluded from meeting face-to-face in a classroom due to pandemic curbs, a digital whiteboard affords both teacher and learners to be collaborating and completing a task from home instead.

Some of the challenges associated with collaborative learning relates to that of individual effort attribution. Social loafing (reduction of personal effort when working in a group) (Harkins & Jackson, 1985) and free-riding (individual bearing minimum work but sharing benefits of the group) (Albanese & Van Fleet, 1985) are some undesirable side-effects of collaborative learning but yet they are difficult to detect and prevent. Assessing only the outcome or final artifact but not the process is another issue which resulted in groups taking the safe and tested instead of some unorthodox and innovative means of solving the allocated tasks. Most digital whiteboards record the historical traces of participants' interactions within the board and these traces are key to uncovering the process that learners took in resolving the problem and for distilling evidence of social-loafing and free-riding. This will be part of our investigation in this study.

2. Related Studies

Glover et al. (2005) evaluated the use of interactive whiteboards (IWBs) undertaken in five elementary schools in England. The responses were collected using questionnaires, observations, and interviews of 35 teachers. The results indicated that students' motivation was enhanced, and the students were more willing to interact and participate. This coincides with another study (Aflalo et al., 2018) where the authors surveyed and observed the students instead of teachers using questionnaires and observations. From the results, the authors concluded high student enthusiasm and satisfaction with correspondingly high active learning and participation rates from the students with the use of IWB. The authors of both studies acknowledged that more research is however required for conclusive results on the benefits of IWBs especially in encouraging and discovering evidence of learners' interactions in classrooms.

Campbell et al. (2019) leveraged a digital whiteboard for promoting graduate social work students to engage and collaborate with their peers on a fully online course delivery mode. A logic model digital worksheet assignment was designed to be completed by the learners on a collaborative basis. The learners then reviewed what their peers had added and answered questions on their experiences with the use of digital whiteboard as a tool for information exchange and learning. The authors employed a cross-sectional design using traditional open-ended survey questions and online focus group to elicit their learners' responses. From the analysis, the authors concluded that their learners engaged more with each other and had a greater motivation to learn with the use of digital whiteboard.

Assessment plays an integral role in driving students' learning appropriately (Knight, 2012). Notwithstanding the pedagogical benefits of collaborative learning, adequate assessment of individual inputs in group work to the final output is difficult as much of the work occurs outside formal teaching sessions. The study by Johnston * and Miles (2004) investigated the use of self and peer assessment to minimize free-riding effects and better evaluate individual contributions to the final artifact. The use of online collaboration tool such as digital whiteboard may triumph over a face-to-face collaborative setting in this aspect as the medium provides a written record of the interactions between learners (Macdonald, 2003). These interaction logs can be analyzed to gain a deeper understanding of the collaboration and interactions of learners for identifying free-riders. The analysis of these logs may offer further insight into the process of problem solving by the learners, allowing for the attribution of marks to the process on top of the final product.

Extending from the aforementioned work, this study thus addresses the following research questions:

RQ 1. What are students' perceptions of the digital whiteboard in terms of acceptance, collaboration and learning motivation??

RQ 2. What are the patterns of collaboration that can be derived from trace data in digital whiteboards?

3. Methodology

A total of 28 year-one university students (by convenience sampling) from the faculty of computer science and information systems undertaking the course of data management participated in this research. The author and another fellow instructor co-teach the class of 34 students (6 students opted out of participating in this research). The data management course spans a total duration of fifteen weeks and covers the fundamentals of relational database which includes data modelling, data design (Entity-Relationship diagram) and database implementation.

The digital whiteboarding software – Miro (http://www.miro.com) was introduced to the students in week 2 of the course. In each lesson, the students were first introduced to the concepts to be covered in the class. Other than instructor elaboration on some demonstration exercises, in-class exercises were dispersed throughout the entire lesson. An active learning approach was undertaken where students worked on the in-class exercises on the digital whiteboard collaboratively in pairs or in groups of three. Due to social distancing measures in our institution, arising from the Covid-19 pandemic, part of the class may be attending the lesson face to face while the others are attending remotely.

The digital whiteboarding software, Miro facilitates discussions and interactions among groups with a mix of remote and in class students. Miro offers collaboration features such as chat and zoom integration and more importantly synchronous interaction features where participants can see each other's annotations and drawings on a real time basis. In a typical lesson, instructors went around the groups and facilitated by asking guiding questions.

The students were also tasked with a group assignment to be submitted by week 8 (which involves designing an ER diagram for a provided scenario) and most students disclosed that they worked in a mix of synchronous and asynchronous fashion for group discussion. The students were allowed to form their own teams with the restriction that teams must comprise of 2 to 3 members.

The students were briefed on the aims and details of the research and only those who provided informed consent filled in an anonymous questionnaire, adapted from (Türel, 2011), on the use and benefits of digital whiteboard. A five-point Likert scale (Likert, 1932) was used for measuring the intensity of responses for the questionnaire.

Students who used Miro for their assignments were also requested on an optional basis to share their digital whiteboard canvas anonymously with the instructors for analysis of the historical interaction events logged by Miro. A total of 7 groups shared their Miro canvas with the instructors. This research is approved by the university's Institutional Review Board (IRB).

The board history feature in Miro stores a list of historical actions performed by the logged in user. The logged information includes the user who performed the action, the date and time of the action and details on the specific action performed e.g. add, edit and delete and the object acted on e.g. text, line, shape, sticky note e.t.c. Miro offers a Representational State Transfer (Fielding & Taylor, 2002) Application Programming Interface (REST API) interface for developers to extend on the current capabilities of Miro but unfortunately does not provide for the retrieval of board histories through REST API. Thus, we developed our own Python program to web scrap (Mitchell, 2018) the board histories of the seven groups' canvases. The scrapped logs are stored as comma delimited files. We then developed dashboards to pre-process and visualize the logs.

From the raw logs, we compiled a count of the actions for each user by the type of actions i.e. edit, add and delete and the average duration between each action. Durations that are more than five minutes are considered as pauses. We regard learners who were working on the same Miro canvas as being in the same team. This is a reasonable assumption as learners were instructed to share and work on their assignment only within their own team. The questionnaire response is summarized in Table 1.

Table 1. *Summary of questionnaire responses on the use of digital whiteboard*

Question	**SA/A (%)**	**Neutral (%)**	**SD/D (%)**	**Category**
I feel comfortable with the use of digital whiteboard tool.	89.29	7.14	3.57	Acceptance
I look forward to my professor's use of digital whiteboard in class.	85.71	14.29	0	Acceptance
Digital whiteboard should be used for more classes.	82.14	14.29	3.57	Acceptance
Digital whiteboard offers more convenience when engaging in collaborative discussions with my course mates.	89.29	7.14	3.57	Collaboration
Digital whiteboard allows me to collaborate better with my course mates.	89.29	10.71	0	Collaboration
Digital whiteboard is useful in pandemic situation as it allows us to collaborate online on exercises and assignments	78.57	17.86	3.57	Collaboration
I can learn more when my professor uses a digital whiteboard.	96.43	0	3.57	Learning Motivation
The use of digital whiteboard increases my interest in class	82.14	14.29	3.57	Learning Motivation
Digital whiteboard allows me to learn more effectively.	60.71	35.71	3.57	Learning Motivation

The Cronbach's alpha (Cronbach, 1951) values in Table 2 are above 0.8 for all three categories of Acceptance, Collaboration and Learning Motivation, thus indicating high internal consistency among the questions within each category.

Table 2. *Summary of questionnaire responses on the use of digital whiteboard*

Category	**Mean**	**Standard Deviation**	**Cronbach's alpha**
Acceptance	4.153	0.767	0.897
Collaboration	4.31	0.618	0.849
Learning Motivation	3.99	0.757	0.935

The high averages of 4.0 and above for all three categories indicate the students' high acceptance of digital whiteboard, high perceived ease of collaboration and learning motivation (RQ1) with the use of digital whiteboard.

From the processed historical action logs, we constructed the visualizations listed below to address RQ4.

- Total duration in days (from the date of their first action to the date of their last action on the canvas) that each team spent on the assignment.
- Number of actions by day across teams.
- Number of actions and average duration between actions across users by teams.

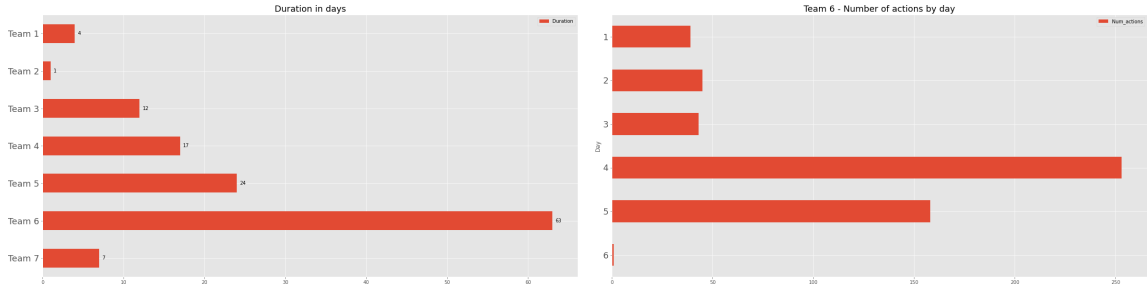


Figure 1. Left: Number of days spent on assignment, Right: Number of actions by day (team 6)

From Figure 1(left), we can see that team 2 worked on their Miro canvas for only a single day which is much shorter than the other teams. On the other hand, team 6 worked on the assignment for a longer duration as compared to the rest (almost triple that of the team 5). From the processed logs, we uncovered that team 6 started on the assignment earlier than the rest of the teams and worked on it till the due date. In Figure 1 (right), we drill into Team 6 for a breakdown of the number of actions by day. We sequence the day for the dates which the team worked on the canvas and summed up the number of actions performed by day. Thus, we can summarize from Figure 1 (right) that team 6 worked on the canvas consistently for at least five days. When we looked at similar graphs for other teams (not shown due to constraint of space), most of the other teams were working on the canvas for only two days.

To elucidate the contributions of individual members for the group assignment within each team, we constructed scatter plots of the number of actions by mean duration between actions by the individual members of the various teams (with the exception of team 2). We eliminated the records with has more than 50 actions as we noted that the initial copying and pasting of shapes and text from the template generated a huge number of actions. We are only interested in the edits and addition done by the students after this initial phase of copying and pasting. The duration between actions (in minutes) helps to distinguish between actions that does not require thinking e.g. moving of shapes from actions that require some time to contemplate (e.g. deciding on the relationship cardinality) which we will call "long actions".

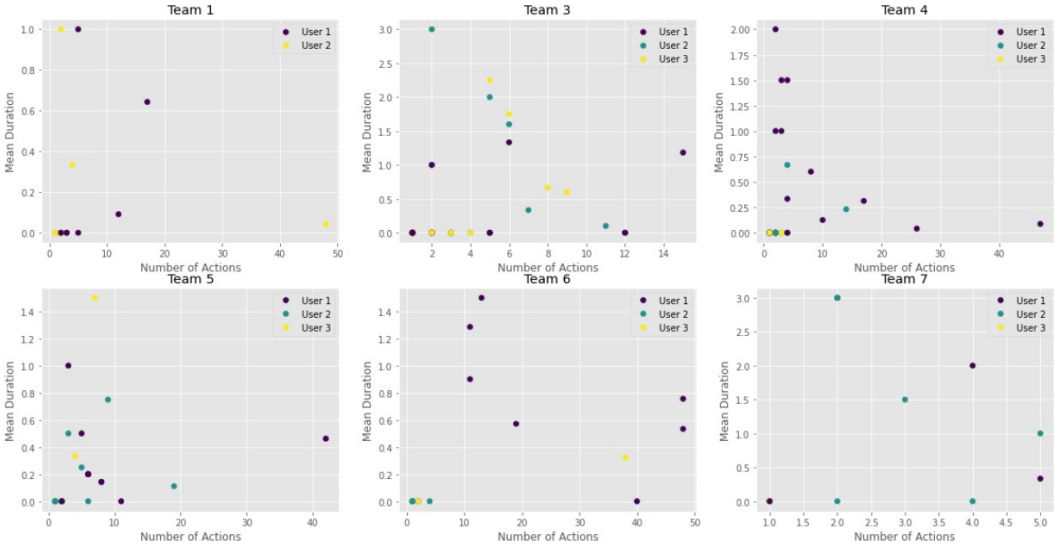


Figure 2. Number of actions and average duration between actions across users by teams

We can surmise from Figure 2 that across the teams, most of the members contribute to the assignment with non-zero number of actions with the exception of "User 3" in team 4 and "User 3" in team 7 (with close to zero actions). This offers us a quick way to identify potential free-riders which we may have to confirm further (e.g. by interviewing the teams or peer review). Focusing on the "long actions", we discover that the teams do have a mix of "long" and "short" actions with teams 4, 5 and 6 having more "long actions" as compared to the rest. When we further evaluated the ER diagrams of teams 4, 5 and 6, they are found to be more complete which correlates with more deep-thinking actions. We can also identify different team dynamics across the teams e.g. the contributions of members for some of the teams are evenly distributed but are uneven for other teams. To illustrate, both "User 1" in team 4 and team 6 dominates with proportionately more actions as compared to the other members. We postulate that the team dynamics factor could be further investigated on whether it influences the performance of the team.

4. Conclusion

From the results of the survey, we can conclude that the digital whiteboard is not only well-accepted by the learners but also perceived as fostering collaboration and leading to higher motivation. These address research question 1 which we set out in Section 2. In addition, we contend that assessment is difficult in collaborative work as much of the collaborative work occurs outside formal teaching hours. Typically, a major proportion of the assessment will be on the final produced output or artifact which leads to issues such as free-riding and difficulty attributing marks to individual contributions. We proposed to visualize the process of collaboration from the action logs recorded by the digital whiteboard Miro.

With these visualizations, we were able to discover differing interaction and collaboration practices among the teams as well as the contributions of individual team members within each team (e.g. some teams exhibiting uneven effort distribution).

In summary, we believe that our findings are significant as to the best of our knowledge, this is the first study which uses historical action logs from digital whiteboard to gain insights into the process and evidence of collaborative learning through visualization techniques. This work is just the tip of the iceberg in the use of digital whiteboard for collaborative learning. We acknowledge that the small sample size and use of only a single digital whiteboarding software may limit the generalizability of the study results. We intend to extend this work to cover diverse collaborative learning scenarios e.g. collaborative writing and with different digital whiteboard software.

References

Aflalo, E., Zana, L., & Huri, T. (2018). The interactive whiteboard in primary school science and interaction. *Interactive Learning Environments*, *26*(4), 525–538. https://doi.org/10.1080/10494820.2017.1367695

Albanese, R., & Van Fleet, D. D. (1985). Rational behavior in groups: The free-riding tendency. *Academy of Management Review*, *10*(2), 244–255.

Armstrong, V., Barnes, S., Sutherland, R., Curran, S., Mills, S., & Thompson, I. (2005). Collaborative research methodology for investigating teaching and learning: The use of interactive whiteboard technology. *Educational Review*, *57*(4), 457–469. https://doi.org/10.1080/00131910500279551

Astin, A. W. (1977). *Four Critical Years. Effects of College on Beliefs, Attitudes, and Knowledge.*

Campbell, M., Detres, M., & Lucio, R. (2019). Can a digital whiteboard foster student engagement? *Social Work Education*, *38*(6), 735–752. https://doi.org/10.1080/02615479.2018.1556631

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Fielding, R. T., & Taylor, R. N. (2002). Principled design of the modern web architecture. *ACM Transactions on Internet Technology (TOIT)*, *2*(2), 115–150.

Glover, D., Miller, D., Averis, D., & Door, V. (2005). The interactive whiteboard: A literature survey. *Technology, Pedagogy and Education*, *14*(2), 155–170. https://doi.org/10.1080/14759390500200199

Harkins, S. G., & Jackson, J. M. (1985). The role of evaluation in eliminating social loafing. *Personality and Social Psychology Bulletin*, *11*(4), 457–465.

Johnston *, L., & Miles, L. (2004). Assessing contributions to group assignments. *Assessment & Evaluation in Higher Education*, *29*(6), 751–768. https://doi.org/10.1080/0260293042000227272

Knight, P. (2012). *Assessment for learning in higher education*. Routledge.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Loes, C. N., Culver, K. C., & Trolian, T. L. (2018). How Collaborative Learning Enhances Students' Openness to Diversity. *The Journal of Higher Education*, *89*(6), 935–960. https://doi.org/10.1080/00221546.2018.1442638

Macdonald, J. (2003). Assessing online collaborative learning: Process and product. *Computers & Education*, *40*(4), 377–391. https://doi.org/10.1016/S0360-1315(02)00168-9

Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*. O'Reilly Media, Inc.

Piaget, J. (1952). *The origins of intelligence in children.* (M. Cook, Trans.). W W Norton & Co. https://doi.org/10.1037/11494-000

Price, E., De Leone, C., Henderson, C., Sabella, M., & Hsu, L. (2008). Archiving Student Solutions with Tablet PCs in a Discussion-based Introductory Physics Class. *AIP Conference Proceedings*, 175–178. https://doi.org/10.1063/1.3021247

Smith, B. L., & MacGregor, J. T. (1992). *What is collaborative learning*.

Tinto, V. (1997). Enhancing learning via community. *Thought & Action*, *13*(1), 53–58.

Türel, Y. K. (2011). An interactive whiteboard student survey: Development, validity and reliability. *Computers & Education*, *57*(4), 2441–2450.

Vygotsky, L. S. (1980). *Mind in Society: Development of Higher Psychological Processes* (M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4

Webb, N. M. (1982). Group composition, group interaction, and achievement in cooperative small groups. *Journal of Educational Psychology*, *74*(4), 475–484. https://doi.org/10.1037/0022-0663.74.4.475

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

The Development of Online Learning Readiness Scale for Junior Students (OLRS-J) in Taiwan

Chien CHOU[a*], Ruo-Yu LI[a], Li-Fang KO[a] & Min-Ling HUNG[b]
[a]*Institute of Education, National Yang Ming Chiao Tung University, Taiwan*
[b]*Teacher Education Center, Ming Chuan University, Taiwan*
*cchou@nycu.edu.tw

Abstract: This research seeks to develop the Online Learning Readiness Scale for Junior Students (OLRS-J) for 4-12 graders to measure their readiness for online learning. The authors first revised the existing Online Learning Readiness Scale by conducting intensive in-depth interviews. Secondly, the Fuzzy Delphi Method (FDM) was used for peer-review to establish the scale content validity. The dimensions, exemplary items, and future research work of the OLRS-J are presented and discussed in this paper.

Keywords: Online learning readiness, Fuzzy Delphi Method, Grade 4-12 students

1. Introduction

The pandemic of COVID-19 in 2020 has changed the way how people live, learn and communicate worldwide. Students around the world from kindergarten to higher education must go online to continue their formal education. For example, from mid-May to late June in 2021, Taiwan government decided the "suspension of classes without suspending learning" policy, asking all students to stay home to take online learning as the sole formal learning method. Facing this new policy, are students—from elementary to university—ready to do the online learning?

Online learning as one of the instructional/learning channels has been adopted and implemented for learners in higher education as well as life-long education for decades. Past research has also focused on these more-mature students' readiness for online learning. For example, Hung, Chou, Chen, and Own (2010) developed an instrument—Online Learning Readiness Scale (OLRS)—to study college students' preparation for this relatively-new way of learning. The OLRS is comprised of five dimensions: self-directed learning, motivation for learning, computer/Internet self-efficacy, learner control, and online communication self-efficacy. The results showed there were no gender difference but higher graders had significant greater readiness in some dimensions than lower graders.

Over the past 11 years, the OLRS has been cited frequently (i.e., more than 250 times in Web of Science and over 900 times in Google Scholar, such as Korkmaz, 2022; Tang et al., 2021). We as authors consider it is time to revise the scale for two reasons. First, the rapid change from face-to-face learning to online learning caused by COVID-19 not only have impacted college students but also elementary schoolers (6-12 year olds) and middle schoolers (12-18 year olds). It necessitates a boarder coverage of students by the OLRS. Secondly, the original 18 items in five dimensions in original OLRS, therefore, may not be enough for examining the readiness of elementary and middle school students. It is necessary to review/revise all dimensions and items of OLRS. Last, since the online learning platforms and social media available have grown abundantly in the past decade, examples and words in the OLRS need to be updated.

The purpose of this study is to develop a new instrument entitled "Online Learning Readiness Scale for Juniors (OLRS-J or J-Scale) for higher graders in elementary schools and middle schools. In other words, the target students of OLRS-J are 4-12 graders (i.e., children and adolescents aged 10-18 years old). This paper mainly describes the revision process: intensive in-depth interviews for more scale items and Fuzzy Delphi Method (FDM) for peer-review to establish its content validity. The up-to-date dimensions, exemplary items, and future research work will also be presented and discussed.

2. Methods

2.1 In-depth Interviews

To adapt the original OLRS, the present authors—all experts in e-learning theories and practices—have brainstormed to provide more possible items. Then we conducted 6 in-depth interviews: one e-learning professor, two school teachers (one for elementary school and one for high school), two experienced parents, and one high school student. These face-to-face interviews took 30 minutes to one hour long. Interviewees were either asked to freely talk about their students' or children's online learning experience in May, or were presented by and gone through our revised scale. The results of interviews helped draft the J-scale which contains 55 items in seven dimensions. Beside the original 5, two dimensions are added: Internet device affordance (4 items) and Family support (9 items).

2.2 Fuzzy Delphi method (FDM)

The purpose of FDM questionnaire is to examine the importance and appropriateness of 7 dimensions and 55 items. Dimensions and items with consensus values higher than a threshold would be included in our final version for student questionnaire. We invited 12 experts for FDM: 4 professors in online learning, 2 parents, and 6 teachers from elementary schools, middle schools, and high schools (2 have PhD in online learning). All of them have experienced in either conducting online learning themselves, developing online learning materials, or assisting their children to do online learning. The comments from experts were for consideration of new items and word/example selection.

We adopted two triangular fuzzy numbers to calculate the consensus values for all items. Specifically, the gray zone test was suitable for evaluating the consistency of the opinions given by experts. In addition, a scree plot was served to identify the threshold and screen the less important items, which ultimately helped us to obtain the priority order of key items. In terms of the research process, we conducted two rounds of the FDM questionnaire. In the first one, 6 items did not reach the consensus; some experts also provided comments on verbal expressions, items to combine, and extra items to consider. Therefore, we had 18 revised items and 2 new items for the second round.

3. Results and discussion

After two rounds of FDM, the final version contains 51 items in 7 dimensions. Table 1 shows the dimensions and exemplary items in OLRS-J.

Table 1. *Dimensions and exemplary items*

Dimension	Exemplary item
1. Computer and Internet self-efficacy (4 items)	• I feel confident in performing the basic functions of online tools (e.g., Google Classroom, Whiteboard, Google Jamboard, Kahoot). • I feel confident in search necessary information on the Internet (e.g., Google search).
2. Self-directed learning (6 items)	• I can set up my learning goals. • I have more free time to arrange my learning activities since the start of online learning.
3. Learner control (in an online context) (16 items)	• I am not distracted by other online activities when learning online (e.g., FB, IG, YouTube, web sites, online games). • I repeated the online instructional materials based on my needs.
4. Motivation for learning (in an online context) (7 items)	• I improve from my mistakes in assignments and exams. • I am open to new ideas (e.g., online learning)

Dimension	Exemplary item
5. Online communication self-efficacy (8 items)	• I feel confident in online communication with others effectively. • I feel confident in posting/answering questions in online discussions.
6. Internet device affordance* (4 items)	• I have enough computer equipment (e.g., desk computer, notebook, tablet, cell phone) for online learning. • I have steady and enough bandwidth for my online learning.
7. Family support* (6 items)	• My family members will help me manage online learning progress. • My family members consider "suspension of classes without suspending schooling policy" very important.

*new dimensions compared to original OLRS (Hung, Chou, Chen, & Own, 2010)

As Table 1 shows, compared with original 5 dimensions in OLRS for college students, 2 dimensions are added: Internet device affordance (4 items) and Family support (6 items). In addition, OLRS-J has 16 items in the Learner control dimension, compared to 4 in same dimension for the OLRS. This addition is because literature (e.g., Korkmaz, 2022) and all experts in interviews and FDM considered this dimension to be important and critical in the success of online learning. Moreover, the J-Scale divided motivation to intrinsic motivation (4 items) and extrinsic motivation (3 items) according to motivation theory (Deci, Koestner, & Ryan, 1999). Overall, the OLRS-J has examples and words updated to meet the new development of online learning applications and student usage scenarios; its 7 dimensions and 51 items with content validity are considered more appropriate to investigate younger students' online learning readiness.

The next step of this study is to distribute OLRS-J to Taiwan sample of students in grades 4-12. By doing so, we can establish its validity and reliability, conduct cross analyses, provide implementations for educators and parents, and share the scale and results with other researchers around the world.

Acknowledgements

This study was supported by the Ministry of Science and Technology, Taiwan (Project No. MOST108-2511-H-009-011-MY4). The authors would like to thank all experts in the interviews and FDM for their contributions to the development of OLRS-J.

References

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627–668 (discussion 692-700). doi:10.1037/0033-2909.125.6.627

Hung, M. L., Chou, C., Chen, C. H., & Own, Z. Y. (2010). Learner readiness for online learning: Scale development and student perceptions. *Computers & Education, 55*, 1080-1090. doi:10.1016/j.compedu.2010.05.004

Korkmaz, S. (2022). To what extent are preparatory school students ready for online learning? *Journal of Theoretical Educational Science, 15*(2), 239-263. doi:10.30831/akukeg.991017

Tang, Y. M., Chen, P. C., Law, K. M. Y., Wu, C. H., Lau, Y. Y., Guan, J., … Ho, G. T. S. (2021). Comparative analysis of student's live online learning readiness during the coronavirus (COVID-19) pandemic in the higher education sector. *Computers & Education, 168, 104211.* doi:10.1016/j.compedu.2021.104211

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Analysis of CSCL for Mathematical Proof Based on The Log Data of Learners' Verbal and Nonverbal Communications

Masataka KANEKO[a*], Hironori EGI[b] & Takeo NODA[c]
[a]*Faculty of Pharmaceutical Sciences, Toho University, Japan*
[b]*Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan*
[c]*Faculty of Science, Toho University, Japan*
*masataka.kaneko@phar.toho-u.ac.jp

Abstract: In this study, we investigate the process by which university students produced mathematical proofs in a CSCL environment. Students discussed how to prove the addition theorem for trigonometric function while manipulating mathematical content generated by a dynamic geometry system. The log data of their utterances and manipulations was stored and visualized. Besides the annotated text data of their communication and videotaped images of their activity, we analyzed the fine-grained log data. Analyzing how students produce proofs is difficult because we must recognize the critical moment of the proof by observing their activities while they are still engaged in the proof work. Furthermore, their insufficient background in mathematics prevents them from fully explaining their thinking processes. The result of this study indicates that the comparison of qualitative data of learners' communication and those multimodal log data can help us unfold their convergent conceptual change.

Keywords: Mathematical Proof, Dynamic Geometry, Log Data of manipulations, Log Data of Utterances, Multimodal Analysis on the Collaborative Work

1. Introduction

This study explores the process of how students prove the addition formula for trigonometric function in CSCL environment. As seen in some CSCL studies in mathematics (Oner, 2013; Schwarz, et al., 2018), using a dynamic geometry system can help learners visualize and simulate mathematical objects. The system used in this study is CindyJS (https://cindyjs.org/) which enables us to manipulate mathematical content on a web browser. A Moodle plug-in of CindyJS has been implemented to store the log data of learners' manipulations (Kaneko, et al., 2020). We also utilize another system to record the sound of each learner's utterance and store the log data of its volume (Ishikawa et al., 2019). The resources for mathematical activities are composed of many artifacts including gestures as a conceptual metaphor, written modes, spoken discourse, and visualization on digital media (Lakoff & Nunez, 2001). From the commognitive perspective, learners' discursive action makes them individualize the newly signified mathematical objects which are realized in those modalities. In the case of this study, learners experienced some commognitive conflict while generalizing their prior knowledge to newly emerged situation. Learners' participation in some discursive processes (metadiscourse) to handle the relevant commognitive conflict leads to the reification of those processes through which a new object is constructed (Sfard, 2008). From the commognitive perspective, learners' convergent conceptual change is identified with the change in pattern of their discursive process. Since the subtleties in those changes tend to be missed in simple field observations, fine-grained transcriptions of their communication and repeated viewings/hearings of recordings are needed (Koschmann & Zemel, 2009). This study proposes that high resolution multimodal data of learners' verbal and nonverbal communications can supplement their transcribed discussion.

2. Research Method and Results

The CindyJS content used in this study is shown in Figure 1 (left). By moving the red points on the unit circle, the angles α(red) and β(green) can be changed. The log data of the fluctuation of these angles are stored on Moodle with a time stamp. Three students were asked to manipulate the content on an iPad collaboratively. They were three first grade female students in a Japanese university. They had already learned the proof of the addition formula in the rudimentary case of when all angles are acute. Since only one student could manipulate the content at a time, they had to discuss each manipulation and each step of the proof. Using wearable microphones each student's utterances were recorded separately. Moreover, the students' activities around the iPad were videotaped. Before their collaborative work, they were asked to fill in a worksheet including the static figure in Figure 1 (right) to review that rudimentary case separately. The time they needed to fill in the worksheet was measured to estimate their prior skill level. The utterances and behaviors of students were annotated chronologically. The log data of the volume of each student's utterance was graphed with the log data of their manipulations. A previous study indicated that an utterance produced in unison might reflect the fact that it was the most obvious way for participants to describe the phenomenon as they understood it (Koschman & Zemel, 2009). In view of this suggestion, the concordance of some participants' utterances was specified. Using the manipulation data, the process through which students led to those utterances was examined. Simultaneously, using the sound data derived when some participants produced concordant utterances, the power balance between them together with their intention was presumed.

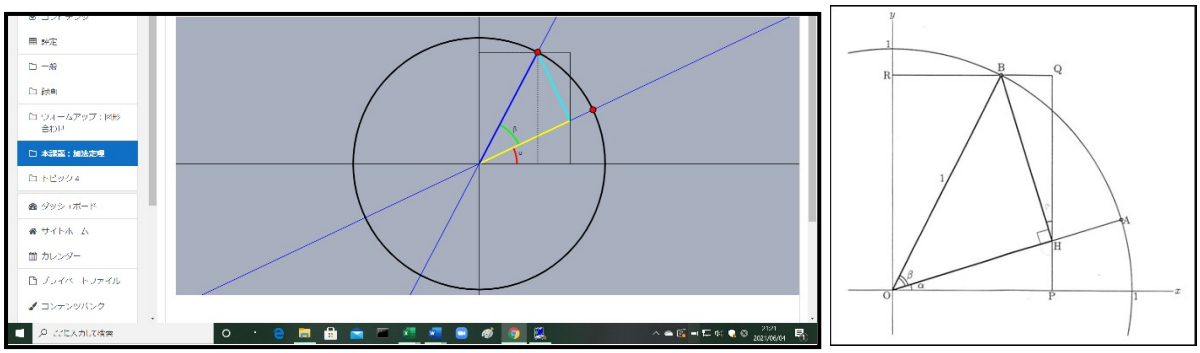


Figure 1. The CindyJS content used in this study (left) and static drawing in printed material (right)

The log data of the students' manipulations is visualized in Figure 2 (left). The horizontal axis pointing to the right represents the passage of time and the fluctuation of the angles α and $\alpha+\beta$ is represented in black and red respectively. Comparing this data to the annotated text of communication, it can be observed that participants' discussion had reached an impasse after completing the first step of the proof where the angle $\alpha+\beta$ was obtuse while the angles α and β were kept acute. In fact, the reason for the impasse was that the participants tried the complex case in which the angle β was negative and the instructor gave them some advice at 15 minutes after the beginning. Therefore, the authors analyzed only for the initial 10 minutes. The log data of the volume of each participant's utterance is visualized in Figure 2 (right). The data corresponding to the participants A, B, and C are represented in green, red, and black respectively. While it was observed that their prior skills were ranked as C>A>B, the result in Figure 2 (right) shows that the utterance of B was most dominant in the process of communication.

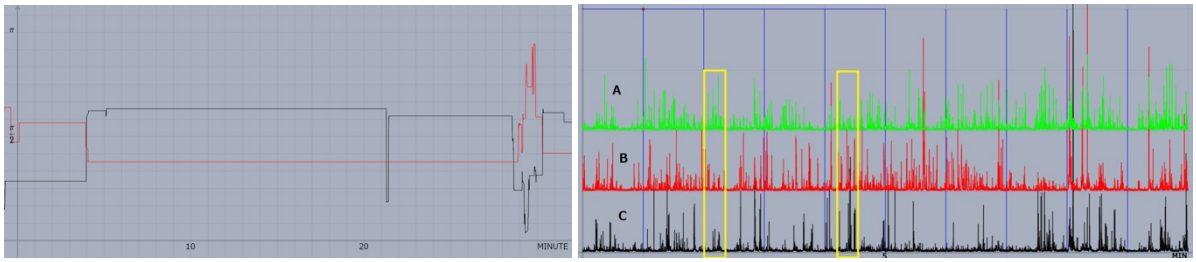


Figure 2. Visualization of manipulation (left) and the volume of each participant's utterance (right)

Their initial concern was to identify △BOH on the iPad screen. Then their attention moved to △HOP and △BHQ. In the discussion concerning △HOP, they had argued that the length of the leg OH was equal to $\cos\beta$ as in the rudimentary case. During this discussion, B often repeated the statements of A or C. Just after this discussion, A and C said "the leg is …" in unison when they argued about △BHQ. This first synchronous utterance occurred at 131 seconds after the beginning and it is specified with one of the two yellow rectangles on the left in Figure 2 (right). Following this communication, A said, "All right, also it (the fact that the length of the leg BH of △BHQ is equal to $\sin\beta$) is not different (from the rudimentary case)." Immediately after that, B repeated it. Judging from the annotated text, all participants could certify that the workflow on the worksheet was applicable to the relevant case until 210 seconds. The next trial (at 240 seconds) to prove the addition formula in case when the angle β is negative caused their struggle to identify the angles β and $\alpha+\beta$ which made it difficult to recognize the counterpart of the side OH of △HOP. The second synchronous utterance occurred in this situation (at 260 seconds) which is specified by the yellow rectangle on the right in Figure 2 (right). After B said, "the length of OH is…", B and C said "cosine beta ($\cos\beta$)" in unison. Figure 2 (right) shows that the volume of this second synchronous utterance was far higher and quicker than the first one.

3. Conclusion

The annotated text and videotaped record of the participants' communication show that the participants repeatedly alternated between referring to the drawing of those elements and referring to their symbolic representations. This repetitive pattern reflects the process in which all the components of the proof are organized and the narratives about that organization are reified into an object-like metarule. The remarkable point of the participants' concordant utterances is that they are concerned with △BHQ. Since the positional relationship of it to other geometric elements is highly dependent on the choice of the angles α and β, it can be assumed that those utterances reflected the participants' commognitive conflict as they try to accommodate their current discussion to that in the rudimentary case and triggered the emergence of convergent conceptual change. These results indicate that high resolution multimodal data of learners' verbal and nonverbal communications can supplement the annotated text and videotaped record to help analyze CSCL.

Acknowledgements

This work was supported by JSPS KAKENHI 18K02872, 19K03175, and 21K02752.

References

Oner, D. (2013). Analyzing group coordination when solving geometry problems with dynamic geometry software. *International Journal of Computer-Supported Collaborative Learning*, *8*, 13–39.

Schwarz, B., Prusak, N., Swidan, O., Livny, A., Gal, K., Segal, A. (2018). Orchestrating the emergence of conceptual learning: A case study in a geometry class. *International Journal of Computer-Supported Collaborative Learning*, *13*, 189–211.

Kaneko, M., Nakahara, T., Noda, T. (2020) Temporal analytics of log data derived from students' manipulating math objects. *Companion proceedings of LAK20*, 86–88.

Ishikawa, N., Okazawa, T., Egi, H. (2019) Diana-AD: Dialog analysis system for adjusting duration on real time in discussion. *Lecture notes in computer science*, *11677*, 212–221.

Lakoff, G., & Nunez., R. (2001) *Where mathematics comes from*. Basic Books.

Sfard, A. (2008) *Thinking as communicating: Human development, the growth of discourses, and mathematizing*. Cambridge University Press.

Koschmann, T., & Zemel, A. (2009). Optical pulsars and black arrows: Discoveries as occasioned products. *The Journal of Learning Sciences*, *18*(2), 200–246.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Estimating Divergence-or-Convergence in Discussions Based on Abstractness of Words in Utterance

Ryunosuke NISHIMURA*, Taisei MURAOKA, Risa IHARADA & Hironori EGI
Graduate School of Informatics and Engineering, The University of Electro-Communications, Japan
*n2230106@edu.cc.uec.ac.jp

Abstract: Estimating divergence-or-convergence of discussions is expected to increase the productivity of the discussions. Divergence means expanding ideas and convergence means summarizing ideas. However, previous estimating methods can only be applied to limited discussions. In this study, divergence and convergence are estimated using the abstractness of utterances. Divergence is assumed to be a concrete state and convergence an abstract state, respectively. The estimated results in the three discussions are compared to human judgment. As a result, agreement rates are 69.2%, 87.9% and 54.3%. Additionally, the abstractness of utterances in two discussions are low in divergence and high in convergence. The activeness in three discussions may cause the difference of the abstractness in utterances. These results suggest that abstractness of utterances has a potential to estimate divergence-or-convergence in general discussions.

Keywords: Discussion divergence, discussion convergence, discussion support system, abstractness of words, dialog analysis

1. Introduction

Dialog analysis in educational technology is used to support group learning. Especially, researchers have focused on estimating divergence-or-convergence of discussions, the expanding or summarizing of ideas respectively. This definition is based on divergent thinking and convergent thinking (Guilford, 1967; Cropley, 2006).

Estimating these two states, learners can learn strategies to diverge or converge discussion because learners understand good examples of utterances that contribute to divergence-or-convergence of discussion. For example, Mochizuki et al (2007) shows that estimating two states can help self-assessment of learners. Attritionary, good reflection is conceptualizing concrete experiences (Moon, 2013). From the above, learners can learn strategies to make divergent arguments converge by reviewing which of the two states the participants' comments enhance. This can increase the productivity of the discussion. In this study, we proposed a method for estimating the divergence and convergence of a discussion using the abstractness of words in the utterance. We also verified the effectiveness of the proposed method.

2. Related works

Abstractness of words is related to idea generation. Schön (1979) found that metaphors facilitate idea generation. This suggests that divergent thinking is encouraged when abstract problems are embodied by metaphors. Gonalves, Cardoso and Badke-Schaub (2012) quantitatively investigated the quality of idea generation for high abstractness text. In this study, it was observed that the group given a text about "Wizard of Oz" generated more ideas than those with a text on "Straddling Bus" (an example of transportation in the future). This suggests that the abstraction of object movement promotes idea divergence. Therefore, expanding ideas can be considered as embodying ideas.

On the other hand, in the idea convergence method such as the KJ method (Scupin, 1997), ideas related to a theme are written on cards, grouped, and headings are added to converge ideas. The headings

provide a concise summary of the contents of the grouped cards. Therefore, the abstractness of the word in the heading is considered higher than the idea on the card. That means summarizing ideas can be considered as abstraction of ideas.

From these studies, assuming that the same relationship holds in the discussion, estimating the divergence and convergence of arguments by calculating the portions where the abstractness of words in an utterance switch from high to low and otherwise is possible. Therefore, in this study, we use abstractness of words to estimate divergence-or-convergence of discussions. However, for an argument that produces a concrete answer, the abstractness of the words in the utterance may not be high upon discussion convergence. Therefore, this study focuses on discussions in abstract themes that do not have a single answer.

Additionally, previous divergence-or-convergence estimation methods can apply only to a specific discussion. Ichino (2011) estimated the divergence and convergence of an argument using nonverbal information from time-series utterance patterns. Here, the divergence-or-convergence of an argument was estimated by supervised learning using five features: frequency of utterances, percentage of overlapped utterances, length of utterances, the time between utterances, and the number of speaker transitions. She applied the results to a system that supports meetings in terms of both divergence and convergence. The results showed that the agreement among the three raters ranged from 77% to 85%. However, this study was conducted for a discussion on the 20-questions game, in which participants guessed the words that the contestant had in mind by answering 20 yes/no questions. Therefore, introducing the system into a discussion without specifying the format is difficult. In this study, we apply the method to general discussions.

3. Estimating divergence-or-convergence of a discussion based on abstractness

In this study, we considered abstractness as a value corresponding to a word. This abstractness may be changed by people's general and specific knowledge. Therefore, an abstractness dictionary based on crowdsourcing is used. The abstractness can be calculated using the dictionary.

Additionally, we assume divergence as embodying and convergence as abstraction. In other words, absolute abstractness may not be low in divergence and high in convergence. For this reason, we estimate divergence-or-convergence using relative abstractness of words. In this study, the validity of the estimation method is examined. Additionally, the results of estimating the divergence-or-convergence of discussion using the proposed method were compared with those of human recognition. We verified the relationship between the divergence and convergence of an argument and the abstractness of words in an utterance from the comparison results.

4. Methods

Figure 1 shows the method of estimation. First, a voice recording of the discussion was obtained and converted into text data using natural language processing and cleaning. In the cleaning process, the stilted were removed and misrecognized parts of speech were replaced with correct words based on the recording. Next, the obtained text data were divided into utterance units by meaning of the utterance and extracted nouns using morphological analysis for each utterance. Then, nouns with little relevance to the discussion, such as function words were removed. Next, the abstractness of each utterance and the abstractness of difference which represents the relative variation of abstractness were calculated.

In this study, utterances with relatively low and high-abstractness contributed to divergence and convergence, respectively. Therefore, the estimate of the divergence and convergence of the discussion with the negative abstractness of difference diverged and that with the positive abstractness of difference are converged. Notably, a three-valued or higher classification such as divergence-or-convergence, or neither was not used because labeling by a human concentrated in intermediate categories and appropriate evaluation of estimation accuracy was impossible. In the following, we described how we calculated the three abstractions used in the estimation of the divergence and convergence in the discussion.

4.1 Abstractness of a word

The abstractness of a word is a quantification of abstractness at the word level. The database of the abstractness of words for common Japanese words (Social Computing Laboratory, NAIST, 2019) was used for this study. There are two types of Japanese abstractness dictionaries: AWD-J Core (15220 words), in which abstractness is assigned by humans, and AWD-J EX (437299 words), in which abstractness is assigned by machine learning using the AWD-J Core as training data. In this study, AWD-J EX was used as the abstractness of a word dictionary because of its noun coverage. The abstractness of a word is a real number ranging from 1 to 5. The lower abstractness of a word indicates that the word is more concrete, whereas the higher abstractness of a word indicates that the word is more abstract. For example, the abstractness of "Tokyo" is 1.3, the abstractness of "city" is 2.0, and "metropolis" is 2.09.

4.2 Abstractness of each utterance

The abstractness of each utterance is defined as the average of the abstractness of a word in the utterance, where only words included in the AWD-J database are considered in the calculation of abstractness of each utterance. For example, "I have dogs and cats." has "dogs" and "cats" whose abstractness of a word is 1.45 and 1.54 respectively, the abstractness of the above utterance is (1.45 + 1.54)/2 = 1.495.

4.3 Abstractness of difference

The abstractness of difference is the relative abstractness of each utterance. It is defined as the difference between the abstractness of the utterance and the average of the abstractness of the four preceding utterances. If the four preceding utterances include utterances that do not have an abstractness of each utterance, they are excluded from the calculation. For example, it assumes that the abstractness of utterance A is 2.0 and the abstractness of the five preceding utterances are 2.0, 1.0, null (no abstractness), 2.0, and 2.0. Then, the abstractness of difference for utterance A is calculated without the utterance with null abstractness. Therefore, $(2.0 - (2.0 + 1.0 + 2.0 + 2.0))/4 = 0.25$ is the abstractness of difference for utterance A. The first four utterances have no abstractness of difference because they do not have any preceding four utterances.

In this study, we estimate divergence-or-convergence of discussions based on abstractness of difference. If the abstractness of difference is negative, the utterance is estimated as the divergence. If the abstractness of difference is positive or 0, the utterance is estimated as the divergence.

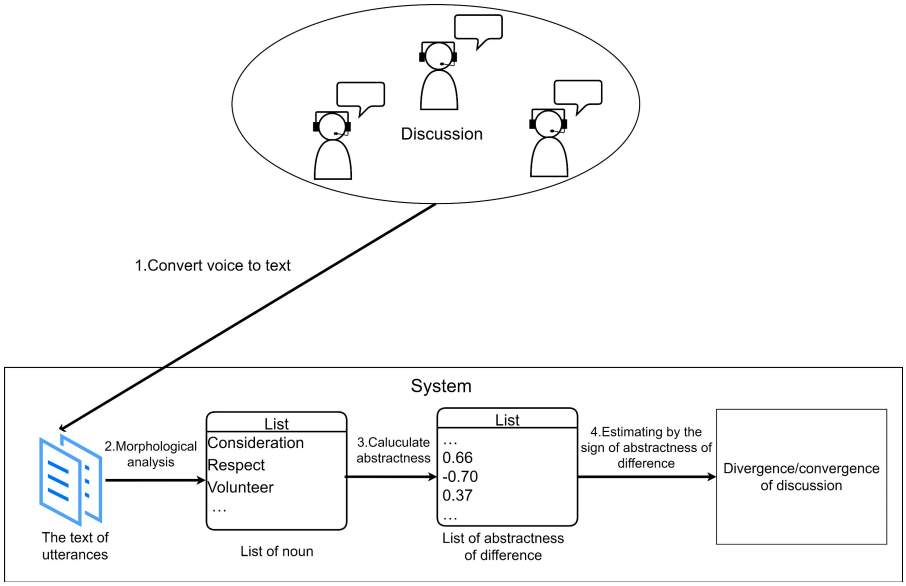


Figure 1. Procedure for estimating divergence or convergence discussion

5. Experiments

An experiment was conducted to survey the relationship between abstractness and divergence-or-convergence in a discussion and to evaluate estimation accuracy. Participants are undergraduate and graduate students of a science and engineering university in Japan. First, we collected three sets of 30-minutes discussion data under the topic "What is a thoughtful consideration?" from two groups of three participants and "Why is youth turnout low?" from one group of three participants. In this theme, turnout means voter turnout in elections. We define two discussions of "What is thoughtful consideration?" and a discussion of "Why is youth turnout low?" as Consideration 1, Consideration 2, and Turnout 1.

After the discussion, we asked the undergraduate and graduate students who did not participate in the discussion to label each utterance and collected the data of divergence and convergence as recognized by humans. These participants are instructed on how to determine the divergence and convergence of the discussion. For example, utterances were tagged as divergence when ideas were expanded and those with the immediately preceding summary were reexplained, whereas they were tagged as convergence in those utterances in which ideas were summarized and referred to the rules and regularities of "What is thoughtful Consideration?" (or "Why youth turnout is low?"). In this labeling, the participants seated in front of the monitor, read the text of the four previous utterances displayed on the input screen, and labeled each utterance as divergence-or-convergence. The reason for limiting the range of human speech reading to the four immediately preceding utterances was the need to restrict the information used for the recognition (estimation) of divergence and convergence. For instance, if there is no restriction on the range of utterances, participants will focus on different parts of the discussion depending on their comprehension of the content of the argument and their value criteria. Additionally, skimming may result in inaccurate results. For these reasons, participants labeled divergence and convergence as contextualized by the four immediately preceding utterances, which are the same as the system's estimation criterion.

In the labeling, the subjects were given printed materials containing the same contents as those explained in the divergence-or-convergence of discussion, so they checked the definitions when necessary. Additionally, subjects were told to label the utterance they judged contributed more when they were unsure whether to label an utterance as divergence-or-convergence. The participants took breaks and asked questions on the definitions of divergence and convergence at any time during the experiment. After the labeling, they were asked about any difficulties they had in labeling. The number of participants who had to label the discussions were 31 for the theme of "What is thoughtful consideration?" and 30 for the theme of "Why youth turnout is low?".

6. Results

6.1 Relationship between abstractness of each utterance and divergence-or-convergence of discussion.

We examined the difference in abstractness of each utterance between utterances labeled divergent or convergent by 80% of participants. The percentage was not set to 50% (over half) or 100% (unanimous) because the result of labeling was subjective. Here the groups of utterances labeled divergence and convergence by more than 80% of all subjects are referred to as the divergence and convergence set, respectively. The Brunner-Munzel test (Brunner and Munzel, 2000) was used as the statistical test method because the distribution of abstraction has two peaks (Brysbaert, Warriner and Kuperman, 2014), unequal sample sizes for the divergence and convergence set, and homoscedasticity of the population cannot be assumed. Table 2 shows the results of the Brunner-Munzel test for the divergence and convergence set for Consideration 1 and 2 and Turnout 1. There is a significant difference at the 1% level in the mean ranking of abstractness of each utterance between the divergence and convergence set in the "What is thoughtful consideration?" discussion, as shown in Table 2. This indicates that in the "What is thoughtful consideration?" discussion, abstractness of each utterance is low in the divergence of discussion and high in the convergence of discussion. The 95% confidence intervals for

the estimates of P(X1 < X2) + 0.5P(X1 = X2) were 0.657-0.909, 0.757-0.950 for Consideration 1 and 2, respectively, and 0.393-0.803 for Turnout 1.

Table 2. *The result of Brunner-Munzel test between the divergence set and the convergence set*

Group	p-value	Estimation values of P(X1 < X2) + 0.5P(X1 = X2)
Consideration 1	3.924×10^{-11}	0.783
Consideration 2	2.754×10^{-13}	0.853
Turnout 1	0.3351	0.598

6.2 Agreement between the labeling and estimation

Table 3 shows the agreement rates for both divergence and convergence in Consideration 1, Consideration 2, and Turnout 1.

Table 3. *Agreement between the labeling and estimation (divergence and convergence)*

Group	# of matches	Number of utterances with the same label by 80% of participants	# of utterances	Agreement (%)
Consideration 1	36	52	178	69.2
Consideration 2	58	66	158	87.9
Turnout 1	19	35	102	54.3

7. Discussion

7.1 Relationship between abstractness of each utterance and divergence-or-convergence of discussion.

Table 2 shows that the abstractness of each utterance is high and low for the divergence and convergence of discussion, respectively. The data confirmed our hypothesis in the discussion that has an abstract theme. This implies that people may talk about concrete topics like examples or experiences in divergence and talk about abstract topics like generalization or abstracted rules in convergence.

However, no significant difference exists between the divergence and convergence set in Turnout 1. The abstractness of the theme possibly affects the abstractness of the discussion. That is, embodying ideas may be hampered by the concrete theme because it is difficult to embody concrete concepts. Additionally, if divergence is insufficient, convergence may be also insufficient because discussions state alternately switches from divergence to convergence and from convergence to divergence.

7.2 Agreement between the labeling and estimation

Table 3 shows that the agreement between the abstractness method and human recognition is 69.2% at Consideration 1, 87.9% at Consideration 2, and 54.3% at Turnout 1. Therefore, estimating divergence-or-convergence of discussion by abstractness is reliable because the accuracy of the estimation is 50% if we choose the two states randomly. However, the accuracy of Turnout 1 is lower than that of Consideration 1 and 2. This can be attributed to the number of utterances since the number of utterances in Turnout 1 is about 100 utterances, whereas those in Consideration 1 and 2 are from 160 to 180. Nonetheless, the discussion time of Turnout 1 equals that of Consideration 1 and 2. Since the number of utterances per time is small, Turnout 1 does not diverge and converge sufficiently. The stagnant may lower the agreement of Turnout 1 because labeling is difficult for the participants.

The agreement in Consideration 2 is 20% higher than that in Consideration 1. This difference may be attributed to participants' speaking types. In the following, utterances in the divergence set that has high-abstractness of each utterance are referred to as the high-abstractness divergence set, whereas utterances in the convergence set that has low-abstractness of each utterance are referred to as the low-abstractness convergence set. For example, "In short, I think it is considerate to give what others need, but not expecting reward," is the low-abstractness convergence set in Consideration 1. This utterance has a concrete assumption that "It is considerate to give what others need." This decreases the abstractness of the utterance than "Consideration is not expecting reward." Another example, "If it continues to go unrecognized, it becomes commonplace," is the high-abstractness divergence set in Convergence 1. This utterance has few nouns such as recognition and commonplace. As a result, the abstractness of each utterance may be relatively high. Summarily, the type of utterance in Consideration 1 affects the estimation accuracy.

8. Conclusion

In this paper, we showed the relation between the divergence-or-convergence of discussion and the abstractness of words. Here, two significant results were obtained: the abstractness of words is low in divergence and high in convergence and the abstractness of words is fundamental to estimating the discussion state. Our study gives a quantitative perspective of abstractness in the discussion using qualitative analysis of abstractness and divergent/convergent thinking.

However, this study has some limitations. In this study, we used only the abstractness of difference. Therefore, it may be difficult to compare the contribution of utterances to divergence or convergence. Therefore, the abstractness of each utterance can be included in estimating as an indicator of absolute abstractness.

Acknowledgements

This work has been partly supported by the Grants-in-Aid for Scientific Research (Nos. 19K03175, 21K02752, 21K18527 and 22K02951) by MEXT (Ministry of Education, Culture, Sports, Science and Technology) in Japan

References

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 42(1), 17-25.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3), 904-911.

Cropley, A. (2006). In praise of convergent thinking. *Creativity research journal*, 18(3), 391-404.

Gonalves, M., Cardoso, C., & Badke-Schaub, P. (2012). Find your inspiration: exploring different levels of abstraction in textual stimuli. *In DS 73-1 Proceedings of the 2nd International Conference on Design Creativity Volume 1.*

Guilford, J. P. (1967). The nature of human intelligence. New York: *McGraw-Hill*.

Ichino, J. (2011). Discriminating divergent/convergent phases of meeting using non-verbal speech patterns. *In ECSCW 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work*, 24-28 September 2011, Aarhus Denmark (pp. 153-172). Springer, London.

Moon, J. A. (2013). A handbook of reflective and experiential learning: Theory and practice. *Routledge*.

Mochizuki, T., Kato, H., Fujitani, S., Yaegashi, K., Hisamatsu, S., Nagata, T., Nakahara, J., Suzuki, M., & Nishimori, T. (2007). *Promotion of self-assessment for learners in online discussion using the visualization software. In User-centered design of online learning communities* (pp. 365-386). IGI Global.

Schön, D. A. (1979). Generative metaphor: A perspective on problem-setting in social policy. *Metaphor and thought*, 2, 137-163.

Scupin, R. (1997). The KJ method: a technique for analyzing data derived from Japanese ethnology. *Human organization*, 56(2), 233-237.

Social Computing Laboratory, NAIST. (2019) awd-j: abstractness of word database for Japanese common words. Retrieved from https://sociocom.naist.jp/awd-j/.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

The Influence of Awareness of a Difference between Concept Maps on Transfer: Experimental Investigation on the Efficacy in Collaborative Learning

Shigen SHIMOJO[a*], Yoshimasa OHMOTO[b], Junya MORITA[b] & Yugo HAYASHI[c]
[a]*Graduate School of Human Science, Ritsumeikan University, Japan*
[b]*Faculty of Informatics, Shizuoka University, Japan*
[c]*College of Comprehensive Psychology, Ritsumeikan University, Japan*
*cp0013kr@ed.ritsumei.ac.jp

Abstract: This study investigated two types of feedbacks which we defined as directive feedback and indirect (facilitative) feedback for collaborative learning on concept explanation task. In this task, we investigated how these different types of feedback will influence on the performance. The results show that learners who received directive feedback outperformed the learning performance of those who did not receive interventions for both performance on learning and transfer performances. Also, learners who received facilitative feedback that present difference did not outperform those who had no intervention in terms of the learning performance. However, learners who receive facilitative feedback outperformed those who did not received an intervention with regard to the transfer performance. In a future study, we will need to investigate the transfer performance in detail and combination of directive and facilitative feedback.

Keywords: Collaborative learning, feedback, concept maps, transfer.

1. Introduction

1.1 Background

The Intelligent Tutoring System (ITS) focuses on learning support system based on learners' model (Anderson, Boyle, & Reiser, 1985). Some studies have recently investigated learning support in collaborative learning (Hayashi, 2020; Olsen et al., 2014). The features of such support based on learners' model are aimed at leading to a correct answer using small steps by revise misconception. By contrast, computer-supported collaborative learning (CSCL) focuses on how to facilitate the learning performance during collaborative learning using a computer (Dillenbourg, 1999). Such research has investigated how to facilitate a collaborative learning process not leading to an answer. Consequently, some feedback leads to an answer and others are not. However, there are no studies that have investigated the effect of intervention when comparing two feedbacks and how each feedback influences learning experimentally.

1.2 Need for Intervention in Collaborative Learning Using Concept Mapping

In collaborative learning, it is necessary to acquire the knowledge of others as a prerequisite. Concept maps are used as a tool to help with this. In Interactive-Constructive-Active-Passive (ICAP) framework, learners are facilitated by building concept maps individually and learners who build those collaboratively outperformed than they who build those individually (Chi & Wylie, 2014).

They have been used in a wide range of fields and investigated not only in practice but also in laboratory experiments. For example, in practical situations, support in collaborative learning using concept maps has been shown to improve the learners' understanding by helping instructors recognize

gaps in such understanding using Kit-Build concept maps (Pailai et al., 2017). In addition, Engelmann & Hesse (2010) used tools to visualize the knowledge of others during collaborative learning. These studies have mainly focused on an interface of learning support system. However, it is assumed that learners will be able to use and discuss these awareness tools to support collaborative learning, and that successful collaboration depends on the learners' abilities (Clarebout & Elen, 2006). Therefore, it is necessary for a facilitator to foster the learning performance using not only an interface (e.g., awareness tools) but also feedback (Janssen & Bodemer, 2013). Consequently, this paper investigates the effect of two feedbacks for learning collaboration in the context of using concept maps.

1.3 Directive and Facilitative feedback

The study to investigate feedback have shown the effectiveness of directive feedback and facilitative feedback (Shute, 2008). Directive feedback present information about correction and facilitative feedback present the information about provision (Archer, 2010). In the former, for example, Koedinger et al. (1997) investigated how to provide hints and a procedure for leading to answer when learners make a mistake. As a result, understanding of algebra was facilitated by learners following such feedback based on correct answer. However, some learners do not deepen their understanding. In the latter, for example, Hayashi (2020) investigated feedback that provides the information of the collaborator's eye movement and meta suggestion while learners learn a concept. As a result, learners received the other's information (e.g., knowledge and gaze) and a suggestion regarding the purpose of the task, which fostered their understanding of the concept. In addition, Shimojo & Hayashi (2021) indicated that intervention that learners externalize one own information and exchange the information facilitates the learning performance. In collaboration, learners need to acquire group members' knowledge (Janssen & Bodemer, 2013). Therefore, facilitative feedback is efficacy because learners are aware of group members' different knowledge. As the efficacy of facilitative feedback, the learners receive a collaborative learning process and transfer their understanding and knowledge to another task. Also, near transfer need to identify the basic element and structure of the task (Hung, 2013).

Consequently, two feedbacks are effective for near transfer because directive feedback provides the structure of the task and facilitative feedback provides help to identify basic elements and structure of the task by collaboration. Learners provided the latter feedback is also influenced by individual differences because it is difficult for learners to obtain a collaborative learning process. This paper investigates the effect of two feedbacks in learning performance and transfer performance.

1.4 Study goal and hypothesis

The study goal of the present study is to investigate the effect of a directive feedback and facilitative feedback during collaborative learning. First, we adapted the learning performance. It is expected that learners who have directive feedback will outperform learners who are without such feedback in terms of the learning performance (H1-1). In addition, it is expected that learners who have facilitative feedback will outperform learners without such feedback (H1-2). Furthermore, it is expected that learners who have directive feedback will outperform learners who have facilitative feedback (H1-3). Second, we adapted the transfer performance, and it is expected that learners who have directive feedback will outperform learners who are without feedback in terms of the transfer performance (H2-1). In addition, it is expected that learners who have facilitative feedback will outperform learners without such feedback (H2-2). Finally, it is expected that learners who have facilitative feedback will outperform learners with directive feedback (H2-3).

2. Method

2.1 Participants

In this study, there were 58 university students majoring in psychology (15 male, 43 female), the average age of which was 19.16 (SD = 0.91).

We adapted a factor between participant design. For the control condition, learners were without feedback. For the follow condition, learners were presented correct answer. Finally, for the relative condition, learners were presented difference.

2.2 Experiment Material

The learning material was a causal attribution of success and failure based on attribution theory (Weiner, 1985). As the attribution of success, a person infers a cause from three dimensions (internal/external, stable/unstable, controllable/uncontrollable). Learners were asked to infer the anxiety of a student (anxiety of a new semester) using concept maps individually and collaboratively, explaining one own idea to other.

We used two PCs and monitors during the experiment. In addition, the learners made a concept map by using concept map tools that we developed in C#. During the experimental task, we collected the facial expression, eye movement, and audio information using a video camera (Sony, HDR-CX680) and microphone (BeeFly, MF-MKF2020-JP). Figure 1 shows the learning situation and experiment device used.

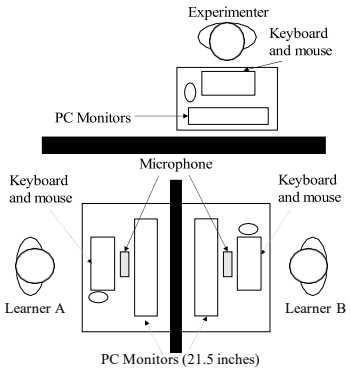


Figure 1. Learning situation and experiment device.

2.3 Procedure

The experimental procedure consisted of 8steps. In the present study, the leaners first answered questions regarding attribution theory to confirm whether there is difference in prior knowledge among the conditions (1). Next, the learners were instructed how to create concept map and what concept map is (2). Also, they learned attribution theory by reading learning text (3). After that, the learners learned a pre-test to investigate the experiment task (4). During the experiment task, the learners were asked to make concept maps collaboratively (6) after making such maps individually (5). Learners see each other concept maps that they made individually on the right side and made the concept map of collaboration on the left side. Two different points among the three conditions were whether the learners were provided prompts or not and whether they were provided prompts based on an answer or not. Details will be described in section 2.4 (Structure of feedback system). After the experiment task, the learners took a post-test (7). Finally, the learners made a concept map individually during the transfer task (8). In transfer task, episode was about procrastination and different with experiment task.

2.4 Structure of Feedback System

Learners under a follow condition were provided feedback presenting “correct answer”, leading to “correct concept maps”. We adapted classification that classified cause in episode based on three dimensions as a “correct answer” by using table 3 of Nasu (1989). Based on the classification, we discussed and create simple concept maps as a “correct concept maps”. Therefore, we divided the correct maps into nine steps. Figure 2 shows the correct concept maps. When the learners cleared one step of the nine steps, the system provided feedback for the next step. If they proceeded according to the steps, they were able to make the concept of the correct answer. Also, collaborative concept maps in a follow condition were divided into three parts because learners have to write phenomenon (e.g., anxiety) at the top, links (e.g., internal) at the middle, and cause (e.g., effort) at the bottom based on

figure 2 to create the correct map. Example prompts were "Create a node called Anxiety at the top of the three divisions. In addition, place the cause of the node at the bottom and the link at the center."

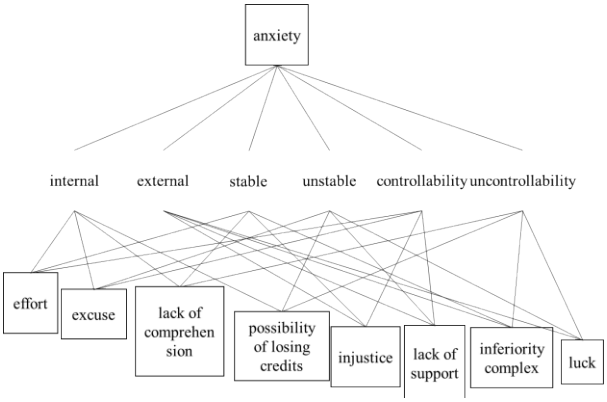


Figure 2. A correct concept map.

Under a relative condition, learners were provided feedback not based on an answer. In particular, the system provided information of different nodes and links between a pair in an experimental task (pre-individual).

2.5 Dependent Variables

This study used the learning and transfer performances as dependent variables by scoring the concept maps during the experiment and transfer tasks. The coder scored using 1 to 5 points. Here, 1 point indicated that the concept maps did not include three dimensions, whereas 5 points indicated that concept maps included three dimensions (internal/external, stable/unstable, controllable/uncontrollable) and associated three dimensions. First coder scored all of data. We calculated Krippendorff's alpha coefficient between first coder and second coder to investigate the reliability of coding by first coder based on (Schneider & Pea, 2014). The second coder scored 20% of all data. As a result, the coder's matching rate was 0.63. Therefore, that the coding was reasonably reliable and the coding of first coder was adopted.

3. Result

3.1 Evaluation of Experiment Task

In this section, we compared three conditions in the evaluation of the concept maps during the experiment task to test H1-1, H1-2, and H1-3. Figure 3 shows a comparison of the three conditions in the evaluation of the concept maps. We conducted a one-way ANOVA and there was a significant difference between conditions (F (2, 55) = 3.55, $p < .05$, and *partial* $\eta^2 = 0.11$). As a result of multiple comparisons in Shaffer method, the evaluation under the follow condition was higher than that under the control condition ($p < .05$). In addition, there were no significant differences between the control and follow condition, or between the follow condition and relative condition ($p = .20$, $p = .20$).

Consequently, H1-1 was supported, but H1-2 was not. In addition, H1-3 was partially supported because the evaluation under the follow condition was only higher than the control condition.

3.2 Evaluation of Transfer Task

In this section, we compared three conditions in an evaluation of the concept maps during a transfer task to test H2-1, H2-2, and H2-3. Figure 4 shows the comparison of the three conditions in the evaluation of the concept maps. We conducted a one-way ANOVA and there was a significant difference between conditions (F (2, 55) = 7.51, $p < .01$, and *partial* $\eta^2 = 0.22$). As a result of multiple comparisons in Shaffer method, the evaluation under the follow condition and the relative condition was higher than that under the control condition ($p < .01$, $p < .01$). In addition, there were no significant differences

between the follow and relative conditions (p = .20). However, there were no significant differences between the follow and relative conditions (p = 0.36).

Consequently, H2-1 and H2-2 were supported, but H2-3 was not because there was no significant difference between the follow and relative conditions in the transfer performance.

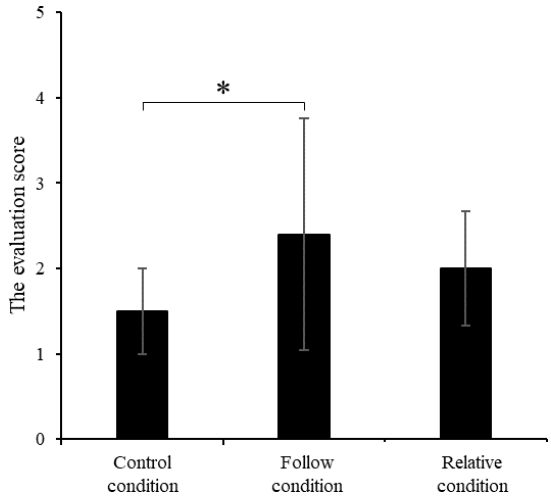


Figure 3. Comparison of three conditions in evaluation of concept maps of experiment task. Error bars are the standard deviation and * represents $p < .05$.

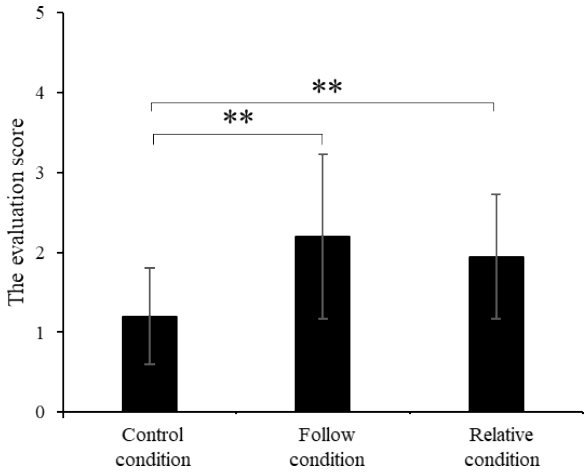


Figure 4. Comparison of three conditions in evaluation of concept maps of the transfer task. Error bars are the standard deviation and ** represents $p < .01$.

4. Discussion

The present study investigated how directive feedback based on an answer and facilitative feedback not based on an answer influence learning by comparing their feedback. We focused on the learning and transfer performances. H1-1 was supported, but H1-2 was not supported. In addition, H1-3 was partially supported because the evaluation under the follow condition was only higher than that under the control condition. For the transfer performance, it was expected that learners who have directive feedback would outperform learners without feedback in terms of the transfer performance (H2-1), learners who have facilitative feedback would outperform learners without feedback (H2-2), and learners who have facilitative feedback would outperform learners with directive feedback (H2-3). H2-1 and H2-2 were supported, but H2-3 was not supported. Consequently, each feedback was effective on the transfer performance.

Koedinger et al. (1997) indicated that directive feedback based on an answer is effective in terms of the learning performance. On the other hands, Hayashi (2020) indicated that facilitative

feedback is effective. This study indicated that directive feedback is effective in terms of the learning and transfer performance while making a concept map. However, the results also differed from those of previous studies (e.g., Hayashi, 2020) in terms of the non-solution-based feedback. The effectiveness of the feedback was only confirmed in the transfer performance. Hung (2013) showed that near transfer needs to identify the basic element and structure of the task. Therefore, directive feedback is effective because learners acquired the basic element and structure through referring correct concept map. Also, facilitative feedback is effective because learners acquired those through referring other concept maps and externalizing based on other's idea.

This result indicated that both feedbacks have each effectiveness and contribute to how to facilitate collaborative learning using computer. Although the effectiveness of each feedback was confirmed, because the transfer task was done immediately after the experimental task it is necessary to examine the effect of not the near transfer but the far transfer after a longer period of time. Also, we will need to investigate the learning performance in greater detail and combination of directive and facilitative feedback in the future.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP20H04299.

References

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science*, *228*(4698), 456-462.

Archer, J. C. (2010). State of the science in health professional education: effective feedback. *Medical education*, *44*(1), 101-108.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, *49*(4), 219-243.

Clarebout, G., & Elen, J. (2006). Tool use in computer-based learning environments: Towards a research framework. *Computers in Human Behavior*, *22*(3), 389-411.

Dillenbourg, P. (1999). What do you mean by collaborative learning?. In P. Dillenbourg (Ed.), Collaborative-learning: Cognitive and Computational Approaches (pp. 1-19). Oxford: Elsevier.

Engelmann, T., & Hesse, F. W. (2010). How digital concept maps about the collaborators' knowledge and information influence computer-supported collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning*, *5*(3), 299-319.

Hayashi, Y. (2020). Gaze awareness and metacognitive suggestions by a pedagogical conversational agent: an experimental investigation on interventions to support collaborative learning process and performance. *International Journal of Computer-Supported Collaborative Learning*, *15*(4), 469-498.

Hung, W. (2013). Problem-based learning: A learning environment for enhancing learning transfer. *New directions for adult and continuing education*, *137*, 27-38.

Janssen, J., & Bodemer, D. (2013). Coordinated computer-supported collaborative learning: Awareness and awareness tools. *Educational Psychologist*, *48*(1), 40-55.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., & Mark, M. A. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, *8*(1), 30-43.

Nasu, M. (1989). A study of Weiner's attribution theory of achievement motivation. *The Japanese Journal of Educational Psychology*, *37*(1), 84-95.

Olsen, J. K., Belenky, D. M., Aleven, V., & Rummel, N. (2014). Using an intelligent tutoring system to support collaborative as well as individual learning, *Proceeding of the 12th International Conference on Intelligent Tutoring Systems (ITS2014)*, 134-143.

Pailai, J., Wunnasri, W., Yoshida, K., Hayashi, Y., & Hirashima, T. (2017). The practical use of Kit-Build concept map on formative assessment. *Research and Practice in Technology Enhanced Learning*, *12*(1), 1-23.

Schneider, B., & Pea, R. (2014). Toward collaboration sensing. *International Journal of Computer-Supported Collaborative Learning*, *9*(4), 371-395.

Shimojo, S. & Hayashi, Y. (2021). Facilitating explanation activities using a concept map in collaborative learning: Focusing on coordination and argumentation process, *Cognitive Studies*, *28*(4), 499-521.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review*, *92*(4), 548-573.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

The Effect of Individual Ideation before Discussion on Computer Supported Collaborative Argumentation in a Primary Classroom

Wenli CHEN*, Qianru LYU, Junzhu SU
National Institute of Education, Nanyang Technological University, Singapore
*wenli.chen@nie.edu.sg

Abstract: This study investigated the effect of an individual preparation activity before collaboration on students' collaborative argumentation process and outcome. Conducted in a fifth-grade English class in a Singapore primary school, this study analyzed students' computer supported collaborative argumentation behaviors in two conditions: immediate collaboration condition and individual ideation before discussion condition. The results from statistical analyses and lag sequential analyses show a higher quality of collaborative argumentation in the individual ideation before discussion condition than in immediate collaboration condition. The results from content analyses and process-oriented bubble diagram show that students generated new ideas more frequently in the immediate collaboration condition. They engaged in behaviors of reading and revising existing ideas more frequent in individual preparation before collaboration condition. The findings contribute to the current understanding of CSCL scripts in terms of individual preparation activity before collaboration on the following collaboration outcome as well as the process. The implications on how to design and implement collaboration scripts in authentic teaching and learning scenarios are discussed.

Keywords: Collaboration scripts, individual preparation before collaboration, computer-supported collaborative learning, computer-supported collaborative argumentation

1. Introduction

The literacy of logical reasoning, arguing, and critical thinking are some of the most vital education objectives nowadays (Noroozi et al., 2012). Argumentation could be a vehicle for collaborative learning processes such as meaning-making and knowledge construction, with an interactive and collaborative nature (Baker, 2003). Collaborative argumentation has been integrated into real-world classroom contexts, identified as a fruitful and engaging process for students (Stegmann et al., 2012; Wang, 2014). Following the shift from page to screen, various computer-supported systems with unique learning affordances have been developed to facilitate collaborative argumentation processes. The learning approach of computer-supported collaborative argumentation (CSCA) aims at scaffolding productive argumentative learning by supporting sharing, constructing, and representing arguments in multiple formats (Kirschner et al., 2012). Research that closely examined CSCA illustrated its impact on the development of argumentation skills (Hsu et al., 2016; Lu & Zhang, 2013), critical and elaborative discussions (Scheuer et al., 2014), domain-specific knowledge (Stegmann et al., 2012) and problem-solving (Gillies & Khan, 2009). However, researchers found that collaborative argumentation remains challenging for primary school students despite the technological support. Appropriate instructional support is still required to engage students in quality argumentative practices (Harney et al., 2017; McNeill & Krajcik, 2006).

One effective approach to support CSCA is collaboration scripts, which consist of "sets of coordinated scaffolds that specify and sequence individual and collaborative learning activities and thereby aim to facilitate knowledge acquisition" (Stegmann et al., 2012, p. 301). Some collaboration scripts examples include the JIGSAW scripts (Aronson et al., 1978), ConceptGrid (Jermann et al., 1999), and the WiSim (De Jong & Van Joolingen, 1998). One common characteristic of the existing

collaboration scripts is a combination of individual work and joint work (e.g., the MURDER Scripts) with the purpose to prepare individual-level contributions (Mende et al., 2021). Though being designed in multiple ways, the integration of individual work and joint work shared one similar temporal structure: individual work before joint work. With an interest in the mechanism of collaboration scripts (Dillenbourg & Hong, 2008) to adapt to the different learning contexts, it is worthy of a deep examination of the effect of the particular phase shared by multiple collaboration scripts, individual preparation before collaboration.

Studied in various research contexts, individual preparation before collaboration refers to "providing learners with time to perform activities directed at processing the instructional material on their own before the collaboration" (Mende et al., 2021, p. 30). Empirical studies reported that with individual preparation, learners can prepare for the subsequent discussion and collaboration, e.g., reflect on their perspectives, create their own arguments, and prepare their individual solutions which they can be compared and integrated during the subsequent collaboration (e.g., Chen et al., 2021; Asterhan & Schwarz, 2007; Van Boxtel et al., 2000; van Dijk et al., 2013). However, Tsovaltzi et al. (2015, 2017) identified negative effect of individual preparation on subsequent collaboration. When there was an individual preparation activity before collaboration, some students experienced idea solidification without ongoing communication and ultimately. Other students presented lower quality of knowledge co-construction and less knowledge acquisition than without individual preparation condition.

Several under-studied areas are identified from the literature. First, there are mixed results reported in existing empirical studies, offering an unclear understanding of the effect of individual preparation before collaboration on collaborative learning, especially in CSCA activities. Second, most studies focused on students' individual learning outcomes or collaboration products without looking into the collaborative learning process in CSCL context, making it difficult to explain how and why students collaborated differently and produced different products. Therefore, this study aims to expand the existing understanding of the effect of individual preparation before collaboration on students' collaborative learning process and results, represented in collaborative argumentation product and the moment-by-moment on-screen behaviors. The following questions are proposed:

RQ1: What is the effect of the individual ideation phase on students' collaborative argumentation outcome?

RQ2: What is the effect of the individual ideation phase on students' CSCA process?

RQ3: What are the sequential patterns of CSCA behaviours emerged in immediate collaboration condition and individual ideation before discussion condition?

2. Method

2.1 Participants and Learning Context

A quasi-experimental design was employed in this study to explore the effect of the individual ideation activity on students' collaborative learning. A total of 28 fifth graders from the same class of a Singapore Primary school participated in this study. All participants were females, and their average age was 10 years. The teacher was a senior English language teacher who had taught the participating students for more than two semesters.

2.2 CSCL Environment

This study leveraged the AppleTree (Chen et al., 2013; Tan & Chen, 2022), a web-based system (www.appletree.sg) to support students' CSCA. The system was developed by the first author's research team at National Institute of Education, Nanyang Technological University. The AppleTree system supports both individual and group argumentation. The graph-based argumentation workspace of the AppleTree system allows the participants to construct the graph-based argumentation in a mind-map structure. Figure 1 shows the graph-based workspace of the AppleTree system. When engaging in argumentation on AppleTree, user would construct evidence (rectangle shape) to support claims (oval shape) or tentative ideas/claims (cloud shape). The arrows connecting evidence to claims dementated

the reasoning of the argumentation - green arrows indicate supporting evidence, red arrows indicate counterevidence, and grey arrows indicate neutral evidence.

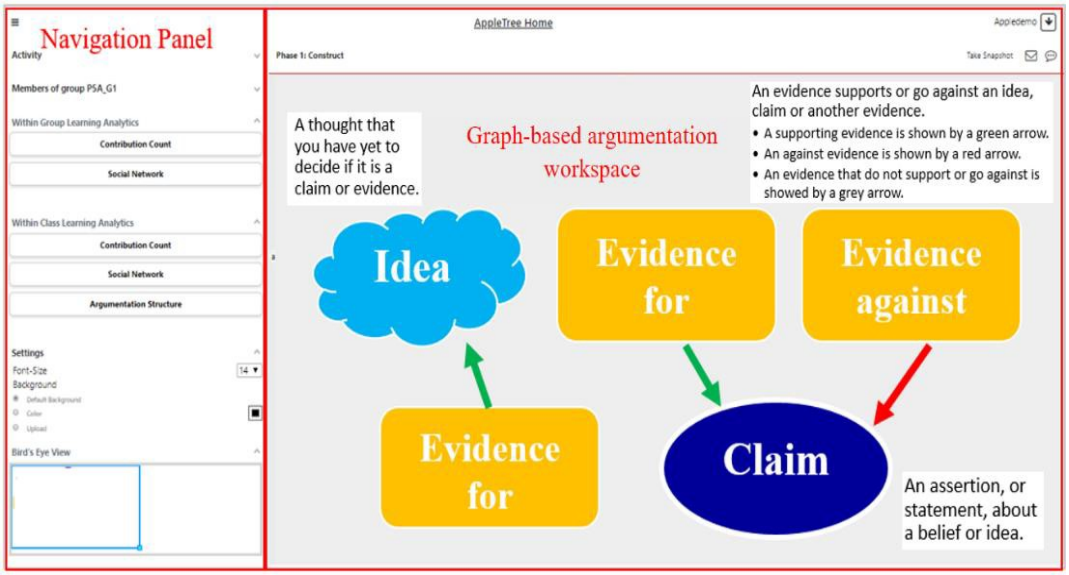


Figure 1. User interface of AppleTree

2.3 Design and Procedure

The two different collaboration conditions were conducted in two English lessons given on two separate school days. Designed by the schoolteacher and researchers of this study, the two lessons involved two comparable collaborative argumentation tasks that required the same amount of time (24 minutes) to complete. The first collaborative argumentation task is ranking three toys in terms of their education value: mystery board game, remote control car, and cube. The other task is ranking three class outing activities regarding how effective it will help class bonding. In each task, the two collaborating students sit together and work on AppleTree using their personal tablets.

On the first day, the immediate collaboration condition (control condition) was implemented. All the students worked in dyads and collaborated on the argumentation topic for 24 minutes. On the second day, the individual ideation before discussion condition (experimental condition) was implemented. For the first 10 minutes' individual ideation, students were asked to independently think and work on their individual argumentation graph (see Figure 2) on the AppleTree platform, during which no communication between students was allowed. After the individual ideation activity, the students were asked to start the discussion and co-construct their argumentation graph (see Figure 3).

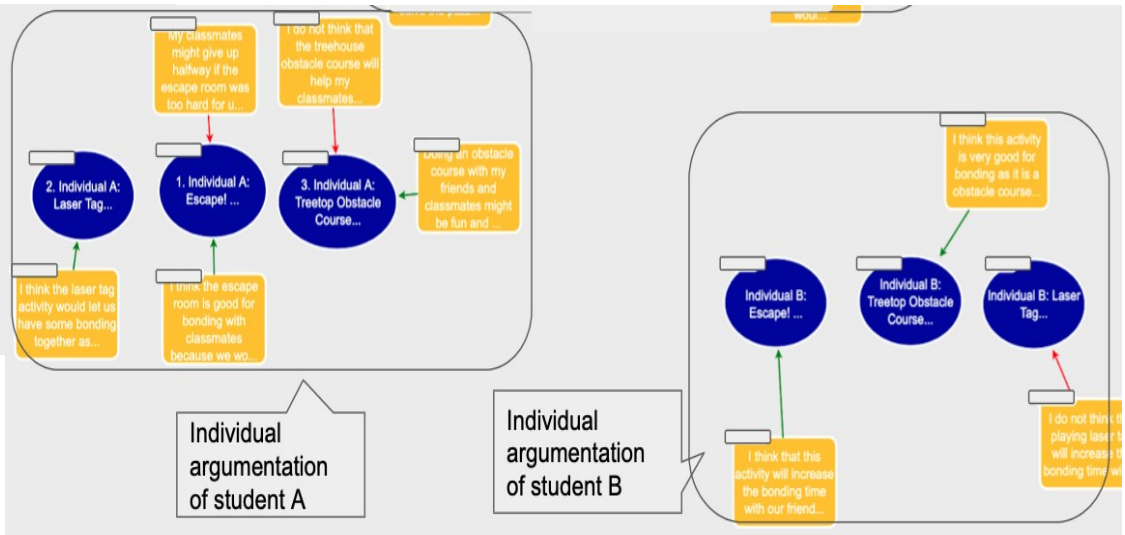


Figure 2. Individual argumentation for experimental condition

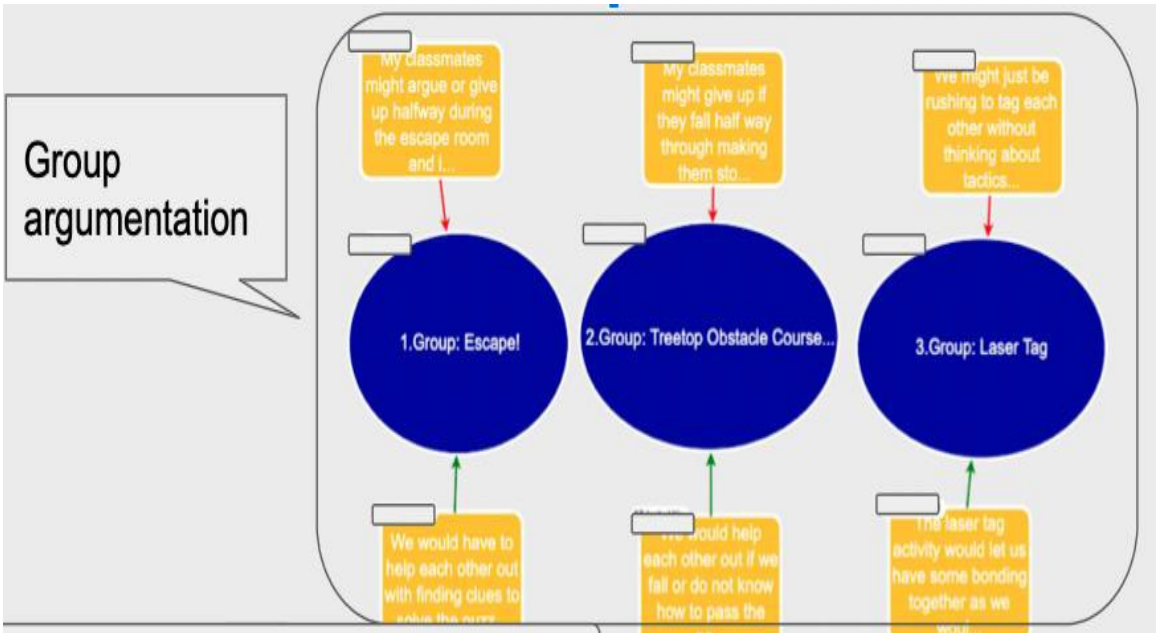


Figure 3. Group argumentation for experimental condition

2.4 Data Collection

To answer research question 1, all group's collaborative argumentation artefacts on the AppleTree platform were collected and analysed. To answer research questions 2 and 3, each individual student's screen recordings were collected. The video data was imported to the video analysis software Datavyu for further process-oriented analysis.

2.5 Research Instruments

2.5.1 Coding of Argumentation Artefacts

Content analysis was performed to examine the quality of the argumentation artefacts. The unit of analysis was the collaborative argumentation graph constructed by each group on AppleTree. Each group's CSCA artefacts on AppleTree platform were scored in four dimensions: clear position, multiple reasons to support the position, elaboration of reasons, and balance of supportive and objective evidence (counterarguments). Each dimension received a score ranging from 0 to 3. The scoring rubric was adapted from the basic argument schema developed by (Reznitskaya et al., 2007), informed by Toulmin's model (Toulmin, 1958). This basic argument schema was chosen for this study because it was developed and applied for evaluating fourth- and fifth-grade students' argument writing in both individual and collaborative learning contexts (e.g., Chen et al., 2016; Silverman et al., 2021).

Two trained coders coded the argumentation artefacts in the four dimensions according to the scoring rubrics. The sum of the four dimensions represents the overall performance of each group's collaborative argumentation artefact. The scores of both researchers were compared to check for inter-coder reliability. The inter-coder reliability was high (Cronbach's alpha = 0.75).

2.5.2 Coding of CSCA Process

To answer the second research question on CSCA process, this study analysed the CSCA behaviours for each student based on the screen recording data. To derive the coding scheme, the researchers coded the screen recordings for several students using a grounded coding approach. Then the researchers continuous adjusted coding scheme until the coding categories were applicable and suitable for further coding (see Table 1 for the coding scheme). The unit of analysis for the on-screen actions was each action taken by the student, such as creating one piece of evidence, editing on one piece of evidence, reading a partner's evidence, and messaging others via the chat-box function. One online action was

the same action from initiated till the next action was initiated. At the same time, the starting timepoint and ending timepoint of each coded action were noted down along with one specific code. In the end, the duration of each coded action was calculated for further analysis.

The two researchers coded all students' screen recordings. Both inter-coder reliability between two trained coders (Cohen's k = 0.807) (Landis & Koch, 1977) and intra-coder test-retest reliability for each coder for 10% of the data (95% of identical scores) were sufficiently high.

Table 1. *Coding Scheme of CSCA Behaviours*

Behavioural category	**Code**	**Description**
Generating Ideas	NISE	Generating a new idea individually without discussing with others.
	NIP	Generating a new idea based on partner's idea, not discussed at the group level.
	NISH	Generating a new idea that has been discussed at the group level.
Reading existing ideas	RS	Revisiting and reading ideas created themselves.
	RP	Revisiting and reading ideas created by the partner.
Editing existing ideas	CO	Copying and pasting the ideas without making changes
	CUT	Cutting existing ideas.
	EXP	Revising existing ideas by expanding them
Managing	CLAIM	Ranking on the given claims by editing the claim.
	LA	Checking learning analytics function on the platform.
	SEARCH	Searching information on the Internet that related to the argumentation.
	GA	Adjusting graph.
Off-tasking	ID	Idling on the platform.
	MESS	Off-task messaging with classmates from other groups via the chat function.
	PS	Playing the platform, such as moving around the argumentation artefact, zooming in and out without purpose.

3. Results

3.1 Effect of Individual Ideation Before Discussion on CSCA Outcome

A paired-sample t-test was conducted to compare the collaborative argumentation performance of each dyad under two different conditions. As shown in Table 2, the result of the paired-sample t-test presented a significant difference in the overall argumentation quality for the immediate collaboration condition (M= 5.571, SD=3.23) and the individual ideation before discussion condition (M= 7.929, SD= 3.050) (t (13) = -4.359, p < .01). The effect size of the significance test was calculated with Cohen's d of 1.44 which suggested a large effect size. This result indicates that students co-constructed higher quality collaborative argumentation artefacts when there was an individual ideation activity before discussion, compared to immediate collaboration.

Table 2. *Paired-Sample t Test of Argumentation Quality*

Condition	Control Group (without individual preparation)		Experimental Group (with individual preparation)		*t*
	M	SD	M	SD	
Score of quality of argumentation artefacts	5.571	3.23	7.929	3.050	-4.359**

**p< .01.

3.2 Effect of Individual Ideation Before Discussion on CSCA Process

To answer the second research question on the individual ideation's effect on the collaborative learning process, this study first adopted the content analysis to examine the nature of students' behaviours demonstrated in the CSCA process. A Chi-square test was conducted to compare the frequency distribution of the five categories of collaboration behaviours in the two conditions. As shown in table 3, there was a significant difference in the online collaboration behaviours between the control condition and experimental condition: X^2 (4, N = 14) = 217.5023, p < .01. Students demonstrate more frequent behaviours of "generating new ideas", "reading existing ideas", "editing existing ideas", and "managing" behaviours in the experimental condition than the control condition.

Table 3. *Chi-Square Analysis Results*

Behaviours	Immediate Collaboration		Individual Ideation before Collaboration			Chi-Square
	Frequency	Proportion	Frequency	Proportion	Total	
Generating Ideas	146	35%	199	29%	345	
Reading existing ideas	26	6%	128	19%	154	
Editing existing ideas	17	4%	102	15%	119	217.5**
Managing	94	22%	204	30%	298	
Off-tasking	140	33%	55	8%	195	
Total	423	100%	688	100%	1111	

** p <.01

On top of content analysis, the study employed a bubble diagram (Figure 4), a process-oriented analytical approach to examine students' CSCA behaviours in chronological order. The rows of the bubble diagram represent different behaviour categories. Different sub-dimensions of behavioural categories were assigned with different colours. Each bubble on every row indicates one behaviour happening at one certain time point, with its size indicating the duration of one behaviour.

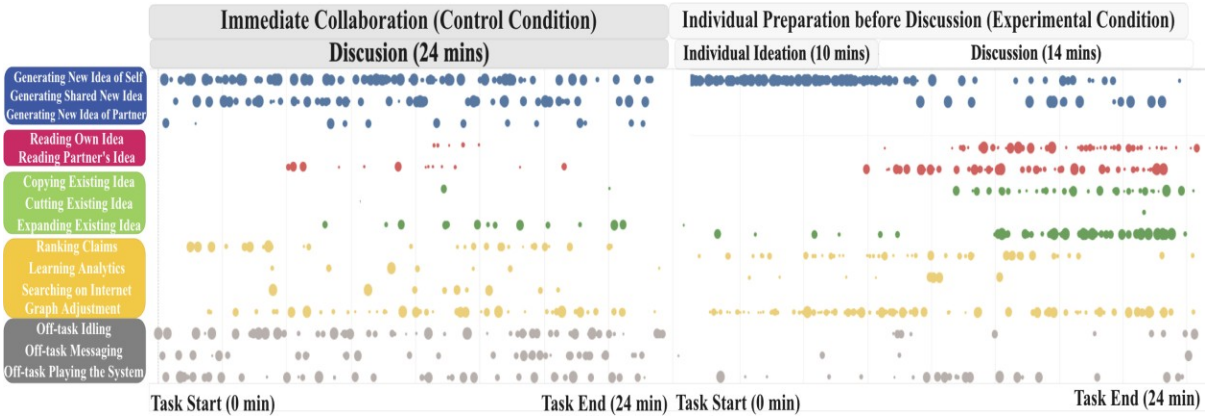


Figure 4. Distribution Diagram of CSCA Behaviours.

Figure 4 shows that there were differences in the distribution of different codes between the two conditions. Students generated new ideas (blue lines) in different timepoint as well as approaches. In the control condition (immediate collaboration), students tended to generate new ideas on their own (1st blue line) as well as from shared discussion (2nd blue line), a few new ideas were directly borrowed from their partner (3rd blue line). In the experimental condition (with individual preparation), most new ideas were produced at the individual ideation phase, based on students' own thinking processes (1st

blue line). There were a few new ideas generated during their discussion phase (2nd blue line). There wase no idea directly borrowed from partner (3rd blue line).

More action of reading existing ideas (red lines) were identified in experimental conditions. In the immediate collaboration condition, students randomly read existing ideas, either those written by themselves (1st red line) or the partner (2nd red line). In contrast, there were much more reading actions identified in the experimental condition, happening mainly during the discussion. Students tend to read ideas written by themselves (1st red line) and those written by their partner (2nd red line) much more frequently till the end of the task.

More editing actions (green lines) were identified in the experimental condition. In the immediate collaboration condition, students barely did actions of copying (1st green line) or cutting idea (2nd green line) though they expanded existing ideas (3rd green line) a few times. In comparison, students spent much more time editing existing ideas in the experimental condition. In particular, there was a higher frequency and longer duration of copying (1st green line) and expanding (3rd line) existing ideas during the discussion.

Students in the immediate collaboration condition showed more off-tasking actions (grey lines) than in the experimental condition, including idling, messaging others, and playing the system.

3.3 Sequential Analysis and Behavior Transition Patterns

To answer the third research question, the lag sequential analysis (LSA) (Bakeman & Gottman, 1997) was applied to examine the statistical significances of sequential patterns of behaviours under the two conditions. Lag sequential analysis takes into account of the occurrence, temporal order, as well as the sequential correlation between every learning behaviours (Kapur, 2011). GSEQ 5.1 is a computer program designed to analyse sequential data. The authors imported the coded collaborative learning behavior sequences into the GSEQ 5.1 and employed the table statistical calculation function, summarizing the frequency continuity of all behavior categories as well as the remaining results after adjustment of every transition. The z-score of each transition from one behavior to another, indicated by the adjusted residual results, determines the significance of the specific behavior transition (z score > 1.96, Bakeman & Gottman, 1997). Exported from the GSEQ 5.1, the lag sequential analysis results were presented in a symmetric matrix, in which the significant transitions were visualized in a transition diagram Figure 5. There were 19 and 21 significant behavioral sequences identified in the control condition and experimental condition respectively.

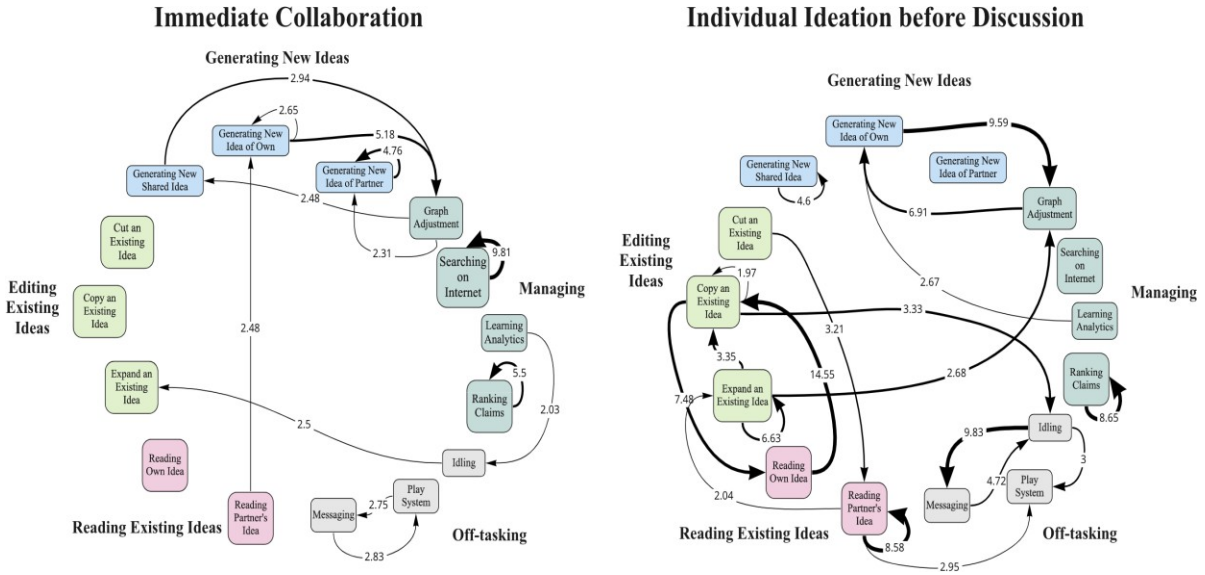


Figure 5. Behavioral Transition Diagram in Control Condition (left) and Experimental Condition (right).

Some transition patterns appeared in both conditions were "generating new idea of own" → "graph adjustment", which is understandable as students worked in a graph-based argumentation platform, and they tended to organize the argumentation graph as they added new content to it. Another

shared transition was the repeated "ranking claims", meaning students in both conditions made decisions on the argumentation claim together. These two paths were aligned with the specific CSCL context and the task requirement (ranking three toys or activities as a group).

There were distinctive paths in the transition diagram of the control condition. There was a central role of the dimension "generating new ideas", including generating ideas of own, partner, and shared. There were repeated occurrences of generating new idea of themselves, as well as new ideas of partner, indicating the students' generated new ideas intensively in control conditions. This finding echoed the findings on the frequency result and distribution trend identified in research questions 2 and 3. Besides, there was only one significant transition under the editing existing ideas dimension, from "idling" to "expanding existing ideas", indicating the few expanding actions were usually executed after some idling actions.

There are several transition features identified in the experimental condition. The first feature was the transitions between different "editing existing ideas" actions and the different "reading existing ideas" actions. Several transitions surrounding the "expanding existing idea" included its self-repetition and "reading partner's idea" → "expanding an existing idea". The self-repetition paths indicated that students' improvement of existing ideas usually constantly happened in the experimental condition. The path "reading partner's idea" → "expanding an existing idea" indicated that students were likely to work upon them by adding their input after reading the partner's ideas.

Another online collaborative behaviour that stands out in multiple significant transitions was "copying an existing idea" in the experimental condition. Keeping in mind its high frequency and even distribution identified in previous research questions, this diagram indicated that it not only happened continuously ("copy one's idea" → "copy one's idea"), but also strongly interacted with "reading one's own idea" ("copy an existing idea" → "reading one's own idea", "reading one's own idea" → "copying an existing idea"). These paths represented that apart from "expanding existing idea", students were also likely to copy and paste their written ideas produced during individual work without any changes to their group artefacts. In addition, the "copying an existing idea" action was also likely to lead to off-task action "idling", usually happening at the very end of the task according to the distribution graph. This inferred that some students tended to go off-task following the copying and pasting actions during the latter part of the task.

The diagram of the experimental condition indicates several significant transitions among the different off-tasking actions: idling to play system, idling to massaging, messaging to idling. Taking the results on the LSA, frequency account and distribution graph together, the findings show that students tended to get disengaged at the end of individual ideation activity as well as the overall task although they had significantly less off-tasking actions in the experimental condition than in the control condition.

4. Conclusions and Discussion

This study examined the effect of the collaboration script, individual preparation before collaboration on students' collaborative argumentation outcomes and processes. In terms of collaboration outcomes, there were significant improvements in students' collaborative argumentation, including the position, breadth, and balance dimensions. This finding contributes to existing literature on the effect of individual preparation before collaboration on students' collaboration products.

Proposed by (Lam & Kapur, 2018), the preparatory mechanism identifies the key roles of some cognitive preparation activities before future collaboration and knowledge consolidation. A preparation activity before collaboration would activate learners' knowledge, expose them to knowledge gaps, motivates them to engage in collaboration, and increase sensitivity to noticing relevant information (Lam & Kapur, 2018). In this study, the preparatory role of individual work before joint work was also reflected in the different collaborative learning behaviours. The immediate collaboration condition presented the central role of generating new knowledge throughout the overall task, and most of the ideas were barely improved further. In comparison, students were more prepared for in-depth discussion when there was individual ideation before the discussion. According to the frequency results and LSA, the findings on increasing reading of existing idea behaviors and editing existing idea behaviors indicated that students were able to not only generate new ideas, but also increase sensitivity to others' input, and put efforts to elaborate on the ideas.

The findings of the study resonated with the pre-requisite for meaningful discussion proposed by previous scholars of CSCL (Wichmann & Rummel, 2013). Accumulating research identified that quality communication in CSCL context is not easy to occur, as there lacks a "common focus" of collaboration, which is the basis for group coordination, generating shared ideas, gaining mutual understandings (Hord & Tobia, 2015). As found in primary school settings by Barron (2003), group members did not pay attention to others' opinions, interrupted them, and rejected other's suggestions without justification. These behaviours inhibited group functioning as well as individual's learning. The study by Ross (2008) also found primary schoolers often failed to realize shared knowledge construction, as the help-seekers were unable to formulate effective requests for help, and accordingly, the help-givers were barely able to provide meaningful help .

The process-oriented analyses results revealed various approaches students generated, read on, and improved ideas on their collaborative argumentation artefacts. With the analysis from multiple perspectives, the results indicated that in the immediate collaboration condition, there emerged a central role of sharing new knowledge and generating new ideas throughout the task. In the individual ideation before discussion condition, students tended to add more elaborations to support their standpoint or ideas and devoted more time to reading and editing existing ideas on the shared space. These findings were insightful for both educators and researchers who are interested in designing or implementing collaboration scripts in real-world classroom settings.

Acknowledgements

This study was funded by Singapore Ministry of Education (MOE) under the Education Research Funding Programme (OER 17/19 CWL) and administered by National Institute of Education (NIE), Nanyang Technological University (NTU), Singapore. The study was approved by the NTU IRB (IRB-2020-04-031). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Singapore MOE and NIE.

References

Aronson, E., Stephan, C., Sikes, J., Blaney, N., & Snapp, M. (1978). The Jigsaw Classroom. Beverly Hills, CA: Sage Publication.

Asterhan, C. S., & Schwarz, B. B. (2007). The effects of monological and dialogical argumentation on concept learning in evolutionary theory. *Journal of Educational Psychology*, *99*(3), 626.

Bakeman, R., & Gottman, J. M. (1997). *Observing Interaction: An Introduction to Sequential Analysis*. Cambridge university press.

Baker, M. (2003). Computer-mediated argumentative interactions for the co-elaboration of scientific notions. In *Arguing to Learn* (pp. 47-78). Springer.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences*, *12*(3), 307-359.

Chen W., Looi, C.K., Wen Y., & Xie, W. (2013). AppleTree: An assessment-oriented framework for collaboration and argumentation. In Kapur, M., Nathan, M., & Rummel, N. (Eds.) *Proceedings of the 10th International Conference on Computer-Supported Collaborative Learning* (pp. 6-9). International Society on Learning Sciences.

Chen, W., Tan, J. S., & Pi, Z. (2021). The spiral model of collaborative knowledge improvement: an exploratory study of a networked collaborative classroom. *International Journal of Computer-Supported Collaborative Learning*, *16*(1), 7-35.

Chen, Y.-C., Hand, B., & Park, S. (2016). Examining elementary students' development of oral and written argumentation practices through argument-based inquiry. *Science & Education*, *25*(3), 277-320.

De Jong, T., & Van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research*, *68*(2), 179-201.

Dillenbourg, P., & Hong, F. (2008). The mechanics of CSCL macro scripts. *International Journal of Computer-Supported Collaborative Learning*, *3*(1), 5-23.

Gillies, R. M., & Khan, A. (2009). Promoting reasoned argumentation, problem-solving and learning during small-group work. *Cambridge Journal of Education*, *39*(1), 7-27.

Harney, O. M., Hogan, M. J., & Quinn, S. (2017). Investigating the effects of peer to peer prompts on collaborative argumentation, consensus and perceived efficacy in collaborative learning. *International Journal of Computer-Supported Collaborative Learning*, *12*(3), 307-336. https://doi.org/10.1007/s11412-017-9263-9

Hord, S. M., & Tobia, E. F. (2015). *Reclaiming our teaching profession: The Power of Educators Learning in Community*. Teachers College Press.

Hsu, P.-S., Van Dyke, M., Chen, Y., & Smith, T. J. (2016). A cross-cultural study of the effect of a graph-oriented computer-assisted project-based learning environment on middle school students' science knowledge and argumentation skills. *Journal of Computer Assisted Learning*, *32*(1), 51-76. https://doi.org/10.1111/jcal.12118

Jermann, P., Dillenbourg, P., & Brouze, J.-C. (1999). Dialectics for collective activities: an approach to virtual campus design. Artificial Intelligence in Education,

Kapur, M. (2011). Temporality matters: Advancing a method for analyzing problem-solving processes in a computer-supported collaborative environment. *International Journal of Computer-Supported Collaborative Learning*, *6*(1), 39-56.

Kirschner, P. A., Buckingham-Shum, S. J., & Carr, C. S. (2012). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer Science & Business Media.

Lam, R., & Kapur, M. (2018). Preparation for future collaboration: Cognitively preparing for learning from collaboration. *The Journal of Experimental Education*, *86*(4), 546-559.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.

Lu, J., & Zhang, Z. (2013). Scaffolding argumentation in intact class: Integrating technology and pedagogy. *Computers & Education*, *69*, 189-198. https://doi.org/10.1016/j.compedu.2013.07.021

McNeill, K. L., & Krajcik, J. (2006). Supporting students' construction of scientific explanation through generic versus context-specific written scaffolds. Annual meeting of the American educational research association, San Francisco.

Mende, S., Proske, A., & Narciss, S. (2021). Individual preparation for collaborative learning: Systematic review and synthesis. *Educational Psychologist*, *56*(1), 29-53. https://doi.org/10.1080/00461520.2020.1828086

Noroozi, O., Weinberger, A., Biemans, H. J., Mulder, M., & Chizari, M. (2012). Argumentation-based computer supported collaborative learning (ABCSCL): A synthesis of 15 years of research. *Educational Research Review*, *7*(2), 79-106.

Reznitskaya, A., Anderson, C., Richard, & Kuo, L. J. (2007). Teaching and Learning Argumentation. *The Elementary School Journal*, *107*(5), 449-472. https://doi.org/10.1086/518623

Ross, J. A. (2008). Explanation giving and receiving in cooperative learning groups. In *The teacher's role in implementing cooperative learning in the classroom* (pp. 222-237). Springer.

Scheuer, O., Mclaren, B. M., Weinberger, A., & Niebuhr, S. (2014). Promoting critical, elaborative discussions through a collaboration script and argument diagrams. *Instructional Science*, *42*(2), 127-157. https://doi.org/10.1007/s11251-013-9274-5

Silverman, R. D., Proctor, C. P., Harring, J. R., Taylor, K. S., Johnson, E. M., Jones, R. L., & Lee, Y. (2021). The effect of a language and literacy intervention on upper elementary bilingual students' argument writing. *The Elementary School Journal*, *122*(2), 208-232.

Stegmann, K., Wecker, C., Weinberger, A., & Fischer, F. (2012). Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science*, *40*(2), 297-323. https://doi.org/10.1007/s11251-011-9174-5

Tan, S. H. J., & Chen, W. (2022). Peer feedback to support collaborative knowledge improvement: What kind of feedback feed-forward? *Computers & Education, 187,* https://doi.org/10.1016/j.compedu.2022.104467

Toulmin, S. E. (1958). *The Philosophy of Science* (Vol. 14). Genesis Publishing Pvt Ltd.

Tsovaltzi, D., Judele, R., Puhl, T., & Weinberger, A. (2015). Scripts, individual preparation and group awareness support in the service of learning in Facebook: How does CSCL compare to social networking sites? *Computers in Human Behavior*, *53*, 577-592. https://doi.org/10.1016/j.chb.2015.04.067

Tsovaltzi, D., Judele, R., Puhl, T., & Weinberger, A. (2017). Leveraging social networking sites for knowledge co-construction: Positive effects of argumentation structure, but premature knowledge consolidation after individual preparation. *Learning and Instruction*, *52*, 161-179.

Van Boxtel, C., Van der Linden, J., & Kanselaar, G. (2000). Collaborative learning tasks and the elaboration of conceptual knowledge. *Learning and Instruction*, *10*(4), 311-330.

van Dijk, A. M., Gijlers, H., & Weinberger, A. (2013). Scripted collaborative drawing in elementary science education. *Instructional Science*, *42*(3), 353-372. https://doi.org/10.1007/s11251-013-9286-1

Wang, T. H. (2014). Implementation of Web-based argumentation in facilitating elementary school students to learn environmental issues. *Journal of Computer Assisted Learning*, *30*(5), 479-496. https://doi.org/10.1111/jcal.12061

Wichmann, A., & Rummel, N. (2013). Improving revision in wiki-based writing: Coordination pays off. *Computers & Education*, *62*, 262-270. https://doi.org/10.1016/j.compedu.2012.10.017

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Interaction and Monitoring Matter: Comparison of High and Low-performing Groups in CSCL

Wenli CHEN*, Junzhu SU, Qianru LYU , Aileen CHAI, Wei Liang TOH
National Institute of Education, Nanyang Technological University, Singapore
* wenli.chen@nie.edu.sg

Abstract: This study explores students' interaction patterns and how monitoring influenced students' interactions when they were engaged in computer-supported collaborative learning (CSCL) activities. A comparison analysis was conducted between two groups of students (one high-performing group and one low-performing group) from a secondary school in Singapore. Content analysis was performed to compare the quality of student-generated artifacts between the two groups. Social network analysis (SNA) and lag-sequential analysis (LSA) were employed to analyze the interactive processes of the students. The results identified three differences in the ways the two groups engaged in collaborative argumentation: (1) quantity of social interaction, (2) diversity of interactive patterns (3) sequence of contributing to group work and seeking input. The findings reveal the possible interaction patterns which promote or constrain the development of collaboration argumentation. The implications on how teachers design and implement small group computer-supported collaborative learning is discussed.

Keywords: Interaction, computer-supported collaborative argumentation

1. Introduction

Collaborative competency is increasingly important in the workplace and life. Learners should learn how to collaborate effectively in classrooms (French & Kottke, 2013). How to promote meaningful group interactions and improve learning performance in computer supported collaborative learning (CSCL) have attracted much attention from researchers and educators in in recent years. One way to design and implement effective collaborative learning is to foster good quality interactions among students. Research studies showed that learning in groups can be more effective than learning alone (Furberg & Ludvigsen, 2008). Dillenbourg and Schneider (1995) defined collaborative learning as two or more subjects build synchronously and interactively a joint solution to solve a problem. The definition infers that the extent and the quality of the exchanges that occur within groups of students in collaborative environments is important (Curtis & Lawson, 2001). Assessment of CSCL should go beyond the learning outcomes; the quality of the collaborative learning processes is also important (Strijbos, 2010). This study explores how students in high-performing and low-performing groups engaged in collaborative argumentation in a CSCL environment to learn by interacting with one another, and to identify interaction behaviors that led to desired learning outcomes.

2. Literature Review

Researchers have been interested in argumentation and how students can benefit from it in CSCL (Stegmann et al., 2007). "Collaborative argumentation" has been important because students can learn critical thinking, elaboration, and reasoning from it (Andriessen, 2006). In argumentative dialogue, learners are expected to put forward suggestions for the analysis and solution of the problem, challenge their proposals, back them up with theories, rebut opposing views on theoretical grounds, and weigh the available theoretical evidence that favors or disfavors possible solutions (Resnick et al., 1993).

Effective collaborative learning requires mutual engagement of learners constructing knowledge and solving problems with joint effort (Dillenbourg, 1999). In collaborative learning

activities, students need to actively make sense of the subject matter by articulating relevant concepts, considering multiple perspectives, and discussing alternative solutions to the problem (Slof et al., 2010). Interactions within the group help the group move towards or away from its goal, or maintain its status quo (Kapur et al., 2011).

Researchers have also found that the groups with students monitoring their own and their peers' learning and thinking processes seem to have an advantage over groups without doing it (Lee et al., 2015). The fact that different groups showed different interaction patterns in collaborative learning was more often group-related than task-related (Vuopala et al., 2016). High-performing groups focused on regulating the cognitive and social aspects of their collaboration, which was not the case with the low-performing groups. The low-performing student groups did not activate the socially shared regulation (Malmberg et al., 2015). A comprehensive understanding of students' collaborative learning requires analysis of both learning process and outcomes (Tan & Chen, 2022). However, the interaction patterns of higher and lower performing groups in collaborative argumentation remain underexplored.

This study examines how students from high- and low-performing groups interact within group in computer-supported collaborative argumentation (CSCA). The objectives of this study are to investigate the high- and low-performing groups interactive network and the behavioral characteristics of students in online collaborative learning activities. Two research questions are specified as follows:

1. What are the differences in terms of interaction network between the high- and low-performing groups in CSCA?
2. What are the differences in terms of behavioral characteristics between the high- and low-performing groups in CSCA?

3. Method

This study employed the exploratory case study method. It examined the collaborative argumentation process of two groups of students (N=9) from an English language class of a secondary school in Singapore.

3.1 Participants

There were ten groups of three to five Secondary-3 students (aged around 15 years old) in the class participating in computer-supported collaborative argumentation. Content analysis was conducted to assess the high- and low-performing groups based on students' written arguments at the end of the activity. The groups with the highest quality argumentation and lowest quality argumentation were included in the data analysis for this study.

3.2 CSCA environment

The CSCA activities were conducted in an online collaborative learning environment named AppleTree (Chen et al., 2021; Chen et al., 2019), a graph-oriented, computer-assisted application to support students' collaborative argumentation. The process of participants interacted in a intra-group synergy phase was analysed. During the intra-group synergy phase, the participants were asked to deliberate and integrate group members' ideas in order to be able to present quality argumenation via a group mind map. Figure 1 shows the screen capture of one group of students working on CSCA in AppleTree.

Instead of the traditional teacher-centered instructional approach, the CSCA approach was adopted. In this study, students in groups co-constructed argumentation in the form of an argumentation graph with claim, evidence, and reasoning. During collaborative argumentation in the English language lesson focused on argumentation, students sat in groups and worked on an online collaborative argumentation platform, Appletree. They brainstormed and put forth their opinions to the question "Technology is the solution to climate change. Do you agree?" in the AppleTree system. At the end of collaborative argumentation task, each student completed an argumentation writing for 30 minutes.

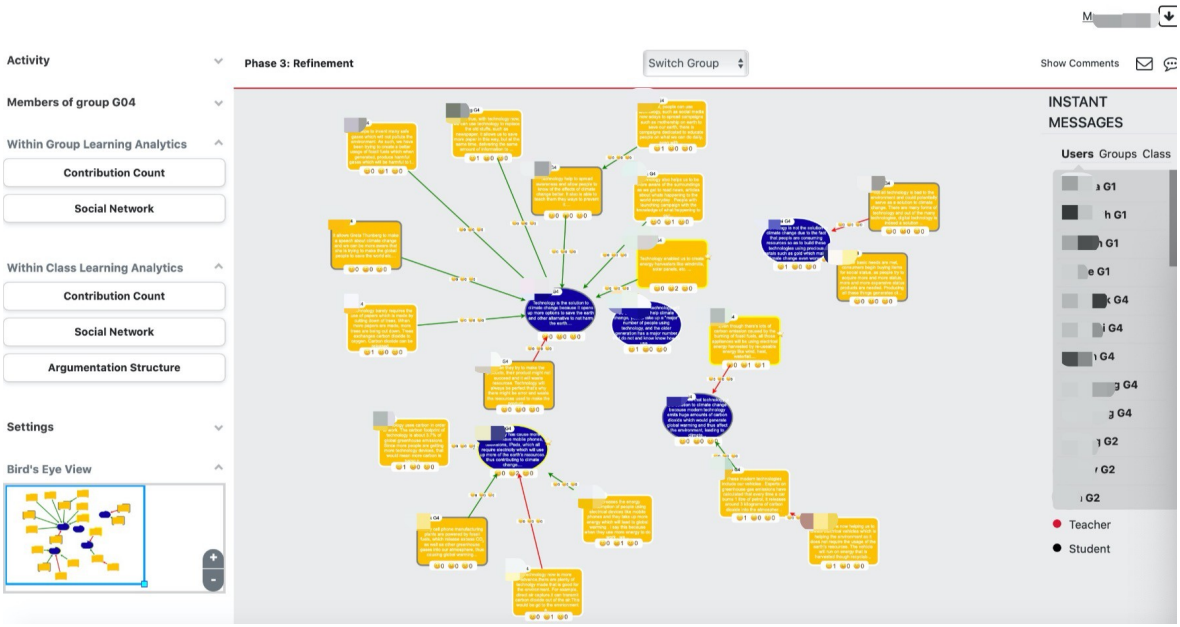


Figure 1. The screen capture of one group's argumentation in Appletree System with Learning analytics on the left and chat function on the right side

3.3 Data Collection and Instrument

The students' online interaction data during the activities and individual argumentative writing pieces at the end of the activities were collected. When students co-constructed arguments on the Appletree system, all the online interaction data were stored automatically in the Appletree System. Figure 2 shows the online artefacts generated by the two groups. The sequence of the online interactive behavioral were captured by the backend of the system. The two groups of students' online behaviors were screen recorded and transcribed.

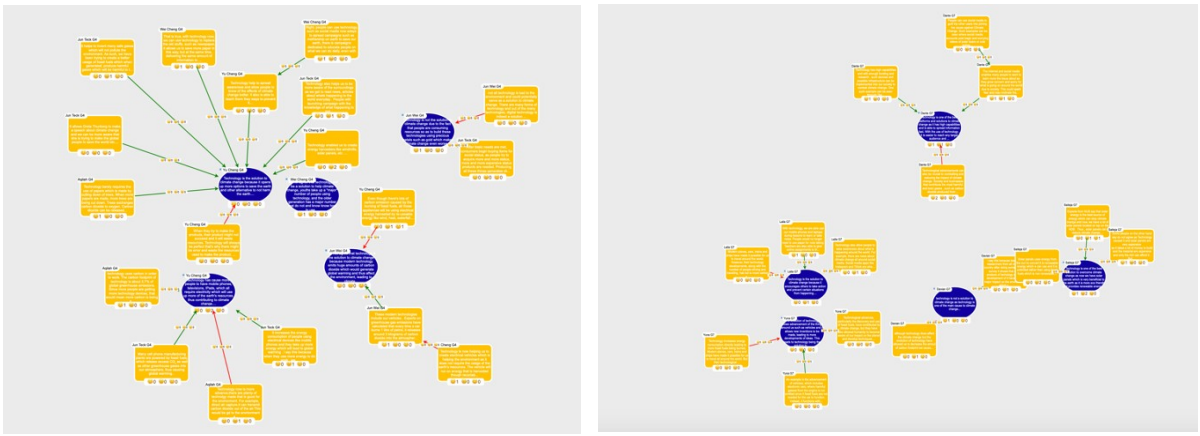


Figure 2. The mind map of high-performing group (Left) and low-performing group (Right) on the Appletree System

The students' learning outcome(argumentative writing) was coded according to the code scheme, including four dimensions (Table 1). Among them, Logical reasoning was modified based on Toulmin's way (1958) of viewing argumentation, Depth was modified based on Fan et al (2020)'s assessment rubric for argumentation with a closer look at thinking skills, Clarity was adopted from (Andrade et al., 2010) focus on the quality of writing/meaning-making of students' argumentative artifacts, Persuasiveness (Connor, 1990) looks at the effectiveness of rhetorical techniques including logical appeals, affective appeals, and ethical appeals, and writers' ability to keep audience in mind represented in students' argumentative artifacts.

Table 1. Coding scheme of quality of learning outcome (argumentative writing)

Dimension	0	1	2	3
1. Logical reasoning	No clear point or has so many errors in logic that it is invalid.	Makes a point but is not well articulated, or not supported by evidence; contains some errors in logic.	The logic is acceptable; provides some logical supporting reasons.	Well-articulated and logically strong; sound and cogent; nuanced arguments consider multiple perspectives logically.
2. Depth	Irrelevant ideas.	A superficial argumentation, more implicit than analytical or explanatory, unexamined hunch or borrowed ideas, lack of explanation and interpretation.	Reflects some in-depth thinking process such as reflecting, evaluating, and considering different perspectives.	A thorough, explanatory, and sophisticated argumentation; supported, justified; consider and evaluate multiple perspectives to back up the claim; deep and broad.
3.Clarity	The central idea and the clarity of purpose are absent.	The central idea is basically expressed though it may be vague or too broad.	The central idea and the clarity of purpose are generally evident.	The central idea and the clarity of purpose are well developed; very compelling statement.
4.Persuasiveness	No persuasive appeals used (e.g., logos, pathos, ethos) / not attuned to audience.	Make use of simple persuasive appeals; not very effective word choice and language use attuned to audience.	Make use of several persuasive appeals and techniques; relatively effective word choice and language use attuned to audience.	Make good use of several persuasive appeals and techniques; sophisticated word choice and language use attuned to audience.

The students' online interactions were coded by adopting the coding schemes of Curtis & Lawson (2001) and Popov et al., (2019). There were seven categories of online collaborative behaviors (Table 2): organizing, contribution, seeking input, monitoring learning analytics, refining and revising, social interaction and activity-related individual behavior. Each category of collaborative behavior had several subcategories. In total, there are 17 subcategories of collaborative behavior.

Table2. Coding of on-screen behavior

Behavior categories	**Code**	**Description**
Organizing	OGM	Organize Group mind map
	OIM	Organize individual mind map
Contributing	ASE	Add supporting evidence
	AOE	Add opposing evidence
	AI	Add idea bubble
Seeking input	RGW	Read Group members' work (within Group)
	ROW	Read other's work (other Groups)
	MCC	Monitoring contribution count
Monitoring learning analytics	MSNA	Monitoring SNA (within Group/within class)
	MMS	Monitoring mind map structure

Refining and Revising	RPW	Revising previous work
Social interaction	GM	Group message chat (unrelated to activity)
	RARE	Read activity requirements and extracts
Activity-related individual behavior	NAV	Navigate the system
	INT	Internet information search
	TH	Thinking, drafting, idling

Inter-coder reliability. Two researchers coded all the data independently by using the aforementioned coding scheme. An overall Cohen's Kappa of .706 was found, a satisfactory result.

4. Results

4.1 The differences in terms of interaction networks between the high- and low-performing groups in CSCA?

Social interaction behaviors in CSCL have been found to be effective in predicting learning (Kang et al., 2018). Social Network Analysis (SNA) helps researchers in the analysis of these patterns by illuminating the 'flow' of information that are exchanged among students. SNA was used to answer the first research question by examining the dyadic interaction of the participants. Figure 3 shows the SNA results of the two groups. The node represents each student collaborator. The line represents the relationship between collaborators. The interaction means the exchanges between students. For example, the line between students indicates they build on each other's idea by adding ideas or evidence on each other's ideas. The diagrams in Figure 3 show how students connect to the members in the group and the patterns of collaboration (one-to-one or many-to-many). According to SNA results, there were regular interactions among students within high-performing group (N=27). In contrast, there was only one interaction relationship between two students within the low-performing group. The results also indicated high cohesiveness and tight interaction relationship among students in the high-performing group.

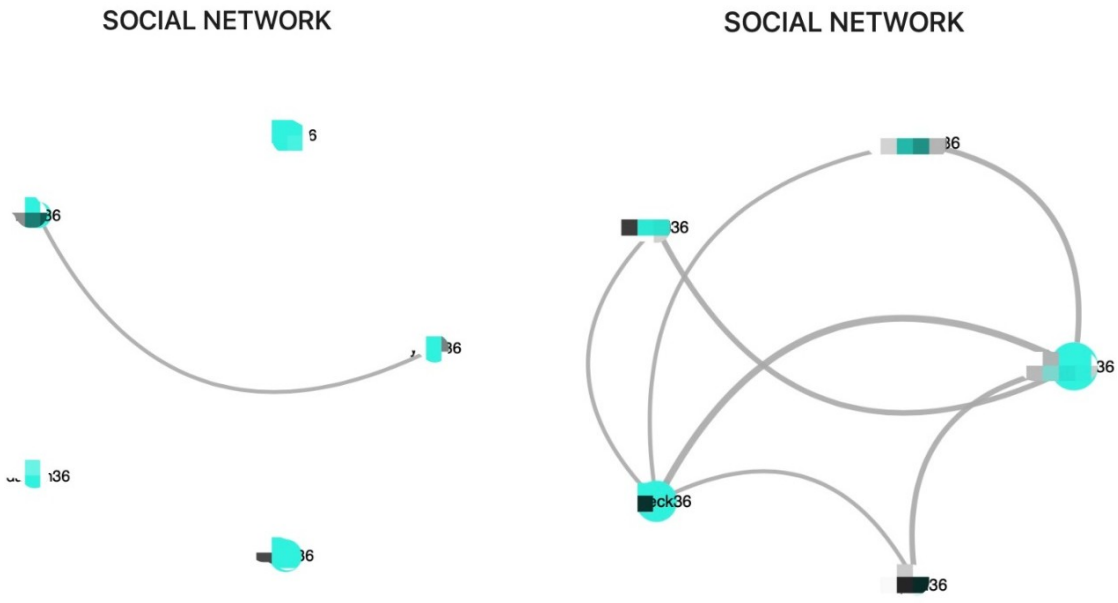


Figure. 3 Comparison of the Social Networks of high-performing group (Left) and low-performing group (Right) (the dots represents students, the thickness of the lines represent exchanges of information)

4.2 The differences in terms of behavioral characteristics between the high- and low-performing groups in CSCA

To answer the second research question, Lag Sequence Analyses (LSA) was conducted. LSA tests the probability that one behavior occurs after another behavior, and whether it is statistically significant (Bakeman & Gottman, 1997). GSEQ 5.1 was used for LSA. GSEQ5.1, has been widely used in many behavioral pattern studies (Hou & Wu, 2011). Firstly, we imported the coded collaborative behavior sequences into GSEQ 5.1 and saved it as an independent document. Then the function of calculating table statistics in GSEQ 5.1 was used to summarize the frequency of behavioral types and the adjusted residual results of transitions (Z - score > 1.96 indicates the behavior path has significance). Finally, we illustrated the behavior transition diagrams based on all the sequences that were statistically significant.

The LSA results of the two groups were shown in Figure 4 to illustrate the behavioral patterns of knowledge construction in group collaboration.

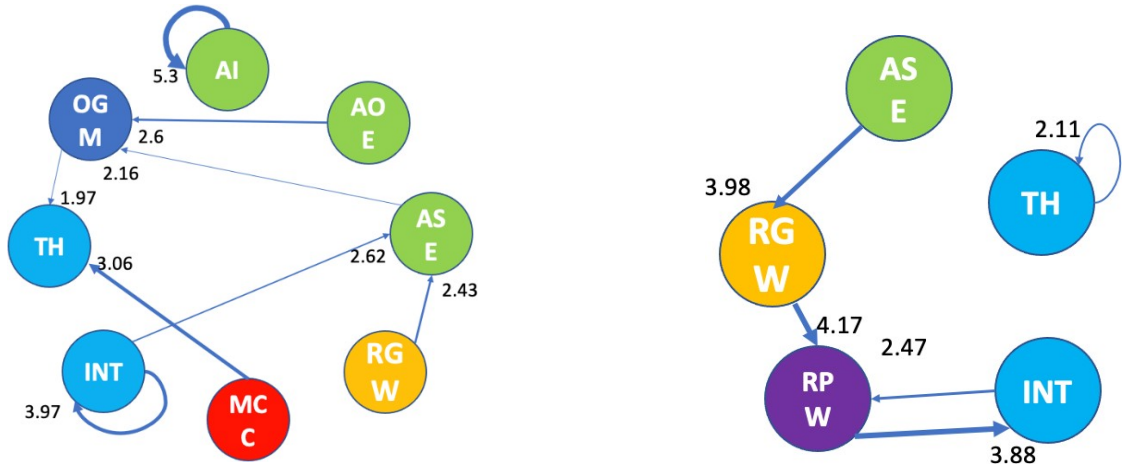


Figure 4. Behavioral transition diagram of high-performing Group (Left) and low-performing group (Right)

Behavior transition diagrams of high-performing group and all sequences that reached statistical significance are shown in the Figure 4. In total, there were eight significant behavioral sequences identified in the high-performing group. Behavioral paths ASE→ OGM, and AOE → OGM showed that after students added new evidence, they would usually organize the group argumentation graph. Behavioral paths MCC → TH and OGM → TH showed that students monitored contribution count or organized group mind map, followed by thinking or idling. Behavioral paths INT → ASE showed that students searched for online information, after which they added supporting evidence. Behavioral paths RGW → ASE showed that they added supporting evidence after reading group members' work.

Behavior transition diagrams of low-performing group and all sequences that reached statistical significance of the group are shown on the right of Figure 4. There were five significant behavioral sequences in this group in total. The Behavioral paths RGW→ RPW, and INT → RPW showed that after students read the ideas from their group members on the AppleTree system or searched online for information, they tended to revise their work individually. Behavioral paths ASE→ RPW showed that students read group work after adding supporting evidence.

Among all the transitions, some behavioral paths yielded much higher z-scores than 1.96. These paths include MCC → TH and INT → ASE in the high-performing group, and RGW→ RPW, INT → RPW in low-performing group. These significant transition paths indicate the main differences between the two groups. High-performing group members tend to add supporting evidence after reading group members' work, in contrast, low-performing group members tend to read group members' work after they add supporting evidence. The results suggested that high-performing group members concentrated on completing, monitoring and improving their group argumentation graph. Besides, low-performing group members tended to refine and revise their previous work after reading group work and searching internet information, while high-performing group members did less refinement and revision.

5. Discussion and conclusion

This study analyzed the interaction patterns of students' collaborative argumentation behaviors in small groups. Firstly, it was found that high-performing group members have more frequent social interactions than low performing group members. This echoed the previous research studies that cognitive processes needs social interaction to take place as only then can new ideas, critical comments, other viewpoints, feedback, shared understanding be exchanged (Kreijns et al., 2003). In addition, high-performing group members tended to have more diverse interactive patterns than low-performing group members.

Secondly, high-performing group members seek input from group members before they work on improving their argumentation. In contrast, low-performing group members put forth their ideas before they seek input from others. Thirdly, unlike the students in high-performing groups, members in low-performing group tended to revise their work based on group members and online information. The high-performing group members concentrated on within-group ideation in collaborative argumentation activities by frequent interaction and monitoring the group work. Compared with the high-performing group members,the low-performing group members focused less about the whole group work, including organizing group argumentation graph, monitoring the argumentation structure. A possible explanation was that students from low-performing group lacked the ability to build on peers' ideas. Previous studies have shown that students do not spontaneously build on peers' ideas, which is considered to play an essential role in collaborative learning (Molinari et al., 2008).

The findings from this study provide further evidence of how students from high-performing group and low-performing groups interact and monitor in collaborative argumentation activities within their groups. The findings have implications for teaching and facilitating collaborative learning. First, teachers should encourage active idea exchange within group members, instead of only building independent work. Second, teachers utilise learning analytics generated by the system to encourage and enable students on monitoring and regulating inter-and intra-group interactions. Extra scaffolds for the lower performing groups in productive interaction and monitoring are needed. Teachers are suggested to provide intervention when students are not building on each other's work in CSCL. This study also has limitations. The sample size was small as this study was exploratory in nature. In addition, the collaboration setting was a specific graph-based collaborative argumentation platform. Bigger-scale studies in wider CSCL contexts are needed to explore the interaction mechanism of small group collaborative learning.

6. References

Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 199-214.

Andriessen, J. (2006). Arguing to learn. K. Sawyer (Ed.) Handbook of the Learning Sciences (pp. 443-459). In: Cambridge: Cambridge University Press.

Bakeman, R., & Gottman, J. M. (1997). *Observing interaction: An introduction to sequential analysis*. Cambridge university press.

Chen, W., Tan, J. S., & Pi, Z. (2021). The spiral model of collaborative knowledge improvement: an exploratory study of a networked collaborative classroom. *International Journal of Computer-Supported Collaborative Learning*, *16*(1), 7-35.

Chen, W., Zhang, S., Wen, Y., Looi, C.K. & Yeo, J. (2019). A Spiral Model of Collaborative Knowledge Improvement to Support Collaborative Argumentation for Science Learning: Technological and Pedagogical Design. In *13th International Conference on Computer Supported Collaborative Learning (CSCL)* (pp. 240-247). Lyon, Singapore: École Normal Supérieure de Lyon.

Connor, U. (1990). Linguistic/rhetorical measures for international persuasive student writing. *Research in the Teaching of English*, 67-87.

Curtis, D. D., & Lawson, M. J. (2001). Exploring collaborative online learning. *Journal of Asynchronous learning networks*, *5*(1), 21-34.

Dillenbourg, P. (1999). What do you mean by collaborative learning? In: Citeseer.

Fan, Y., Wang, T., & Wang, K. (2020). Studying the effectiveness of an online argumentation model for improving undergraduate students' argumentation ability. *Journal of Computer Assisted Learning, 36*(4), 526–539. https://doi.org/10.1111/jcal.12420

Furberg, A., & Ludvigsen, S. (2008). Students' meaning-making of socio-scientific issues in computer mediated settings: exploring learning through interaction trajectories. *International Journal of Science Education*, *30*(13), 1775-1799.

Hou, H.-T., & Wu, S.-Y. (2011). Analyzing the social knowledge construction behavioral patterns of an online synchronous collaborative discussion instructional activity using an instant messaging tool: A case study. *Computers & Education*, *57*(2), 1459-1468.

Kang, L., Liu, Z., Su, Z., Li, Q., & Liu, S. (2018). Analyzing the relationship among learners' social characteristics, sentiments in a course forum and learning outcomes. 2018 seventh international conference of educational innovation through technology (EITT),

Kapur, M., Voiklis, J., & Kinzer, C. K. (2011). A complexity-grounded model for the emergence of convergence in CSCL groups. In *Analyzing interactions in CSCL* (pp. 3-23). Springer.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfalls for social interaction in computer-supported collaborative learning environments: a review of the research. *Computers in Human Behavior*, *19*(3), 335-353.

Lee, A., O'Donnell, A. M., & Rogat, T. K. (2015). Exploration of the cognitive regulatory sub-processes employed by groups characterized by socially shared and other-regulation in a CSCL context. *Computers in Human Behavior*, *52*, 617-627.

Lizzio, A., Wilson, K., & Simons, R. (2002). University students' perceptions of the learning environment and academic outcomes: implications for theory and practice. *Studies in higher education*, *27*(1), 27-52.

Malmberg, J., Järvelä, S., Järvenoja, H., & Panadero, E. (2015). Promoting socially shared regulation of learning in CSCL: Progress of socially shared regulation among high-and low-performing groups. *Computers in Human Behavior*, *52*, 562-572.

Molinari, G., Sangin, M., Nüssli, M.-A., & Dillenbourg, P. (2008). Effects of knowledge interdependence with the partner on visual and action transactivity in collaborative concept mapping.

Popov, V., Biemans, H. J., Fortuin, K. P., van Vliet, A. J., Erkens, G., Mulder, M., Jaspers, J., & Li, Y. (2019). Effects of an interculturally enriched collaboration script on student attitudes, behavior, and learning performance in a CSCL environment. *Learning, Culture and Social Interaction*, *21*, 100-123.

Resnick, L. B., Salmon, M., Zeitz, C. M., Wathen, S. H., & Holowchak, M. (1993). Reasoning in conversation. *Cognition and instruction*, *11*(3-4), 347-364.

Slof, B., Erkens, G., Kirschner, P. A., Jaspers, J. G., & Janssen, J. (2010). Guiding students' online complex learning-task behavior through representational scripting. *Computers in Human Behavior*, *26*(5), 927-939.

Stegmann, K., Weinberger, A., & Fischer, F. (2007). Facilitating argumentative knowledge construction with computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, *2*(4), 421-447.

Strijbos, J.-W. (2010). Assessment of (computer-supported) collaborative learning. *IEEE transactions on learning technologies*, *4*(1), 59-73.

Tan, S. H. J., & Chen, W. (2022). Peer feedback to support collaborative knowledge improvement: What kind of feedback feed-forward? Computers & Education *187,* https://doi.org/10.1016/j.compedu.2022.104467

Toulmin S. (1958). *The Uses of Argument*. Cambridge University Press: Cambridge.

Vuopala, E., Hyvönen, P., & Järvelä, S. (2016). Interaction forms in successful collaborative learning in virtual learning environments. *Active Learning in Higher Education*, *17*(1), 25-38.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Effects of Group Awareness Tools on Student Engagement and Enjoyment in Online Collaborative Writing

Xiaohui SONG[a], You SU[a*], Jing REN[a]& Rushi GONG[a]
[a]*Beijing University of Posts and Telecommunications, China*
*suyou@bupt.edu.cn

Abstract: Group awareness tools have the potential to promote student engagement and improve the quality of online collaborative learning. However, few studies have explored the effects of group awareness tools on second language learners' engagement and enjoyment in collaborative writing. In this pilot study, we adopted a quasi-experimental design to examine whether the use of group awareness tools led to increased student engagement and enjoyment in online collaborative writing of English argumentative essays. The results revealed that the students in the experimental condition (with a group awareness tool) demonstrated higher behavioral engagement than the students in the control condition (without a group awareness tool), while no statistically significant difference was found in cognitive engagement and social engagement. Additionally, compared with those collaborated with the traditional method, the students using group awareness tools generally indicated higher enjoyment but the difference was only significant in the dimension of student support. Some discussions were made and suggestions for future studies were also provided.

Keywords: Group awareness, student engagement, enjoyment, online collaborative writing

1. Introduction

Online collaborative writing, as an effective instructional task, has been increasingly implemented in the field of second language education (Abe, 2021; Such, 2021; Yang, 2018). Online collaborative writing refers to pair or group writing tasks conducted online in which students interact, negotiate meaning, co-construct text, co-revise texts and jointly produce a writing text facilitated by technology tools (Li, 2018). A large body of research has documented the benefits of online collaborative writing such as raising audience awareness (Lee, 2010), enhancing writing skills (Bikowski & Vithanage, 2016), and promoting writing qualities in accuracy and complexity (Wang, 2015). However, previous studies have also identified that unequal participation or social loafing remains as a challenge that hinders effective collaboration and reduces students' perceived enjoyment of collaborative writing activities (e.g., Ducate, Anderson, & Moreno, 2011).

Researchers started to examine learning groups' participation processes in collaborative learning from the perspective of student engagement (Xu, Chen, & Chen, 2020). According to Fredricks, Blumenfeld, and Paris (2004), student engagement is a multi-dimensional construct which includes factors like behavioral engagement, cognitive engagement, and social engagement. Behavioral engagement refers to the participation and involvement of activities ranged from scholarly activities to social or extracurricular activities (Fredricks, Blumenfeld, & Paris, 2004). Cognitive engagement is the commitment to study on the psychological level (Fredricks, Blumenfeld, & Paris, 2004). Social engagement draws on the quality of social communication with teachers and students in academic tasks and school settings (Hoi & Hang, 2021).

Group awareness tools have been developed to solve the problem of low engagement in online collaborative learning settings (e.g., Peng, Li, Su, Chen, & Jiang, 2022). According to Bodemer, Janssen, & Schnaubert (2018), group awareness refers to perception or understanding of characteristics of learning partners or the collaborating group. The provision of group awareness information also serves as an implicit guidance for collaborative learning processes (Janssen & Bodemer, 2013; Schnaubert & Bodemer, 2019). Researchers have pointed out that being mutually

aware of contribution rates can stimulate motivational processes because students tend to avoid negative social evaluations by the group (Janssen, Erkens, & Kirschner, 2011). Previous studies have developed various group awareness tools and found that these tools can enhance student's engagement in the process of online collaborative writing (e.g., Janssen, Erkens, Kanselaar, & Jaspers, 2007; Janssen, Erkens, & Kirschner, 2011). For instance, to better support group writing processes, Calvo and his team (2011) developed a tool called Glosser that enables a team to understand how each member is participating in the collaborative and use this information to improve their collaboration. In addition, some emerging studies have also documented the benefits of group awareness tool for improving students' attitudes and perceived enjoyment toward collaborative learning in the CSCL environment (Yilmaz & Yilmaz, 2020). However, few studies have examined the effect of group awareness tools on student engagement and enjoyment in language education settings. Technology enhanced collaborative writing has been increasingly used in second language teaching practices but this instructional method has also encountered problems such as low engagement and social loafing (Li, 2018). This study aims to explore whether the incorporation of group awareness tools in collaborative writing can enhance student engagement and enjoyment in the context of online collaborative writing of English argumentative essays.

2. Research Questions

To address these research gaps, we designed an online collaborative writing platform named *CollaWrite* which was embedded with group awareness tools to explore the effect of group awareness on student engagement and enjoyment in an online collaborative essay writing task. The two research questions guiding this study were presented as follows:

1. Did the students supported with group awareness tools show greater student engagement than those without the support of group awareness tools in online collaborative English essay writing?
2. Did the students supported with group awareness tools perceive more enjoyment than those without the support of group awareness tools in online collaborative English essay writing?

3. Methods

3.1 Context and Participants

This study was carried out in a college English course that aimed to improve students' English writing abilities at a key university in northern China. This course used an online collaborative writing task that lasted three weeks. The task required students to work in small groups to jointly write an argumentative essay on the same topic. Following Storch's (2013) guidelines for collaborative writing, participants within a group were required to interact with each other throughout the whole writing process and contributed to the planning, idea generations, editing, and revision. The writing task took the process writing approach and each group should complete three-round of writing. For each round of writing, the essay was randomly assigned to other groups for peer assessment and feedback. The whole collaborative writing task was conducted on an online platform which could offer a collaborative writing service and a synchronous chat room to help students communicate with their peers. In addition, group awareness information (e.g., the number of posts per student) was also provided by this platform.

Participants were 42 freshmen majoring in telecommunications and modern postal service. Since small group writing activities that consists of four or five members are found to be effective and are widely used in second or language education practice (e.g., Bikowski & Vithanage, 2016; Lee, 2010), the instructor in this study also randomly assigned these students into 10 groups of four or five students. Five groups were appointed as the experimental group supported with group awareness information while five groups were set as the control group without group awareness information support correspondingly. They took an English proficiency test before the experiment, and results showed no significant difference between the experimental and the control groups ($t=1.011$, $p=0.319>0.05$).

3.2 System Description

CollaWrite is a platform we developed that consists of three functional areas: collaborative writing section, chat room, and group awareness visualization. The group awareness tools used in this study provided four graphics on groups' behavioral engagement, a radar map of peer assessment scores, and a word cloud (see Figure 1). Firstly, group awareness tools on students' behaviors in the writing process can inform each group of its members' participation (the number of words written by per student), activeness (the number of logins per student), communication (the number of posts per student) and interpersonal relations. Instructor and group members usually use these visualized graphics to identify the individual contribution to the writing task. Next, scores of peer assessment are presented visually as a radar map from five cognitive dimensions: vocabulary, mechanics, content, grammar, and cohesion, aiming to assist group members in recognizing the shortcomings of their essay. In addition, the word cloud indicates the frequency of key words used in collaborative writing. With the help of word cloud, group members can easily observe what key words and key points the whole class was writing.

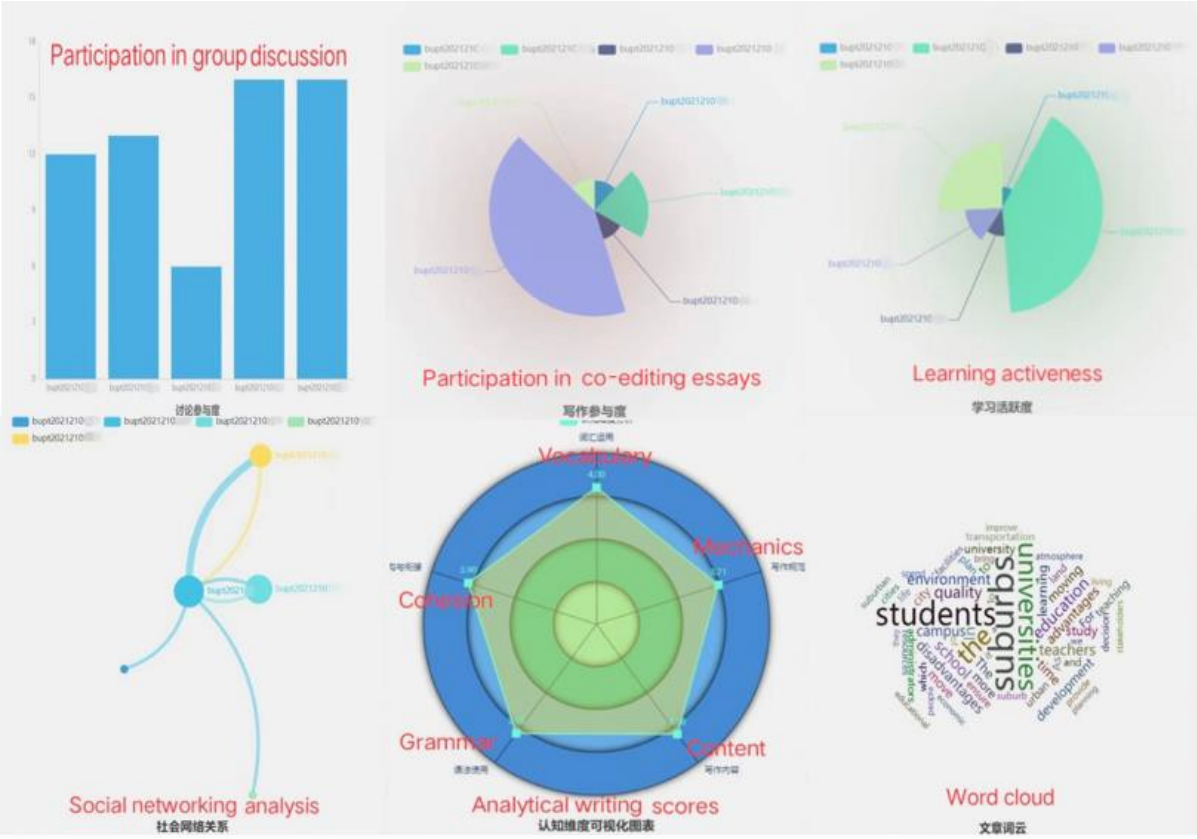


Figure 1. Group awareness tools

3.3 Data Collection and Analysis

According to this study, we adapted the Online Student Engagement Questionnaire (Hoi & Hang, 2021) and English Classroom Enjoyment Scale (Jin & Zhang, 2021) to measure student engagement and enjoyment in online collaborative writing with some modifications of the original instrument. The final version of instrument for assessing student engagement and enjoyment in this research included 32 items and was divided into 5 factors: behavioral engagement (7 items), cognitive engagement (5 items), social engagement (6 items), enjoyment of collaborative writing (4 items) and enjoyment of student support (4 items). Data was collected through a 5-point Likert survey scale ranging from 1 (strongly disagree) to 5 (strongly agree) with an overall Cronbach' s alpha value of 0.861 at the end of online collaborative writing task. The instructor sent this questionnaire to all participants, and altogether 37 valid responses (experimental group=18, control group=19) were received.

We first calculated means and standard deviations of student engagement and enjoyment in the experimental group and the control group. Then, a Shapiro-Wilk test used to explore whether these data obeyed a normal distribution. As the data violated the normal distribution, the present study employed Mann-Whitney U-test to compare student engagement and enjoyment between the experimental group and the control group.

4. Results

4.1 Effects of group awareness tools on student engagement

Table 1 displays the mean and standard deviation of behavioral engagement, cognitive engagement, and social engagement of the students in the experimental and control groups. From the descriptive statistical analysis results, we can see that the students supported with group awareness tools had higher behavioral engagement and social engagement while cognitive engagement was slightly lower than the students in the control group. However, the results of Wilcoxon-Mann-Whitney U-test revealed that statistical significance only existed in the dimension of behavioral engagement (z=-2.48, p=0.01).

Table 1. *Descriptive statistics and Mann-Whitney U-test for student engagement*

	Experimental Group		Control Group		*U*	*z*	*p*
	Mean	SD	Mean	SD			
Behavioral Engagement	3.90	0.40	3.48	0.67	90.00	-2.48	0.01*
Cognitive Engagement	3.77	0.52	3.89	0.59	140.50	-0.94	0.35
Social Engagement	3.22	0.70	3.10	0.74	147.50	-0.72	0.47

*p<.05

4.2 Effects of group awareness tools on student enjoyment

The second research question focuses on the effect of group awareness information on student enjoyment. Outcomes of descriptive analysis and Wilcoxon-Mann-Whitney U-test for student enjoyment are showed in Table 2. Results of descriptive analysis indicated that the students provided with group awareness tools showed more enjoyment in the online collaborative writing task. The result of Wilcoxon-Mann-Whitney U-test demonstrated that the difference in enjoyment of student support between the experimental group and the control group was significant (z=-2.51, p=0.01).

Table 2. *Descriptive statistics and Mann-Whitney U-test for student enjoyment*

	Experimental Group		Control Group		*U*	*z*	*p*
	Mean	SD	Mean	SD			
Enjoyment of Collaborative Writing	3.86	0.73	3.68	0.69	156.00	-0.46	0.65
Enjoyment of Student Support	4.35	0.46	3.80	0.71	89.50	-2.51	0.01*

*p<.05

5. Discussion

The purpose of this research is to explore the effect of group awareness tools on student engagement and student enjoyment in online collaborative writing. The results revealed some differences between the experimental group and the control group in student engagement and enjoyment of collaborative learning. Participants in the experimental group demonstrated relatively higher behavioral engagement and enjoyment of student support in the overall process of online collaborative writing.

This study reveals that group awareness tools can promote students' behavioral engagement, which is consistent with the findings of Liu, Liu, & Liu 's (2018). This finding also corroborates with a previous study of Janssen, Erkens, Kanselaar, & Jaspers (2007) in that visualization of group

members' contribution in online collaborative learning (e.g., the number and average length of discussion entries posted by each group member) is beneficial for increasing students' online participation. In addition, Li, Li, Zhang, & Li (2021) also found that students with group awareness information performed more cognitive behaviors such as negotiation of meaning and construction of new knowledge in their discussions. However, this study does not find positive effect of group awareness tools on cognitive engagement and social engagement.

Regarding student enjoyment, those provided with group awareness tools during online collaborative writing appreciate group member' support more than students without group awareness information, which is in line with Rojas et al.'s (2022) finding that group awareness tools can promote students' positive emotions toward online collaborative learning.

There are some limitations in this study. First, the sample size in this pilot study was relatively small. Future studies are suggested to enlarge the sample size to further explore the effect of group awareness tools on students' cognitive engagement, social engagement, and student enjoyment in online collaborative writing. Second, this study adopted a quantitative approach using self-reported questionnaire data. Future research on effects of group awareness tools needs to adopt a mixed method to collect both quantitative and qualitative data through interview, classroom observation or content analysis.

6. Conclusion

Using a quasi-experimental design, the current research examined the effect of group awareness tools on student engagement and enjoyment in online collaborative writing. The findings revealed that group awareness tools could effectively increase students' behavioral engagement and student enjoyment of *student support* in an online collaborative writing task. However, no statistically significant difference was found in cognitive engagement and social engagement. Additionally, compared with those collaborated with the traditional method, the students using group awareness tools generally indicated higher enjoyment but the difference was only significant in the dimension of student support. This study contributes to our understanding of the role of group awareness tools in computer mediated collaborative language learning.

Acknowledgements

This paper was supported by the Humanities and Social Sciences Fund of Chinese Ministry of Education (18YJC740084) and the Teaching Reform Project of Beijing University of Posts and Telecommunications (2021JXYJ19)

References

Abe, M. (2021). L2 interactional competence in asynchronous multiparty text-based communication: study of online collaborative writing. *Computer Assisted Language Learning*, *34*(4), 409–433.

Bikowski, D., & Vithanage, R. (2016). Effects of web-based collaborative writing on individual L2 writing development. *Language Learning & Technology*, *20*(1), 79–99.

Bodemer, D., Janssen, J., & Schnaubert, L. (2018). Group awareness tools for computer-supported collaborative learning. In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), *International Handbook of the Learning Sciences* (pp. 351–358). Routledge.

Calvo, R. A., O'Rourke, S. T., Jones, J., Yacef, K., & Reimann, P. (2011). Collaborative writing support tools on the cloud. *IEEE Transactions on Learning Technologies*, *4*(1), 88-97.

Ducate, L. C., Anderson, L. L., & Moreno, N. (2011). Wading through the world of wikis: An analysis of three wiki projects. *Foreign Language Annals*, *44*(3), 495–524.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research*, *74*(1), 59–109.

Hoi, V. N., & Hang, H. L (2021). The structure of student engagement in online learning: A bi-factor exploratory structural equation modelling approach. *Journal of Computer Assisted Learning*, *37*(4), 1141–1153.

Janssen, J., Erkens, G., Kanselaar, G., & Jaspers, J. (2007). Visualization of participation: Does it contribute to successful computer-supported collaborative learning?. *Computers & Education*, *49*(4), 1037–1065.

Janssen, J., & Bodemer, D. (2013). Coordinated computer-supported collaborative learning: Awareness and awareness tools. *Educational Psychologist*, *48*(1), 40–55.

Janssen, J., Erkens, G., & Kirschner, P. A. (2011). Group awareness tools: It's what you do with it that matters. *Computers in human behavior*, *27*(3), 1046–1058.

Jin, Y., & Zhang, L. J. (2021). The dimensions of foreign language classroom enjoyment and their effect on foreign language achievement. *International Journal of Bilingual Education and Bilingualism, 24*(7), 948–962.

Lee, L. (2010). Exploring Wiki-Mediated Collaborative Writing: A Case Study in an Elementary Spanish Course. *CALICO Journal*, *27*(2), 260–276.

Li, M. (2018). Computer-mediated collaborative writing in L2 contexts: An analysis of empirical research. *Computer Assisted Language Learning*, *31*(8), 882–904.

Li, Y., Li, X., Zhang, Y., & Li, X. (2021). The effects of a group awareness tool on knowledge construction in computer-supported collaborative learning. *British Journal of Educational Technology*, *52*(3), 1178-1196.

Liu, M., Liu, L., & Liu, L. (2018). Group awareness increases student engagement in online collaborative writing. *The Internet and Higher Education*, *38*, 1–8.

Peng, Y., Li, Y., Su, Y., Chen, K., & Jiang, S. (2022). Effects of group awareness tools on students' engagement, performance, and perceptions in online collaborative writing: Intergroup information matters. *The Internet and Higher Education*, *53*, 100845.

Rojas, M., Nussbaum, M., Guerrero, O., Chiuminatto, P., Greiff, S., Del Rio, R., & Alvares, D. (2022). Integrating a collaboration script and group awareness to support group regulation and emotions towards collaborative problem solving. *International Journal of Computer-Supported Collaborative Learning*, *17*, 135–168.

Schnaubert, L., & Bodemer, D. (2019). Providing different types of group awareness information to guide collaborative learning. *International Journal of Computer-Supported Collaborative Learning, 14*(1), 7–51.

Storch, N. (2013). *Collaborative writing in L2 classrooms*. Bristol, UK: Multilingual Matters.

Such, B. (2021). Scaffolding English language learners for online collaborative writing activities. *Interactive Learning Environments*, *29*(3), 473–481.

Wang, Y. C. (2015). Promoting collaborative writing through wikis: A new approach for advancing innovative and active learning in an ESP context. *Computer Assisted Language Learning*, *28*(6), 499–512.

Xu, B., Chen, N. S., & Chen, G. (2020). Effects of teacher role on student engagement in WeChat-Based online discussion learning. *Computers & Education*, *157*, 103956.

Yang, Y. F. (2018). New language knowledge construction through indirect feedback in web-based collaborative writing. *Computer Assisted Language Learning*, *31*(4), 459–480.

Yilmaz, R., & Yilmaz, F. G. K. (2020). Examination of the effectiveness of the task and group awareness support system used for computer-supported collaborative learning. *Educational Technology Research & Development*, *68*, 1355–1380.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Understanding Areas of Parallelograms Through Virtual Geometrical Representations: A Pilot Study

Eva-Maria TERNBLAD
Dept of Cognitive Science, Lund University, Sweden
eva-maria.ternblad@lucs.lu.se

Abstract: Do virtual representations, intended for teaching and learning, afford different constraints and affordances compared to their physical twins, and may this lead to different ways of interaction and understanding? In this line of inquiry, and with the perspective on learning as an extended activity, where learning can be mediated through the interaction with artefacts, the present study has translated the concrete material from a 30-year-old study on geometry to a digital application with virtual geometrical manipulatives. 26 students from the 6th grade worked with the software for two lessons and thereafter completed a test. Learning outcomes are presented and compared to those from the original study, revealing not only substantially less learning gains, but also other unexpected differences.

Keywords: Concrete and virtual representations, digital and physical interaction, geometry, extended cognition, constraints, affordances

1. Introduction

The digitalization of learning environments is one of the major pedagogical shifts in modern education, and today more than 90% of the elementary schools in Sweden use digital devices as an integrated part of their curriculum (Skolverket, 2022). The transition from traditional hands-on experiences to digital and screen-based activities has, during the last decade, led to a series of studies where physical and virtual representations have been evaluated in relation to learning outcomes, revealing very mixed results (Erickson, 2015; Rau, 2020; Zierer, 2019). Here, a variety of theoretical paradigms has been applied, such as theories about learning and motivation/engagement, embodied cognition, cognitive load theory, and theories on visual perception and attention (Rau, 2020). However, another way of contrasting virtual and physical representations would be to focus on their intrinsic qualities, and how these – through interaction – may help reshaping the learner's cognitive state. We might, for instance, evaluate if and how a student actually makes use of a representation, and how different operations (due to constraints and affordances in the material) might mediate the student's understanding of a specific concept. Additionally, we could investigate whether the students' interpretation of a representation is adequate or not. Such an inquiry, borrowing theories and concepts from situated and extended cognition (Clark, 2014; Sawyer & Greeno, 2009), would place the properties of the teaching material – together with the learner's observable actions – in the center, instead of focusing on the cognitive and/or neural processes within the learner's head. Given the difficulty with separating such inner processes from one another in a rich naturalistic learning situation, this approach may be a valuable starting point for future research on cognition, interaction, and learning.

Subsequently, the goal of the present study was to evaluate if two versions (one digital and one physical) of a teaching material – intended to mediate the understanding of areas of parallelograms – would lead to different interactions and insights, and subsequently, to different learning outcomes.

2. Method

As a baseline and control, a thirty-year-old study, investigating the effects of a geometrical teaching material, was utilized (Sayeki, Ueno & Nagasaka, 1991). In this study, students worked with two manipulatives: a deck of cards and a paper frame, reshaping them into different parallel figures (see Figure 1, left). The cards were also put inside the frame, and by tilting the frame and cards together the students experienced how the frontal area of the frame decreased until it reached the frontal area of the deck of cards, learning that the formula for the area of a parallelogram was the same as for the area of a rectangle (that is, the base times the height). This teaching method was concluded to be very effective, leading to a deep understanding of areas of parallelograms with different shapes (ibid.).

Since no existing software was found to be appropriate for these exercises, we decided to translate the physical manipulatives described above into virtual ones and put them into a learning application, containing a two-dimensional virtual deck of cards (where the cards were moved sideways), and a virtual frame with fixed sides but movable corners that could be tilted (see Figure 1, middle and right). The application now and then posed questions about the base, height, and area of the different shapes, making the students aware of constant and variable properties of the material.

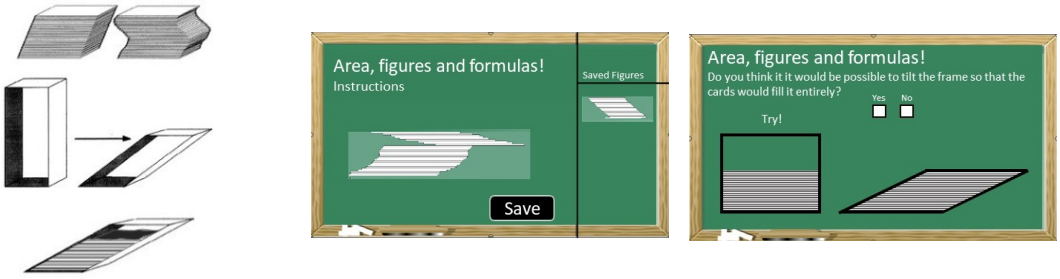


Figure 1. Sayeki and colleagues' (1991) material for learning about areas of parallelograms (left) and the interface for the digital version created for this particular study (middle and right)

The software was then piloted by letting 26 6th-grade Swedish students (aged 11-12 years) use it for two 20-30-minute lessons. The students worked in pairs, sharing one computer, and the experiment was conducted in small groups with four students at a time. This experimental setup made it possible to gain access to the students' thoughts and perceptions through their dyadic conversations (which were recorded together with screen activities). Half a week after the second lesson, the students took a post-test (consisting of multiple-choice questions with suggestions for area formulas for different shapes) and another half week later, the students' conceptual knowledge were explored through discussions with the experimental leader. To control for students' different levels of prior knowledge in math, their teacher assessed their general mathematical skills, resulting in two student groups: Higher-achieving (N=14), and lower-achieving (N=12).

3. Results and discussion

One central question regarding the virtual manipulatives was whether the students would interact with them in the intended – beneficial – way. When analyzing the shapes created with the virtual deck of cards it turned out that the students often created strange spidery figures, quite different from classic parallelograms or more block-like shapes. This was not at all symptomatic for the Sayeki et al's (1991) study. It was primarily lower-achieving students who made these spidery shapes, although higher-achieving students made more figures in total (see Table 1).

When the students interacted with the frame and cards together and were asked if they thought it would be possible to tilt the frame so that the cards would fill it, 86% of the higher-achieving students, but only 45% of the lower-achieving ones, answered 'yes'. All students completed the tilt, and several of them were surprised when seeing the cards and the frame merging into the same area. When it came to answering questions about the figures' base, height and area, the most difficult questions were those regarding the area and height of the frame, and many students confused the height with the length of

the frame's tilted side. The difficulty with grasping the concept of height was also confirmed during the final oral discussion.

Table 1. *Average number of shapes created with the cards by higher- and lower achieving students.*

					Total No of shapes
Higher achieving	0.9	1.9	1.5	1.6	5.9
Lower achieving	1.0	1.0	0.7	1.8	4.5
Sayeki et al (1991)	Yes	Yes	Yes	No	

Results on the post-test reveal that the most difficult area to calculate was the one for a classic parallelogram (Figure 2 in Table 2). Only 50% of the higher-achieving students succeeded on this task – a result quite different from Sayeki et al's (1991) study, where 94% of the students were successful. Areas for other parallel figures (Figure 3 to 5 in Table 2) seemed easier to calculate, although higher-achieving students tended to overthink the formula for the staircase-like shape (Figure 5 in Table 2) by tripling its total area.

Table 2. *Proportion of correct answers on the post-test for higher- and lower- achieving students.*

	Fig 1 	Fig 2 	Fig 3 	Fig 4 	Fig 5 
Higher achieving	1.00	0.50	0.79	0.64	0.57
Lower achieving	1.00	0.27	0.45	0.64	0.73
Sayeki et al (1991)	1.00	0.94		0.91	1.00

There are several potential reasons for the relatively modest learning outcomes in the present study compared to the original one. First of all, the two experiments are not identical, since the students in this study only performed two shorter lessons, while in the original study, they performed 5 lessons and were guided by a teacher. However, this does not explain the differences *within* each experiment, indicating that the classic parallelogram (Figure 2 in Table 2) was the most difficult to calculate in the present study with virtual manipulatives – but not in the original study with physical manipulatives. Notably, parallel figures with non-straight sides (such as spidery ones) were also often created in the virtual, but not in the physical, condition (see Table 1), although the effect of this interaction isn't clear. But the students also had difficulties with interpreting the properties of the virtual frame, possibly due to the level of abstraction. Consequently, one interpretation of the low proportion of correct answers for Figure 2 in Table 2 in the present study is that this figure was interpreted as a frame – ending up with the wrong formula (*area = base* x *length of side*), while the other parallel figures were interpreted as decks of cards – and thereby more often associated with the correct formula (*area = base* x *height*).

In the current era of educational digitalization, the risk for insufficient or inappropriate interaction with different types of manipulatives, together with the potential difficulty to interpret their intrinsic qualities, have to be considered important to investigate – and especially so if the students are young and unexperienced. This might be done by comparing the use of interchangeable virtual/physical representations in controlled but ecologically valid learning environments.

References

Clark, A. (2014). *Mindware: An Introduction to the Philosophy of Cognitive Science.* New York: Oxford University Press.

Erickson, J. J. (2015). To play or to learn? A review of game-based math learning for motivation and cognition. *Int. J. Cyber Behav. Psychol. Learn. 5*(1), 56–74.

Rau, M. A. (2020). Comparing Multiple Theories about Learning with Physical and Virtual Representations: Conflicting or Complementary Effects? *Educational Psychology Review, 32*, 297–325

Sawyer, R. K. & Greeno, J. (2009). Situativity and Learning. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (p. 55-77). Cambridge: Cambridge University Press.
Sayeki, Y, Ueno, N., Nagasaka, T. (1991). Mediation as a generative model for obtaining an area. *Learning and Instruction, 1*(3), 229-242.
Skolverket (2022). *Skolverkets uppföljning av digitaliseringsstrategin 2021.* Skolverket, Stockholm.
Zierer, K. (2019). *Putting Learning Before Technology! The Possibilities and Limits of Digitalization*. New York, NY: Routledge

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Understanding Sociomaterial Encounters in Collaborative Creative Making: A Case Study from a Makerspace in India

Alekh V* & Chandan DASGUPTA
Indian Institute of Technology, Bombay, India
*alekhv@iitb.ac.in

Abstract: Makerspaces are explored as sites for fostering collaboration and creativity, where makers share ideas, use material resources, and create meaningful artefacts. The article aims to contribute to the understanding of sociomaterial entanglements in making, which can inform the design of socio-technical spaces that promote creative outcomes. In this study, we examine how social and material encounters influence the emergence and evolution of ideas in collaborative making when novice makers engage in design problem-solving. We specifically look at data from a maker activity-centered workshop to argue the distributive nature of creative making. Our analysis suggests that social and material elements of the situated context exhibit performative roles in enabling and constraining collaborative making.

Keywords: Making, Distributed Creativity, Makerspaces, Materiality

1. Introduction

Collaborative making activities are featured by the iterative nature, encountering failures, gathering feedback from fellow makers and experts, and sometimes culminating with unexpected outcomes. Making practices bring out skills in tackling uncommon problems and constructively engaging in design (Petrich et al., 2013). Previous research studies (Bevan, 2017; Patton & Knochel, 2017; Timotheou & Ioannou, 2019) have pointed out the need to include elements of maker culture in education settings as the approach nurtures co-creation, innovative design processes, and collaboration. Hands-on makerspace environments promote maker engagement with technology, afford knowledge sharing, and encourage collaborative creative production (Halverson & Sheridan, 2014; Pandey & Srivastava, 2016). In these collaborative spaces, creativity cannot be reduced to individualistic intellect, skills, or mental processes but embedded in collaborative networks (Sawyer & Dezutter, 2009) and dialectical relation between makers and materials in social practices (Tangaard, 2013). As per this viewpoint, creative thinking is *distributed* across the interaction between people, dialog between people and objects, or interactions over time (Glăveanu, 2014; Sawyer & Dezutter, 2009). Placing design as a conversation with the situation (Schön, 1983) expands the opportunity to understand the interplay of social and material aspects in making, where makers confer with fellow makers and materials, leading to creative design prospects (Shotter, 2006). The concept of agency can be used to understand how materials 'talk back' (Schön, 1983) and influence the making processes.

Associating agency with nonhuman entities brings in the queries related to intelligence, intentionality, and purposefulness. Such divisions can be resolved with Latour's conceptualization of agency (1996), where all material entities are possible actants and actively play roles in everyday activities. Consequently, agency is not an attribute of either human or nonhuman elements but emergent manifestations and effects through certain configurations of situated entities (Suchman, 1987). Agency unfolds in practice (Pickering, 1993) and shifts depending on the assemblages of these entities. Recent works in understanding the role of materials in interaction design conceive humans and nonhuman entities as performative agents, driving the activities in design processes via interplays (Jacucci & Wagner, 2003, 2007). These reframed views on agency may expand our understanding of how creativity surfaces in maker and material conversations.

Drawing from all these prior literature, we look into the broader research goal of how maker and material entanglements contribute towards the distributed form of creativity in makerspaces. In this paper, we investigate how sociomaterial encounters influence the emergence and evolution of ideas and designs in collaborative making, when novice makers get involved in design problem-solving. The makerspace setting considered for this study is that of a one-day maker activity-centered workshop rooted in solving design challenges attended by undergraduate mechanical engineering students. Based on the perspective of design as an inherent practice within the making process (Halverson & Sheridan, 2014; Dougherty, 2012), we consider the participants as novice makers and the process they engage in as the making process. We position the creative aspects of the making process as distributed and materially grounded phenomena (Hutchins, 1995; Glăveanu, 2014; Sawyer & Dezutter, 2009).

2. Study Details

2.1 Setting

Data for this study was collected as part of maker workshop series organized at a leading engineering institute in India. These workshops aimed to engage engineering undergraduates with Maker tools and technologies for solving design challenges situated in a typical makerspace environment. The participants of this study were second-year mechanical engineering undergraduate students who successfully completed an earlier introductory maker workshop based on Lego Mindstorms EV3 robotic kit. The workshop organizers invited these participants, to which eight students showed interest, and a total of four teams were formed in the order of their response, with each team consisting of two members. Participants were introduced to digital fabrication techniques, along with a brush-up session on Lego Mindstorms EV3 kit, as part of this workshop. All the participants were familiar with 3D modeling software like Tinkercad and Solidworks. Also, none of the participants were aware of any ideation techniques. This one-day workshop started with introductory sessions on digital fabrication involving hands-on activities with a desktop FDM 3D printer and a 3D printer pen. Later on, participants were given a design challenge: *Conceptualize an assembly line/production line that is semi-automated with static and dynamic robots. Model and build the setup with resources available in the makerspace.* The making sessions lasted for a total of five hours per team. A facilitator was present during the making sessions to support the teams in their collaborative making processes by organizing resources, such as tools and materials, and by providing conceptual and procedural expertise in events where participants were unable to determine how to proceed in solving the design challenge. This study focused on one of the student teams – Team A, consisting of one female student (G) and one male student (B).

The workshop was conducted at a makerspace with a work table, a whiteboard, a computer, and a desktop FDM 3D printer. Each team was provided with two Lego Mindstorms EV3 kits, a 3D printer pen prototyping, 3D printing filaments, prototyping kits consisting of basic hand tools (knives, scissors, screwdrivers), cardboards marking pens, pencils, chopsticks, play-doh modeling compounds, cable ties, tapes, glue, styrofoam sheets, a box consisting of used cables, ropes, wires, paper, defective electronic devices like calculators and earphones.

2.2 Analysis

We conducted the analysis in multiple phases to explore the collaborative processes and assess the outcomes of the creative making in the makerspace. We used the SVS metrics (Shah, Smith, & Vargas-Hernandez, 2003) to assess the outcomes of the making activity. The SVS metrics acknowledge every idea that makers bring into the design stages as meaningful. As per the metrics, creative outcomes are assessed by evaluating the quantity as a measure of total number of ideas, variety as a measure of the explored solution space, quality as a measure of how close an idea satisfies the design specifications, and novelty as a measure of the infrequency of an idea compared to other ideas. As the participants in our study generated one type of idea (building robots for package movement), and since the activity did not constrain to strict design specifications, we only focused on measuring the novelty component, in line with other studies that used certain components of the SVS metrics, according to

the design task (Starkey et .al, 2016; Sinha et. al, 2017). Here, novelty is measured as the infrequency of an idea compared to all the ideas present. Following the SVS approach, idea features were identified and a genealogical tree was developed with the following levels: goal, working principle, functional unit, transfer, motion, and material details. For example, the transfer level of the tree covered the package transfer mechanism idea features incorporated in various functional units of the production line. Each idea was decomposed into the addressed features and grouped to arrive at the genealogical feature tree. The ideas were then independently rated by the author and a fellow researcher, and achieved high inter-rater reliability, Cohen's kappa = 0.893. For the idea *i,* the following equation was used to calculate the feature novelty f_{ij}, where T is the total number of ideas and C_{ij} is the total number of ideas in which the feature j is addressed:

$$f_{ij} = \frac{T - C_{ij}}{T}$$

The novelty score N_i, for each idea was calculated as the ratio of the sum of the feature novelties to the number of features addressed for each idea, where n is the number of features, z_i is 1 if the feature i is addressed by the idea, 0 if it is not.

$$N_i = \frac{\sum_{j=1}^{n} f_{ij}}{\sum_{j=1}^{n} z_i}$$

Once the novelty score for each idea was determined, we looked at the making process of the highest novelty score team, Team A. We began the analysis by content logging and segmenting the data into various phases of the design process such as ideating, information seeking, sketching, prototyping, testing, and refining ideas, to get an overall understanding of the making process in our context. We follow the case study methodology (Merriam, 2007) to help us unpack and follow Team A's making process. A case study is appropriate for this study due to its exploratory nature as we seek to investigate the influence of social and material elements on the creative making process and to provide an in-depth narrative. We used the content logs to conduct interaction analysis (Jordan & Henderson, 1995), focusing on talk and the use of artifacts and technologies. This analysis method is appropriate for our purpose since we use video data along with transcripts and field notes to unpack the moment-by-moment emergent actions during the group work. The idea units are considered for the unit of analysis where we look at episodes with the idea entry, evolution, and exit from the various making phases. While replaying the video and parsing the transcripts iteratively, we focus on dialogic conversations between participants along with their interaction with available material resources.

3. Findings

As per the SVS metrics (Shah, Smith, & Vargas-Hernandez, 2003), Team A scored the highest novelty score of 0.74 in the ideation phase of the making process. We followed the case of Team A's making activities and found that the team started prototyping based on the idea with the highest novelty score of 0.74, but later dropped the particular idea and then continued with another idea, which consisted of common features as that in other teams' ideas and hence with a lower novelty score of 0.56. We found that the making episodes were constitutively entangled with human and non-human elements of the situated context. We present these episodes to illustrate different instances of how ideas get shaped from primary stages to final stages and how agency emerges from the interactions between the makers and materials that influence the creative making process and outcomes. These moments do not represent the final states in the evolution of the ideas but are temporary outcomes that help refine the design ideas and making process incrementally.

3.1 Task definition

The two makers started the first design session by reading the design brief, consisting of the design challenge description printed on a sheet of paper, and began to think of possible ways of solving the design problem, as they sat at the work table. One of the makers suggested a concept for the assembly line, where a static robot picks up a package from a conveyor belt and places it on a dynamic robot, which can be used to move the package to final delivery stations. The following excerpt shows how the makers and material resources become integral in approaching the problem.

Maker B:	So we have to build two robots.. right? [pointing to design brief]
Facilitator:	Yeah.. you might be familiar with the production lines.
Maker G:	Ok.. then we need to make two robots to pick some things and place at other places
Maker B:	What could be a static robot?.. something kind of conveyor belt thing?
Maker G:	[moves arm with elbow fixed on the table] something like this.. picking things and placing.
Facilitator:	Yeah, kind of..
Maker B:	Then.. we will use the static robot to pick from the belt and..
Maker G:	Yeah, start with the static robot.
Maker B:	.. and put the thing on the dynamic robot.. and it moves.

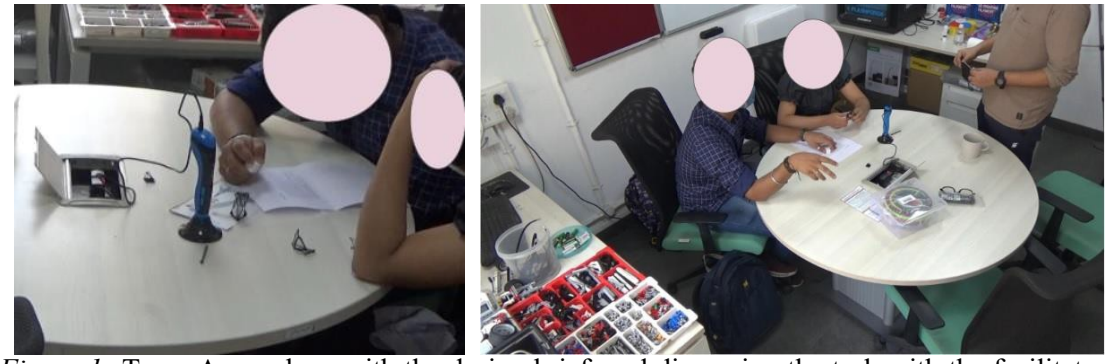


Figure 1. Team A members with the design brief, and discussing the task with the facilitator.

Here the design brief is attended to by the makers and duly considered as a source for defining the task (see Figure 1). The open-ended nature of the problem in the design brief leads the makers to come up with task interpretations. The makers then consult with the facilitator to get more information about the task and develop a broad conceptualization of the assembly line with a set of imaginary conveyor belt and delivery stations. As a material resource, the design brief comes into action as the makers attend to and perform the act of setting boundaries and design possibilities of the task. Similarly, the idea exchanges between the makers and the facilitator help them arrive at a common ground for further progress.

3.2 Design Ideations

As the session progressed, the team started ideating designs for the static robot and dynamic robot according to the proposed solution approach. The makers are found to make use of the whiteboard and markers for sketching their ideas. At every step of the ideation, makers consult with the facilitator for input and to avoid conceptual errors. The following is a representative episode from an ideation phase, showing how a design idea gets shaped in the making context.

Maker B:	[Drawing sketches on the whiteboard] these are poles.. with type of belt kind of thing like that of the printer and the fork can grab the object and

	move in these directions [gestures x,y, z coordinates] and height can be adjusted ..like up and down.. but am not sure how to make it
Facilitator:	Ok.. mm .. for that you will need some guider and rails or some motions like.. you might have seen this.. [takes the slide lock snap-off knife from kit].. this blade can slide over.. similar to telescopic motion.
Maker B:	So more drives.. we have three or four
Maker G:	Or may be we can suspend the fork from top.. with strings
Maker B:	Oh.. then we put strings on these poles and connect to the motors.. it can still work
Facilitator:	Yes.. think of other possibilities too.

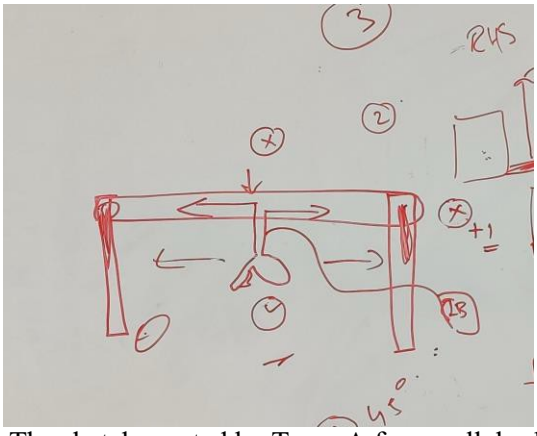


Figure 2. The sketch created by Team A for parallel robot idea.

In the above episode, makers are coming up with ideas for the static robot which can pick and place the incoming package. Maker B starts with the idea of parallel robot, Ni = 0.74, which is to be supported by rectangular frames, and with a hanging 'fork' operated using rope and pulley can pick the package (see Figure 2). The maker points out that the idea is inspired by the 3D printer conveyor movement mechanism. The facilitator pays attention to the idea and questions the mechanism to move the unit up and down, and with the help of a slide lock snap-off knife, explains the idea of telescopic motion, with guide and rails. Here the 3D printer becomes a material source that stimulates the particular idea. On the other hand, the slide lock snap-off knife becomes a source of idea narration and elaboration in the collaborative process. Similarly, the sketches are shared design artefcats, which support the collaborative idea exchange and act as anchors for negotiation.

3.3 Prototyping

As the ideas for building the static and dynamic robots were consolidated, makers started to build on those with available material resources. The team started with the idea of parallel robot, with the rope and pulley mechanism for functioning the fork. Here the makers use a variety of materials including, Lego motors, leg kit parts, chopsticks, earphone wires, play-doh, and tapes. The following excerpt shows how different maker-material configurations impact the externalization of novel ideas in terms of prototypes.

Maker B:	Do we have strings?.. can we use those filaments as strings? [referring to 3D printing filaments]
Facilitator:	Yeah.. you can try. There are some rope pieces in that box [pointing to the box with discarded items]
Maker G:	[Checks the box] there are some wires too..
Maker B:	Aah.. we can use it.. but might slide over.
Maker G:	Or maybe we can cut the ropes.
Maker B:	This earphone .. can we take that.

Facilitator:	Yes, those are junk ones.
Maker G:	[Untangles the earphones]
Maker B:	[Checks the earphones] the lower part looks ok, we can cut that.

In the prototyping session, we observed that the makers trying to blend the lego robotic kit components with mundane materials like chopsticks, earphone wires, play-doh, tapes, etc. The team members start by building the rope and pulley mechanism for the fork. The makers asked for strings (not part of the original kit) to build the mechanism. The makers and facilitator checks the space for the same, but could not find the required kind of thread. But they found a box consisting of used cables, ropes, wires, paper, defective electronic devices like calculators, and earphones, and the facilitator suggested using the rope for the purpose. One of the makers identifies the possibility of using a defective earphone for the same and which is then cut and joined with lego rods for making the pulley assembly (see Figure 3). Similarly, to build the supporting structure, makers used a combination of tape, play-doh, lego motor, and chopsticks so that the fork can follow the up-down motion. Even though makers built an intermediate setup for the same (see Figure 4), it failed during testing as the components fell apart. The team members made further changes by attaching the structure to the pulley using tapes and stabilizing with play-doh, but continued to fail, which stopped them from moving with the particular idea ahead.



Figure 3. The earphone wire – lego components combination for supporting the gripper movement.

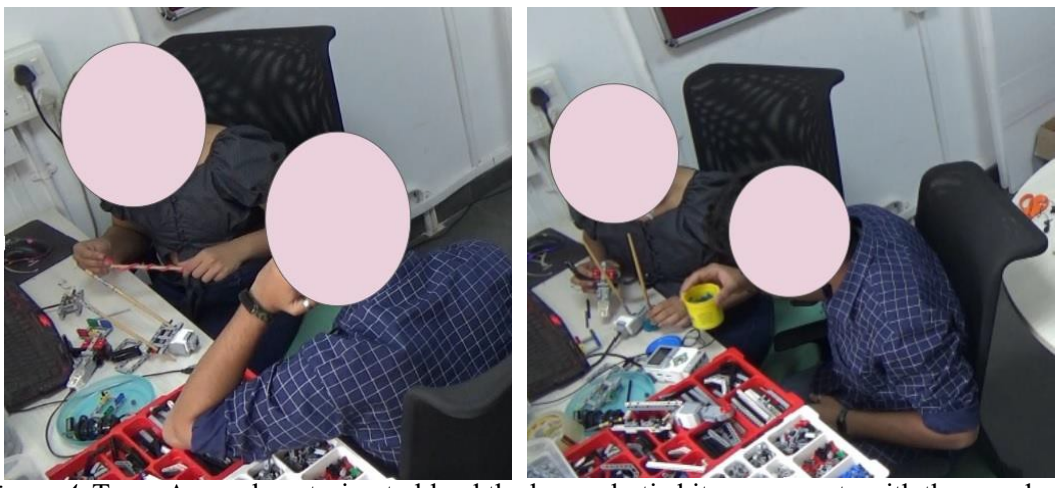


Figure 4. Team A members trying to blend the lego robotic kit components with the mundane materials for building the parallel robot.

Here, we see different modes of maker-material configurations affording and resisting the externalization of ideas in the making process. As for the rope-pulley mechanism, the perceived physical property of the earphone cable is transformed into an emergent quality of a wire-rod combination, performing the acts of fork motions. On the other hand, we find the transformed

combination of lego motors, chopsticks, play-doh etc. resisting the intended actions and forcing the makers to move to alternative solution approaches.

Maker G:	I don't think it will work.. not enough support structure.. drop?
Maker B:	Drop?.. mm .. then we will have to check the crane one.. search in YouTube.. how to build crane.. for picking up
Maker B:	[Pointing search results] got to that.. grab and lift thing
Maker G:	No.. this is something else..
Maker G:	[Opens Lego Mindstorms EV3 Classroom application] there are some models
Maker B:	Hey.. we can do that [points to serial robot model in the manual]
Maker G:	mm.. but look here [points to manual diagrams].. these got some gear systems.
Maker B:	Yeah.. but we can make these.. [picks up and shows gear parts from kit]

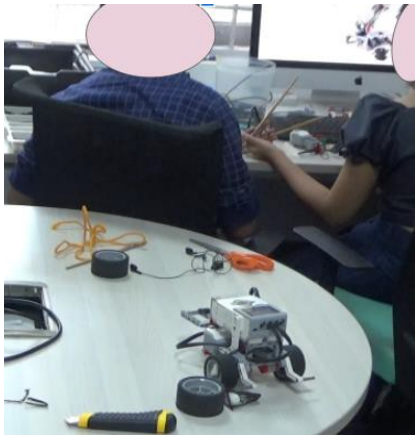


Figure 5. Team A members browsing the Lego Mindstorms EV3 Classroom manual in the computer.

As the idea of the parallel robot seemed to be difficult to realize, the team members chose to look at the other ideas of serial robots. The makers are found to search internet resources for building the serial robot and checked a few YouTube videos. In the dialogic exchanges between the makers, we can see the metaphorical use of cranes to situate the idea of serial robots. The makers agreed to continue with serial robots and, before starting, checked the Lego Mindstorms EV3 Classroom application installed on the computer (see Figure 5). They browsed through the application and found a manual for building a serial robot that can perform pick and place actions. As the makers found the similarities between the sketched serial robot idea, $N_i = 0.56$ (see Figure 6), and that of the one illustrated in the manual, they decided to follow the steps in the manual to complete the robot.

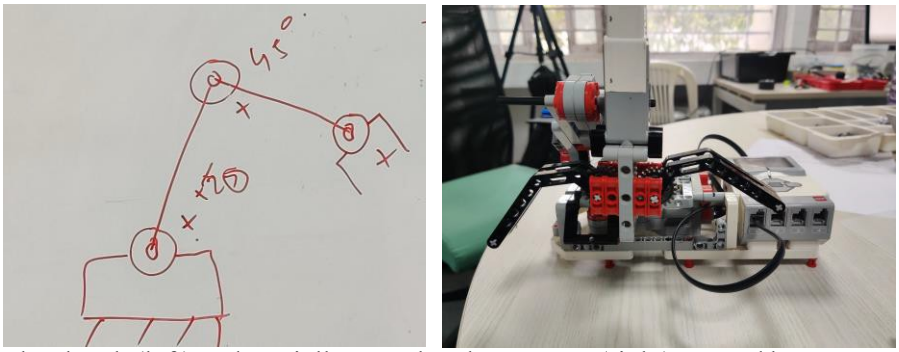


Figure 6. The sketch (left) and partially completed prototype (right) created by Team A for serial robot idea.

Here, the Lego Mindstorms EV3 manual directs the maker actions, that they are following step by step

without checking the useful components of the depicted design in the manual that can be used to address the design problem. As the makers realize that one of the connecting components from the kit is missing, the making process is halted, as one of the robotic links cannot be completed. The problem is reported to the facilitator, who encourages the makers to find a way around it by creating a custom component from the available material resources. The facilitator inspects the prototype and asks a few questions regarding some of the design decisions so that the makers can proceed further without the missing connecting component (see Figure 7).

Facilitator:	Do you really need that extended arm?
Maker B:	Without that… it will not move down, with the gears connected
Facilitator:	Ok.. what if you remove that extra double frames or arm, with that connecting part? [referring to the missing part in the manual]
Maker B:	Aah.. I don't think… it will move
Maker G:	[Points to the manual sketch] not like that.. it can work, we have this motor .. [turns to Maker B] he is asking about that.



Figure 7. Team A members discussing with the facilitator, referring to the manual and partially built serial robot.

4. Discussion & Conclusion

We have presented an empirical case study of making from a makerspace in India, where novice makers engage in solving ill-structured design problems with the resources available. The case demonstrates the ways in which sociomaterial entanglements significantly shape the making paths. With the notion of agency (Latour, 1996), we examined the actions of makers and materials in arriving at creative outcomes as ideas get generated and evolved. The focus on maker-material interaction helps understand how materials play a crucial role in making and how materials "talk back" to makers (Schön, 1983), thus moving beyond the conception of materials as dormant malleable entities. Material resources like the design brief and Lego Mindstorms EV3 manual directed the makers to interpretations and design pathways, the 3D printer acted as a source of idea inspiration, tools like slide lock snap-off knife acted as a source of idea elaboration, and the sketches as anchor points for idea narration and negotiation. In the prototyping sessions, we have seen instances where makers recognize the material properties, which transform into qualities as making progress, and materials are also found to direct maker actions, provide opportunities for idea realization, and offer resistance to the making actions, suggesting the relational and emergent dynamics of agency between artifacts, tools, materials, humans (Pickering, 1993).

Team members and material resources actively contributed to dealing with the design problem, whereas the facilitator supported by providing conceptual and procedural expertise in events of impasses, hence pointing to the distributed form of creative actions (Glăveanu, 2014; Sawyer & Dezutter, 2009). We find that the agency of making gets distributed between makers and materials during the situated interactions. In this case study, we find that the distribution of agency in collaborative making is unpredictable, dynamic, and affects the subsequent creative movements. For example, Lego Mindstorms EV3 manual entry and the blending actions using Lego robotic kit

components with the mundane materials are diverse and unforeseen modes of agency coming into play. We notice that maker-maker interactions, maker-facilitator interactions, and maker-material conversations opened spots for collaborative idea exchange and tinkering, which in turn facilitated and constrained the emergence of novel outcomes. On that account, while using design problem solving as an intervention or in designing learning environments, ample opportunities for social and material encounters are to be ensured to explore the potential of learning by making, especially in dynamic learning environments. With this study, we attempt to understand the emergence of creative outcomes in making by exploring the collaborative processes, maker-material interactions, and assess the outcomes of the creative making using SVS metrics, deviating from the traditional investigations on the effects of individualistic traits on creativity. However, further studies are required to expand our understanding related to the sociomaterial entanglement in creative making contexts.

5. Implications & Limitations

Along with investigating the creative actions and outcomes, the study also shows how making can serve as opportunities for learning through authentic design challenges (Peppler et al., 2016), how materials augment making, and how facilitators can support problem-solving without drastically reducing the wickedness of the design problem. The access to dense and varied material resources, ill-structured problem nature, and less scripted making sessions opened up opportunities to think and make together with materials (Keune & Peppler, 2019). Positioning creativity as a distributed action can provide insight into the distributed mode of learning in makerspace settings. This study sheds light on shared agency in maker-material interaction during making, which calls for careful design and implementation of maker tools and technologies that enact tinkerability and reflective practices (Resnick & Rosenbaum, 2013). Given the sociomaterial lens used for understanding the creative processes in the study, we suggest researchers and practitioners to expand the constructivist views on makerspace education towards the possibilities of post-human approaches (Thiel, 2015; Wohlwend et al., 2017) as it decenters the human roles, and considers other other-than-human entities into account, in learning contexts. These enlarged analyzing lenses can reveal complexities across matter and learners in sociotechnical spaces.

The present data focuses on a typical makerspace setting, where novice makers engage in solving design problems through making. So the findings cannot be generalized across other makerspace settings and sites of collaboration. Future analyses from several making contexts and among diverse makers would give a more refined understanding of the meanings of social and material encounters. It would be also interesting to see how multiple teams work in parallel to solve similar design challenges, sharing the same physical space for making.

References

Bevan, B. (2017). The promise and the promises of making in science education. *Studies in Science Education*, *53*(1), 75-103.

Dougherty, D. (2012). The maker movement. *Innovations*, *7*(3), 11–14

Glăveanu, V. P. (2014). *Distributed creativity: Thinking outside the box of the creative individual*. Cham/Heidelberger: Springer International Publishing.

Gravel, B. E., & Svihla, V. (2021). Fostering heterogeneous engineering through whole-class design work. *Journal of the Learning Sciences*, *30*(2), 279-329.

Halverson, E. R., & Sheridan, K. (2014). The maker movement in education. *Harvard educational review*, *84*(4), 495-504.

Hutchins, E. (1995). *Cognition in the Wild* (No. 1995). MIT press.

Jacucci, G., & Wagner, I. (2003). Supporting collaboration ubiquitously: an augmented learning environment for architecture students. In *ECSCW 2003* (pp. 139-158). Springer, Dordrecht.

Jacucci, G., & Wagner, I. (2007, June). Performative roles of materiality for collective creativity. In *Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition* (pp. 73-82).

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *The journal of the learning sciences*, *4*(1), 39-103.

Keune, A., & Peppler, K. (2019). Materials-to-develop-with: The making of a makerspace. *British journal of educational technology*, *50*(1), 280-293.
Latour, B. (1996). On Interobjectivity. *Mind, culture, and activity*, *3*(4), 228-245.
Merriam, S. B. (2007). *Qualitative research and case study applications in education*. Jossey-Bass.
Pandey, S., & Srivastava, S. (2016). Pop-up maker-spaces: Catalysts for creative participatory culture. In *9th International Conference on Advances in Computer-Human Interactions, Venice* (pp. 50-56).
Patton, R. M., & Knochel, A. D. (2017). Meaningful makers: Stuff, sharing, and connection in STEAM curriculum. *Art Education*, *70*(1), 36-43.
Peppler, K., Halverson, E. R., & Kafai, Y. B. (Eds.). (2016). *Makeology: Makerspaces as learning environments*. Routledge.
Petrich, M., Wilkinson, K., & Bevan, B. (2013). It looks like fun, but are they learning?. In *Design, make, play* (pp. 68-88). Routledge.
Pickering, A. (1993). The mangle of practice: Agency and emergence in the sociology of science. *American journal of sociology*, *99*(3), 559-589.
Resnick, M., & Rosenbaum, E. (2013). Designing for tinkerability. In *Design, make, play* (pp. 181-199). Routledge.
Sawyer, K. R. (2003). Group Creativity. Music. *Theater, Collaboration. New York: Routledge.*
Sawyer, R. K., & DeZutter, S. (2009). Distributed creativity: How collective creations emerge from collaboration. *Psychology of aesthetics, creativity, and the arts, 3*(2), 81.
Schön, D. A. (1983). *The reflective practitioner: How professionals think in action.* Routledge.
Shah, J. J., Smith, S. M., & Vargas-Hernandez, N. (2003). Metrics for measuring ideation effectiveness. *Design studies*, *24*(2), 111-134.
Shotter, J. (2006). Understanding process from within: An argument for 'withness'-thinking. *Organization Studies, 27*(4), 585-604.
Sinha, S., Chen, H. E., Meisel, N. A., & Miller, S. R. (2017, August). Does designing for additive manufacturing help us be more creative? An exploration in engineering design education. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* (Vol. 58158, p. V003T04A014). American Society of Mechanical Engineers.
Starkey, E., Toh, C. A., & Miller, S. R. (2016). Abandoning creativity: The evolution of creative ideas in engineering design course projects. *Design Studies*, *47*, 47-72.
Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.
Tanggaard, L. (2013). The sociomateriality of creativity in everyday life. *Culture & Psychology*, *19*(1), 20-32.
Thiel, J. J. (2015). Vibrant matter: The intra-active role of objects in the construction of young children's literacies. *Literacy Research: Theory, Method, and Practice*, *64*(1), 112-131.
Timotheou, S., & Ioannou, A. (2019). On making, tinkering, coding and play for learning: A review of current research. In *IFIP Conference on Human-Computer Interaction* (pp. 217-232). Springer, Cham.
Wohlwend, K. E., Peppler, K. A., Keune, A., & Thompson, N. (2017). Making sense and nonsense: Comparing mediated discourse and agential realist approaches to materiality in a preschool makerspace. *Journal of Early Childhood Literacy*, *17*(3), 444-462.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Unpacking Emergent Creativity in Online Collaborative Making

Alekh V* **& Chandan DASGUPTA**
Indian Institute of Technology, Bombay, India
*alekhv@iitb.ac.in

Abstract: Making encompasses creative problem solving and artistic self-expression. When makers work in groups by utilizing the available material resources, the creative process can be significantly influenced by social and material interactions. The article aims to contribute to the understanding of creativity as a distributed phenomenon, by discussing an empirical case study of how creative outcomes emerge from collaborative making. We specifically look at data from a graduate course run online, where novice makers collaboratively engage in solving design problems. Our analysis suggests that creativity arising from a collaborative making process cannot be reduced to individualistic traits, but spread across social and material worlds.

Keywords: Making, Collaboration, Distributed Creativity, Collaborative Emergence

1. Introduction

For the last few years, creativity has been discussed in tandem with the Maker movement and Maker culture. In general, making involves creation, collaboration, and sharing, gradually transforming the notion of DIY (Do-it-yourself) to DIT (Do-it-together) (Clapp, 2017). As makers come together in the collaborative making, resources are shared in the situated environment, and influence each other's ideas and design processes, resulting in a collective form of creative outcomes. At the group level, creativity cannot be reduced to individualistic mental processes but distributed across collaborative networks (Sawyer & Dezutter, 2009). As per this perspective, creative processes cannot be considered to be taking place only in isolated minds but also between people, objects, and over time. Such a distributed action considers creativity in terms of interactions, communication, and co-creation (Glăveanu, 2014; Sawyer & Dezutter, 2009). Creativity thus becomes the derivative of the flow of interdependent contextual interventions, and the mechanism of generation within the distributed process can be identified with the concept of collaborative emergence (Sawyer, 2014).

Collaborative emergence refers to the emergence of a new product or creative outcome resulting from collaborating groups that are less constrained (Sawyer & Dezutter, 2009). Collaborative emergence is characterized in group activities as having unpredictable outcomes, with moment-to-moment contingency and the interactional effect of any action can be altered by the subsequent actions of the participants. Collaborative making can also be argued to be distributive in nature, holding the potential for such emergent creative outcomes.

Often, making activities are used as interventions to enhance creativity among learners (Lille & Romero, 2017; Sullivan, 2011), but research studies exploring and documenting collective forms of creativity in making remain scarce (Schmoelz, 2017; Timotheou & Ioannou, 2019). Also, many research approaches have focused only on individual behavior, personality traits, and cognitive process (Kozbelt, Beghetto, & Runco, 2010; Smith & Ward, 2012), leaving out components of interaction with materials. Recent studies have acknowledged the collective and social dimension of creativity. (Aguilar & Turmo, 2019; Clapp, 2017).

In this article, we investigate how creative outcomes emerge out of collaborative making in an online setting, by conceptualizing creativity as distributed and materially-grounded activity, endemic to everyday life (Tangaard, 2013). The setting is that of a graduate course run online, where students work in groups and use maker tools and technologies to solve design problems collaboratively. We consider design as an integral part of the making process (Barniskis, 2014; Dougherty, 2012), and the students taking part in the course as novice makers.

2. Study Details

2.1 Setting

The data for the study is taken from the onscreen video recordings of a graduate course -Tools Lab- run by a leading engineering institute in India, where a series of maker tools-technologies and related activities were introduced to the graduate students. The participants for the course included Ph.D. students and Master's students in an interdisciplinary programme, and were divided into six teams with four members each. For the study, we have considered a making activity: *A co-design activity to reimagine, redesign any of the daily encountered object/device, find pitfall(s) of the object/ device, and ideate for modifications and repurposing, with a stakeholder within the team*. Tinkercad was the recommended tool to engage in 3D modeling as it supported real-time online collaboration for designing. The activity was spread over 3 design sessions, with each session of 3 hours once in a week. Team selection and composition were done by facilitators to ensure balance in terms of disciplinary backgrounds, and prior experience with maker tools, in each team. Teams were directed to design or redesign objects or devices which are tangible and have a form so that 3D designs can be printed. Teams were advised to keep a design journal, but no format was given for the entries. The members engaged in the online making activity of design, through MS TEAMS. We looked for data set that covered teams' making actions from start to end of the activity, who have completed the task. Following these requirements, we have used the video recording of Team K, whose solution approach was found to be novel and addressed the particular design problem, as per the course instruction team. The team comprised of 4 participants, M1- first year Ph.D. student, M2-second year Ph.D. student, M3-first year Master's student, M-4 second year Master's student, and facilitators - F1, F2 as indicated in the excerpts. The members of Team K worked on solving a design problem with the problem statement*: Existing containers/ vessels are not well designed for carrying multiple beverages at the same time, especially in scenarios of meetings where attendees opt for different beverages*. The Team presented ideas and proposed solutions at the end of the design activity.

2.2 Analysis

We analysed the data by content logging and segmenting the data into various phases of the design process such as ideating, information seeking, process planning and organizing, documenting, sketching, 3D modeling, refining ideas, and non-task related activities, to get an overall understanding of the making process in our context. We follow the case study methodology (Merriam, 2007) to examine the making activity of Team K. The online making sessions were recorded using screen recording. The authors were present during the making sessions and made observations to support the data analysis. We used the content logs to conduct interaction analysis (Jordan & Henderson, 1995) focusing on talk, use of artifacts, tools and technologies for communication, and making (Steier, & Davidsen, 2021). While replaying the video and parsing the transcripts iteratively, we focus on dialogic conversations and communications through MS TEAMS between participants along with their manipulation of available resources. We followed the team's making actions in Tinkercad, and actions on other shared artifacts like documents, 3D models, images etc. The idea units are considered for a unit of analysis where we look at episodes with the idea entry, evolution, and exit from the various making phases.

3. Findings

The study revealed evidence of collaborative emergence of ideas across multiple encounters of design sessions, where ideas get accumulated and modified on a moment-to-moment basis, affected by the interaction of or with human and material resources, which translate towards the conclusive proof of concept, capable of addressing the problem in hand. We found the emergence of two core ideas in the making sessions, where the makers attempted to address the previously mentioned problem statement. The two core ideas for the multiple beverage storage container are the idea for a structural unit of the container and the idea for a monitoring unit of the container. The creative making process is found to be

entangled with: People - makers, facilitators, and Materials - 3D models in the Tinkercad shape library, product images from the internet, and sketches.

3.1 Representative episode of collaborative emergence of the idea for monitoring unit

The following episode represents an instance from the ideation phase in the first design session. Participants are using the brainwriting technique to note down ideas in spreadsheets, to solve the problem at hand. Here M3 introduced the idea of using the Internet of Things (IoT) for making the flask design under consideration, a smart device. The other participants are found to build on it and the accumulated ideas are parked for proposing an IOT-based vessel with multiple compartments.

M1 : M3 turns out to be.. he has written IOT based thing [refers to idea written in spreadsheet]

M2 : Oh.. but we will need sensors

M2 : Some software installed in phone and.. then we can control.

M1 : .. we can see the levels.

M3 : Actually.. we have different compartments,. right?.. with different fluids of different temperature.. so helpful with sensing and heating

M1 : Ok.. we keep battery based reheating then.. it can be taken to other places

The dialogic exchanges between the team members show the development of the core idea for smart technology integration in an incremental fashion. The spreadsheet becomes a shared artifact with a collection of ideas that can improve the functionality of the proposed container design. As the sensor-based monitoring is identified by team members, the thread for using IOT for the container is found to be expanding on a moment-by-moment basis. In the final presentation after the activity, the team proposed an IoT-based smart flask, for which a 3D model was created using Tinkercad collaboratively. Even though the model has little reference to IoT components, the final idea and proposed solution consisted of a smart device package and structural package. In the above excerpt, the idea of the monitoring unit cannot be attributed to a team member alone, but as a collectively shaped functionality.

3.2 Representative episode of collaborative emergence of the idea for structural unit

The following episode represents an instance from the 3D modeling phase in the third design session. The team used Tinkercad to realize the design ideas. Figure 1 shows the Tinkercad interface, using which multiple designs were generated. Team members are found to work on a single design at a time, by building on each other's input, refining for further changes, and then moving to a new model.

M1 : .. Just check this M4.. in the chat [shares a link to 3D model, in the chat].. That's a 3d model, but it is cap and all is simple it seems.

M4 : Yeah.. this is pretty good.. we cannot download it.. but can be created.

M1 : mm.. try it. [imports a cylinder shape from Tinkercad Library] .. or can we design a different one.. with cylinder three compartment and common cap .. you got what i said?

M4 : ..not really

M1 : [Moves pointer over the rectangular box model].. u know this hot box right?.. let me show u this picture..[searches images on the internet]. See.. like this..

M1 : [Moves pointer over the cylindrical model] this idea we thought of.. then dropped

for.. some reason

M4 : I was just thinking.. one cylinder and. 3 triangular compartments.. the cap will be something like the one you said and when you rotate it, it changes the compartment

M1 : Initially, we thought about moving the compartment, but F1 said not to use a lot of rotation parts .. the stability issues

M1 : Here.. rotation can be only cap rotation right?

M4 : Yeah..

M2 : [Rejoins after getting disconnected earlier] the problem with inside openers at top will be.. some mechanism required to make the fluid reach top

M1 : Yeah.. correct

M2 : ..But bottom one is helpful in that case

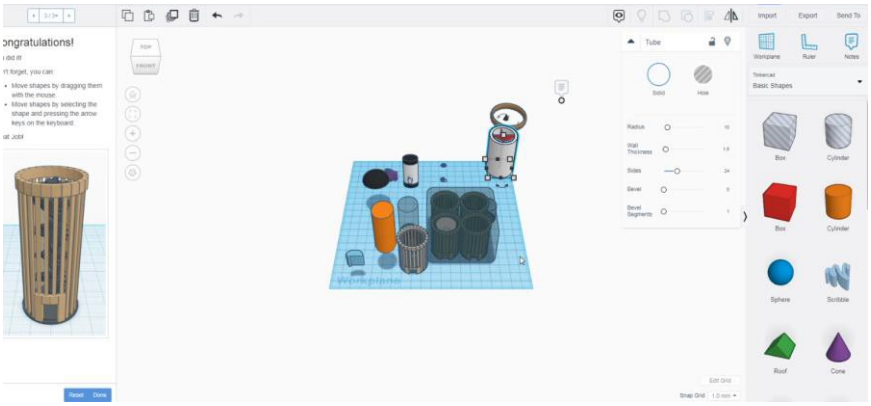


Figure 1. Multiple designs created by Team K during the modelling phase using Tinkercad.

The ideas and designs that are getting refined in the above excerpt are reflected in the collaboratively built 3D model and presented after the activity, as shown in Figure 2. The structural package is found to have 3 compartments, with 3 taps at the bottom and fewer moving parts. Here the dialogic interaction starts with M1 sharing a link via chat, and M4 getting inspired and responding to it. M1 and M4 then jointly decided to drop the rectangular box idea and follow the cylinder idea which is also used in the model shared previously. Also, M1 makes use of hot box product images from the internet, to convey the idea of moving compartments. M1 then points out the suggestions from F1, and are duly considered in the design decisions related to moving parts. It is also seen that M2 who got disconnected earlier due to network issues, rejoins and finds a problem with top opener ideas and suggests bottom ones. This episode of making is enriched by the elements: makers, facilitator, shared 3D model, Internet resources, contributing towards the emergence of the idea for a structural unit. These interactions suggest that making can be considered as a distributed process, with the scope for collaborative emergence of creativity.



Figure 2. 3D model presented by Team K as a final structural package.

4. Discussion

We have presented a case study of online collaborative making where team members worked on solving a design problem. Even though the team has followed established routines of problem-solving like brainwriting, we found instances in the making process where makers collectively contributed to developing the solution, one maker's action followed by the other in adding or refining the idea, facilitator inputs becoming instrumental in the making, material entities affording and mediating the making, thus underlining the distributed form of creativity. With the representative episodes, we have illustrated the emergence of two different core ideas for solving the design problem, from the less scripted collaborative making activity, distributed across the social and material spheres (Glăveanu, 2014; Sawyer & Dezutter, 2009). The first episode shows the interaction among makers taking up creative roles of synthesizers and analyzers, thus playing a vital part in arriving at the concept of using the IoT package for the proposed solution. The second episode showed the interaction of makers, between makers and online resources, and the influence of the facilitator in arriving at the final idea for the structural package, which is reflected in the 3D model created in the making activity. Here, the bidirectional connection between makers in which perspectives are exchanged, and differences are converted into creative expression, highlights the social dimension of creativity. The technological affordances of the tools, communication channels, directive nature of online resource materials, and the corresponding ways of utilization, point to the material line of distribution of creative online making (Literat & Glăveanu, 2018). Although the sociality and materiality aspects of creativity in the collaborative making are mentioned separately, these elements are found to be closely intertwined.

The findings from the study support the conjecture that the collaborative emergence of creative outcomes is shaped by not only social elements but also by the material elements of the situated-making context. The cases presented in the article bring in the materiality aspects of online collaborative making which are often overlooked in understanding creativity in online making contexts. With these initial findings, we suggest to factor materiality to expand the notion of collaborative emergence, which is rooted in sociality (Sawyer, 2014). The study also adds to the positioning of creativity in collaborative making as a distributed phenomenon where the creative process is constituted by the ecosystem of makers, facilitators, materials, and cannot be narrowed down only to individualistic processes. When compared to events of collaborative emergences in theatrical performances (Sawyer & Dezutter, 2009), design acts are of fragmentary in nature. Further studies are required to expand our understanding related to the socio-material entanglement in the creative making, especially that of online contexts, which can inform how the collective action of creativity can be fostered. Similar studies can be potentially helpful in getting insight into the distributed learning that takes place in various sites of collaborative making.

5. Implications & Limitations

Along with investigating creative making actions in online settings, the study also shows the potential of online making for learning. Researchers and educators should attend to design features of learning environments that can afford online collaboration and support tinkering. The materiality of online collaborative making needs to be studied deeper to enhance such distributed modes of learning.

The study focuses on limited data of an online making setting, where novice makers engage in groups to solve design problems. So, the findings cannot be generalized across other making settings and sites of collaboration. Analyses from several making contexts and among diverse makers would have given a wider perspective on collaborative emergence and distributed creativity. It would be also interesting to see how novice makers engage in creative making in physical and hybrid settings, using different modes of digital fabrication, physical materials, and computational units, which bring out more embodied interactions.

References

Aguilar, D., & Pifarre Turmo, M. (2019). Promoting social creativity in science education with digital technology to overcome inequalities: a scoping review. *Frontiers in Psychology*, *10*, 1474.

Barniskis, S. C. (2014). STEAM: Science and Art meet in rural library makerspaces. *IConference 2014 Proceedings*.

Clapp, E. P. (2017). *Participatory creativity: Introducing access and equity to the creative classroom*. Routledge.

Dougherty, D. (2012). The maker movement. *Innovations: Technology, governance, globalization*, *7*(3), 11-14.

Glăveanu, V. P. (2014). *Distributed creativity: Thinking outside the box of the creative individual*. Cham/Heidelberger: Springer International Publishing.

Kozbelt, A., Beghetto, R. A., & Runco, M. A. (2010). Theories of creativity.

Lille, B., & Romero, M. (2017). Creativity assessment in the context of maker-based projects. *Design and Technology Education: An International Journal*, *22*(3), 32-47.

Literat, I., & Glaveanu, V. P. (2018). Distributed creativity on the internet: A theoretical foundation for online creative participation. *International Journal of Communication*, *12*, 16.

Merriam, S. B. (1998). *Qualitative Research and Case Study Applications in Education. Revised and Expanded from" Case Study Research in Education."*. Jossey-Bass Publishers, 350 Sansome St, San Francisco, CA 94104.

Sawyer, R. K., & DeZutter, S. (2009). Distributed creativity: How collective creations emerge from collaboration. *Psychology of aesthetics, creativity, and the arts*, *3*(2), 81.

Sawyer, R. K. (2014). *Group creativity: Music, theater, collaboration*. Psychology Press.

Schmoelz, A. (2017). On co-creativity in playful classroom activities. *Creativity. Theories–Research-Applications*, *4*(1), 25-64.

Smith, S. M., & Ward, T. B. (2012). Cognition and the creation of ideas. *The Oxford handbook of thinking and reasoning*, 456-474.

Steier, R., & Davidsen, J. G. (2021, June). Adapting Interaction Analysis to CSCL: A systematic review. In *14th International Conference on Computer-Supported Collaborative Learning (CSCL): ISLS Annual Meeting 2021 Reflecting the Past and Embracing the Future* (pp. 157-160). International Society of the Learning Sciences (ISLS).

Sullivan, F. R. (2011). Serious and playful inquiry: Epistemological aspects of collaborative creativity. *Journal of Educational Technology & Society*, *14*(1), 55-65.

Tanggaard, L. (2013). The sociomateriality of creativity in everyday life. *Culture & Psychology*, *19*(1), 20-32.

Timotheou, S., & Ioannou, A. (2019, September). On making, tinkering, coding and play for learning: A review of current research. In *IFIP Conference on Human-Computer Interaction* (pp. 217-232). Springer, Cham.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Support System for Understanding How Teammates Grasp Game Situations

Kazuteru Sakamoto[a*], Kota Kunori [a] & Tomoko Kojiri[b]
[a] *Graduate School of Science and Engineering, Kansai University, Japan*
[b] *Faculty of Engineering Science, Kansai University, Japan*
*k511978@kansai-u.ac.jp

Abstract: In team sports, each player's move is affected by their mental state. A successful tactic in team sports is to understand the mindset of teammates. The current situation of the game can provide clues for determining a player's mindset. For instance, if the teammate thinks his team has the advantage, he feels encouraged and can move actively and confidently. Therefore, determining how teammates grasp the game situation is important for tactical success. This study develops a support system for understanding how teammates grasp the game situation. This system consists of two subsystems; the perceived game situation input system is used by individual members to input the perceived game situations of past games and the game situation sharing system shows the current perceived game situation of all team members. By visualizing the perceived game situation in two ways, members are able to recognize the trigger events of grasping negative situations easily.
Keywords: game situation, team sports, visualization, mental situation

1. Introduction

In team sports such as soccer and rugby, the players' actions and movements are planned carefully and tactically. However, no matter how successful they are in practice, these plans sometimes fail during actual play. One reason is that the members' movements are not always the same and can change depending on the player's mental situation (Campo et al., 2018) ; e.g., if they are depressed or nervous, they may not perform as they did in practice. If members are aware of their teammates' mental state, they are able to support each other.

Mental states change depending on how the game flow is perceived (Schmidt et al., 2020), which we call subjective flow. Subjective flow can be determined by the perceived current situation and the prediction of future situations. For example, if the current situation is advantageous and the future is disadvantageous, the subjective flow might be negative and the player's mindset becomes depressed. The perception of the current situation and prediction of future events are determined by the subjective criteria of players and differ between players. If the trigger events of determining the perceived current situation and the prediction of the future situation can be understood, it is possible to infer the mental states of members during the game.

The types of trigger events for the members may be the same, so it is effective to consider the trigger events of past games. As a first step toward grasping the mental situation, we propose a system for collecting the perceived current situation of past games and sharing them with the members. During such activities, it is important to determine how to share the perceived current situation from the viewpoint of understanding the subjective criteria. There are several systems that aim to visualize the knowledge of group members for collaborative learning. Many focus only on sharing current knowledge (Janz et al., 2009) and do not address the trigger for acquiring such knowledge. Our system visualizes the perceived current situation with the game events that trigger the perception.

2. Overall Framework of System

2.1 Mental State and Trigger of Perceiving Game Situation

To grasp the mental states of teammates, we first propose a model representing how mental states are determined. Figure 1 shows the model for identifying the mental situation from game events. The players use the events in the game to perceive the current situation. They also predict future situations from current game events. The difference between the current and future situations establishes the subjective flow of the game and this flow determines the mental situation.

For example, based on the event of losing points and falling behind, a player considers the current situation to be a "disadvantage." Since he is pessimistic and believes his team is going to lose more points in the future, he considers the future situation to be "more disadvantageous." By comparing the current and future situations, the flow appears negative, so his mental state is depressed.

In this process, how the player perceives current and future situations differs for each player based on their personal criteria. Determining the trigger of how the situation is perceived makes it possible for the player to infer the mental states of the other members.



Figure 1. Model for Identifying the Mental Situation by Game Events

2.2 System Overview

This study supports recognizing the triggers of the advantage/disadvantage perception of the current situation. Grasping the future situation, flow, and mental situation remains as our future work.

Since triggers are established depending on the players' criteria and might be common for all games, finding triggers based on past games is effective. This study proposes a support system for understanding how teammates grasp the current game situation by sharing the perceived situation of past games based on game events.

Figure 2 is the overall framework of our system. It consists of two subsystems: the perceived game situation input system and the game situation sharing system. The perceived game situation input system is used by individual members to input the perceived game situation of a past game. The system displays the game's video on each player's interface. The players input the perceived situation, such as whether it is considered an advantage or disadvantage, at each point in time. The input data is saved as perceived situation data. The game situation sharing system shows the perceived game situation of all team members. It arranges the distribution of the perceived data so as to find the players' trigger events or tendencies of teammates easily.

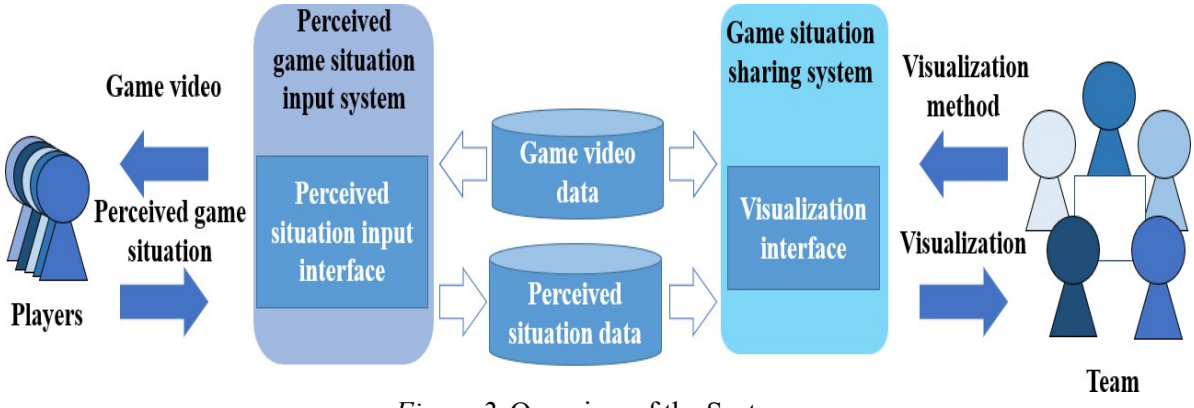


Figure 2. Overview of the System.

3. Perceived Game Situation Input System

The perceived game situation input system allows individual players to input their perceived game situation along with the game flow. Figure 3 shows the interface. The video display area shows the game, allowing the player to watch the video using the start and stop buttons. The perceived situation can be inputted by stopping the timing in the video where the player recognizes the situation, selecting the type of situation using the perceived situation input area, and pushing the input button. At the beginning of the game video, the system recognizes that the player thinks the situation is "neither (advantage /disadvantage)" and, until the perceived situation changes to advantage or disadvantage, it recognizes that the player perceives the same situation. The input situation display area designates the inputted situations along with the time by color. Its horizontal axis corresponds to the time sequence of the game video and the time periods are colored by the perceived situation of the player; that is, blue for "advantage," yellow for "neither," and red for "disadvantage." When the save button is pressed, the input situation is stored as the perceived situation data.

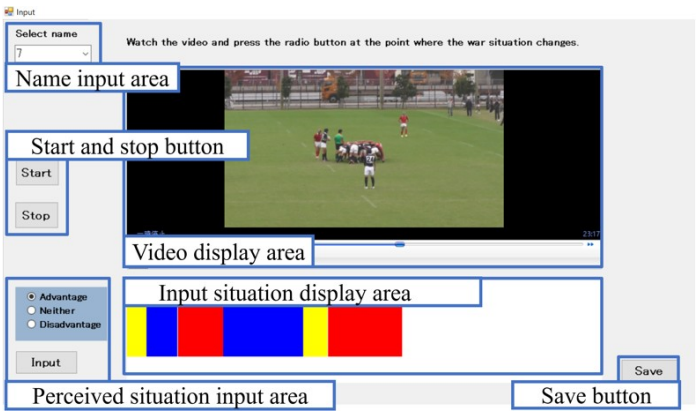


Figure 3. Perceived Situation Input Interface

4. Game Situation Sharing System

The game situation sharing system displays the perceived situation of all members to support them in determining the trigger game events of perceiving the situation. Currently, the system assumes that players tend to grasp two events; one is the trigger event for individual players who perceive situations differently than others and the other is where most of the members perceive the situation as negative. To help members recognize such events easily, this system provides two visualization modes: *whole team mode* and *each player mode*.

Figures 4 and 5 show the interface of the two visualization modes. In whole team mode, the number of players who perceive each situation is counted and indicated as the height of each color, with blue for "advantage," yellow for "neither," and red for "disadvantage." In each player mode, each player is assigned to each line, as shown in Figure 5. In both modes, the horizontal axis represents the time axis and, by clicking the point on the axis, events that occur at that time are highlighted in the list of events area. Table 1 shows the elements of the list of events prepared for a rugby game.

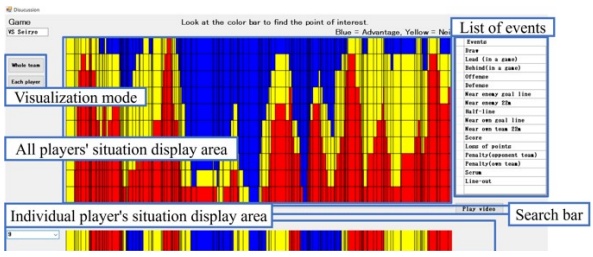


Figure 4. Visualization Interface (Whole Team Mode).

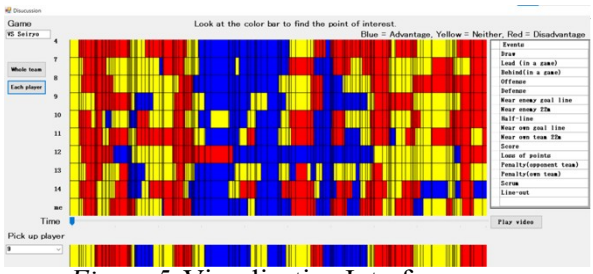


Figure 5. Visualization Interface (Each Player Mode).

Table 1. List of events

Scores: draw, lead, behind
Side: offense, defense
Position: near enemy goal line, near enemy 22 m, half-line, near own goal line, near own team 22 m
Events: score, loss of points, penalty (opponent team), penalty, (own team), scrum, line-out

5. Conclusion

In this paper, we proposed a method for estimating the teammates' mental situations based on perceived current and future situations. In addition, we described the development of a system for grasping the trigger events of perceiving the current situation as advantage/disadvantage using the experience of past games. The system proposes an environment that enables members to express their perception of situations during the game. It also visualizes the distribution of the grasped game situations of the team members for understanding the triggers for determining the situation.

In our future work, we will evaluate the effectiveness of our visualization interface for grasping the trigger events of the current situations of teammates. In addition, we will develop a support method for grasping future situations and mental states.

References

Campo, M., Martinent, G., Pellet, J., Boulanger, J., Louvet, B., & Nicolas, M. (2018). Emotion–performance relationships in team sport: The role of personal and social identities. International Journal of Sports Science & Coaching, 13(5), 629-635.

Janz, B. D., & Prasarnphanich, P. (2009). Freedom to cooperate: Gaining clarity into knowledge integration in information systems development teams. IEEE Transactions on Engineering Management, 56(4), 621-635.

Schmidt, S. C., Gnam, J. P., Kopf, M., Rathgeber, T., & Woll, A. (2020). The influence of cortisol, flow, and anxiety on performance in E-sports: A field study. BioMed research international, 2020.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

GWpulse: Supporting Learner Modeling and Group Awareness in Online Forum with Sentiment Analysis

Yuta NAKAMIZO[a*], Changhao LIANG[a], Izumi HORIKOSHI[a], Rwitajit MAJUMDAR[a], Brendan FLANAGAN[a] & Hiroaki OGATA[a]
[a]*Kyoto University, Japan*
*nakamizo.yuta.83a@st.kyoto-u.ac.jp

Abstract: Collaborative learning in online context has been in educational practice and also gained attention during the emergency remote teaching due to the pandemic. This study follows the Group Learning Orchestration Based on Evidence (GLOBE) framework that considers four phases of technology-supported group work and proposes data-driven services for each phase. In this paper, we present an analysis tool named "GWpulse" that processes learning log data and shares the analysis results as an attribute of the learner model considered in GLOBE and also presents it in the group work dashboard to facilitate group awareness. In the current version GWpulse considered logs of forum activity generated during group activities in the learning management system. Apart from including that data in the learner model, it also visualizes the activity data for the orchestration, evaluation and reflection phases. GWpulse dashboard displays the basic statistics of their students' forum activities such as time interval and the number of posts, as well as a novel indicator of "Assistance Needed Level" calculated using sentiment analysis method that classifies textual statements into positive and negative.

Keywords: Group awareness, learner modeling, learning analytics, sentiment analysis

1. Introduction

Collaborative learning is becoming increasingly important in educational activities (Stahl et al., 2006). The GLOBE framework organizes research to support and analyze collaborative learning, sketching a cycle composed of 1) group formation, 2) orchestration, 3) evaluation, and 4) reflection (Liang et al., 2021). In each phase of the GLOBE framework, learning log data are used and generated by each other. Beyond the GLOBE framework, there is an integrated learning platform called "LEAF", which captures various learning activities as log data for learner modeling and analysis and provides visualized feedback to teachers and students (Franagan & Ogata, 2017). Once group activities under GLOBE framework are conducted, learning log data can be integrated as part of this platform as well and used as evidence in different modules in LEAF.

In this paper, we propose a system named "GWpulse" that stores and visualizes the results of the analysis of conversation logs in Moodle forums as learning log data in the orchestration phase of group activities in light of the GLOBE framework. Our goal is to help teachers and students understand the status of group activities and support their decision making with group awareness information (Schnaubert et al., 2019) to improve performance during the group work. Sentiment analysis is implemented to provide an indicator of how much intervention by the teacher is needed.

2. Related Works of Group Activity Visualization and Sentiment Analysis

There have been several group awareness tools that visualize asynchronous group activities on online forums. For example, an application has been proposed which displays a graph showing the number of contributions and online status of each student, with some patterned sentences to encourage equal distribution of participation when student contributions are unbalanced (Strauß & Rummel, 2021).

Experiments with this application showed that simply checking of the tool by students did not affect the equal distribution of participation, but the satisfaction with the group activity decreased when the equal distribution of participation was biased. Conversely, the satisfaction with the group activity increased when the equal distribution of participation was achieved.

There are various types of sentiment analysis, each with different characteristics. Approaches using Support Vector Machine (SVM) can classify the dataset linearly (Medhat et al., 2014). A framework to find important reviews efficiently by classifying the quality of product reviews on the Internet (Chen & Tseng, 2011) and a framework to discover market intelligence (MI) to support decision makers by aggregating and classifying opinions on microblogs have been proposed by using SVM classification (Medhat et al., 2014). Asari (Nakayama, 2019), an open-source software under MIT license, is one of the SVM modules for Python that performs positive / negative sentence classification of Japanese sentences.

3. System Design

Figure 1 shows the data flow of GWpulse working with Moodle and analysis tools in LEAF, and UI of the group awareness dashboard of GWpulse. In addition to visualizing the results of the analysis of the conversation logs during the group activities, it stores them and makes them available to other phases of the GLOBE framework and other modules in LEAF. The database of learning analytics dashboard (Dashboard DB) stores learner model. The learner model enhancer module in GWpulse periodically copies tables from the Moodle database (Moodle DB) to the common database of Dashboard DB.

When user access to the group awareness dashboard, 1) user selects Forum and/or Discussion. 2) Then, the average of the count of posts, count of characters, and interval between posts of the group, 3) group level Assistance Needed Level (ANL), mean negativeness score for sentiment analysis for each post, divided into four levels with thresholds of [0.00, 0.20], (0.20, 0.40], (0.40, 0.55], and (0.55, 1.00], 4) graphs of count of posts, count of characters and interval of posts for each user, 5) and post level ANL are displayed for each discussion. Sentiment Analysis is performed using "asari", a Python library specialized for positive / negative classification of Japanese sentences.

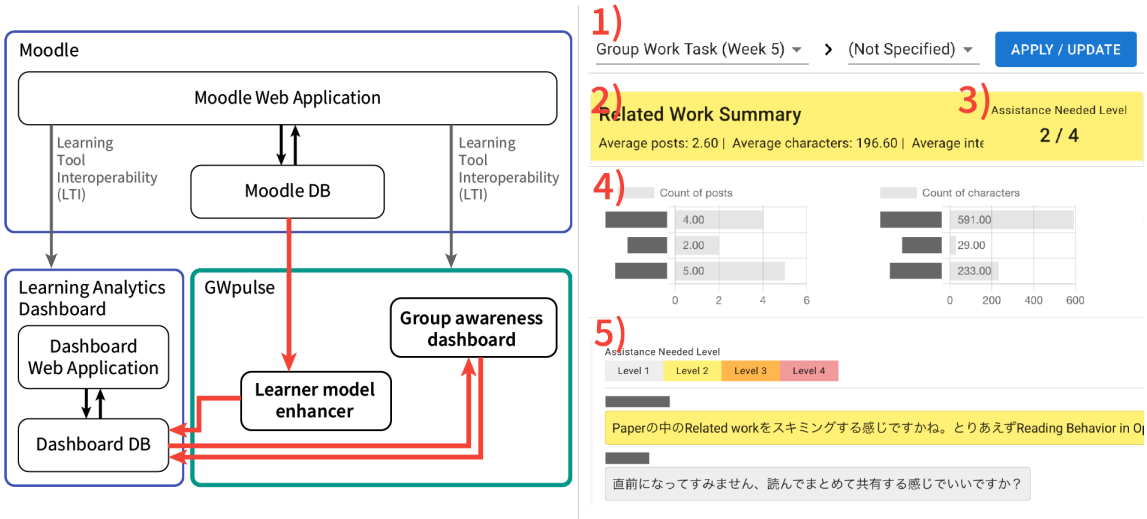


Figure 1. The data flow and the UI of the group awareness dashboard

We expect that GWpulse functions can be utilized in orchestration phase as well as other phases.

Evaluation Phase: By checking the Group awareness dashboard, students can look back at their own, their group's, and other group's activities objectively and visually. In other words, by checking the visualized data by GWpulse in the evaluation phase, we can expect each participant to make evidence-based evaluations.

Reflection Phase: Students can check their own or other students' or other groups' activities with GWpulse when they reflect on their own participation in the group activities after receiving peer evaluation from other students or evaluation from the teacher.

Formation Phase: Teachers can use aggregated indicators calculated from GWpulse that describe the characteristics of students in the group learning process as parameters when creating groups for the next group activity.

4. Discussion and Conclusion

We expect that visualization and confirmation of these indicators can be used as evidence for teachers to decide whether to make interventions such as putting actively participating students in leadership roles and encouraging passive students. Formative indicators such as count of posts can also remind those who are passive to be aware of their low motivation and change attitude toward participation thus improving their group work performance in a timely and informed manner (Coll et al., 2014). It is expected that with data accumulation by GWpulse, we can use them as parameters for creating a learner model in the group formation phase and other learning analytics tools.

The group work module "GWpulse" proposed in this paper is developed as an extension of the existing learning analysis platform that contributes to the accumulation and transparency of learning log data during the orchestration phase of group activities. By combining this system with the existing systems utilized in formation and evaluation phases, it will be possible to accumulate learning log data in all phases of the GLOBE as comprehensive group learning evidence.

In future work, we are planning to enable GWpulse to provide feedback using the learning log data generated by other modules in LEAF, and also to store learning log data such as the history of interventions by the teacher and other decisions made by using GWpulse. In parallel, we are planning to measure the effects of GWpulse during group activities in real classes. And we also need some experiments to compare human-based classification ANL and threshold-based classification ANL because the threshold lacks scientific validation.

Acknowledgements

This research work is supported in part by the following funding: JSPS KAKENHI 20H01722, 22H03902, NEDO Special Innovation Program on AI and Big Data JPNP18013, JPNP20006.

References

Chen, C. C. & Tseng, Y. (2010). Quality evaluation of product reviews using an information quality framework, *Decision Support Systems, 50*(4), 755-768.

Coll, C., Rochera, M. J., & Gispert, I. (2014). Supporting online collaborative learning in small groups: Teacher feedback on learning content, *academic task and social participation. Computers and Education, 75*, 53-64.

Flanagan, B. & Ogata, H. (2017). Integration of Learning Analytics Research and Production Systems While Protecting Privacy, *Proceedings of the 25th International Conference on Computers in Education*.

Liang, C., Majumdar, R. & Ogata, H. (2021). Learning log-based automatic group formation: system design and classroom implementation study, *Research and Practice in Technology Enhanced Learning, 16*(14).

Medhat, W., Hassan, A. & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey, *Ain Shams Engineering Journal, 5*(4), 1093-1113.

Nakayama, H. (homepage on the internet). Hironsan/asari: Japanese sentiment analyzer implemented in Python. (cited 2022, May 31). Available from https://github.com/Hironsan/asari/.

Schnaubert, L., Heimburch, S., Erkens, M. & Bodemer, D. (2019). Cognitive Group Awareness Tools: versatile devices to guide learners towards discrepancies, *Proceedings of the 27th International Conference on Computers in Education., 1*, 158-164.

Stahl, G., Koschmann, T., & Suthers, D.D. (2006)., Computer-Supported Collaborative Learning: A Historical Perspective, Cambridge Handbook of the Learning Sciences, R.K. Sawyer, ed., 409-426, Cambridge Univ.

Strauß, S. & Rummel, N. (2021). Promoting regulation of equal participation in online collaboration by combining a group awareness tool and adaptive prompts. But does it even matter?, *International Journal of Computer-Supported Collaborative Learning, 16*, 67–104.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Improving Spatial Perspective Taking Ability of Middle School Students Using Augmented Reality

Anveshna SRIVASTAVA* & Chandan DASGUPTA
Educational Technology, Indian Institute of Technology Bombay, India
*anveshna.sriv@gmail.com

Abstract: There is great interest in understanding the pedagogical value of Augmented Reality (AR) based systems. Here, we measured the effect of an AR tool - GeoSolvAR - on spatial perspective-taking ability of students. The GeoSolvAR application helps learners to visualize 2D images as 3-D objects. In this pilot study, we report data from 16 middle school students who were individually administered the online Perspective Taking Spatial Orientation Test (PTSOT) both prior to and after performing the GeoSolvAR activities. Our results suggest that the students showed a significant decrease in angular errors on PTSOT in the post-test with a large effect size (Cohen's $d = 1.26$). A cohort-level analysis also revealed larger and equal effect sizes in perspective-taking spatial abilities for both males and females (Cohen's d = 0.69). However, our finding with a subset of 10 students reveal that the GeoSolvAR activities were challenging for the students and that there was a positive but non-significant correlation of students' GeoSolvAR activity with their PTSOT performance. We discuss the limitations, and the way forward for this work, while making a case for using the GeoSolvAR application as a spatial-skill training tool.

Keywords: Spatial Skill, Perspective-taking Ability, Augmented Reality, 3-D Visualization, Middle School Students

1. Learning and Spatial Reasoning

Spatial reasoning is concerned with how we factor the objects in the world, their shapes, their locations, their relations to each other, their potential movement paths, in our decision-making processes (Newcombe, 2017). This reasoning is visible in seemingly inconsequential actions like navigating the spaces in our homes to potentially consequential actions of manipulating the spatial structure of a biomolecule. In the context of learning and social success, spatial reasoning abilities have been particularly demonstrated to play a significant role in performance in school (Lowrie & Diezmann, 2007) and career achievement in STEM (science, technology, engineering and mathematics) and related fields (Wai, Lubinski, & Benbow, 2009).

Given considerable evidence on the long-term impact of spatial reasoning ability on learning and achievement in STEM subjects, recognition of and emphasis on spatial abilities as an essential component of learning foundations, on par with the three *R's* of reading, writing and arithmetic is warranted (Ishikawa & Newcombe, 2021). Just as learners can be trained on and their skills honed for competency in the three R's, they can also be trained to improve their spatial abilities. This follows from multiple streams of research which suggest that spatial skills are sufficiently malleable to permit effective and economically feasible training (Francis, Rothschuh, Poscente, & Davis, 2021).

Literature on spatial ability suggests differentiation of two kinds of spatial transformations- a) object-based transformation, where an object is imagined to transform and, b) perspective-based transformation, where one's point of view is transformed in relation to something else (Bryant & Tversky, 1999). In this study, we focus on the latter form of spatial ability which is concerned with mentally representing and perceiving a situation from a different viewpoint than one's own (Galinsky, Ku & Wang, 2005).

Spatial perspective-taking is a two-step process. In the first step, one mentally simulates the action of situating one's body at a different location from where it actually is and this facilitates the

second step of imagining the perspective one would have by being located in the new space. Thus, spatial perspective-taking is effort intensive (Tversky and Hard, 2009). Further, since different perspectives might help one uncover specific nuances of a problem, it is argued that perspective-taking promotes learning by filling in gaps in comprehension and enabling mental manipulation of the frame of reference.

2. Perspective Taking Spatial Orientation Test

The Perspective Taking Spatial Orientation Test was first developed by Kozhevnikov and Hegarty (2001). It is a measure of the ability to make egocentric spatial transformations in which one's egocentric frame changes with respect to the environment (Thurstone, 1950). For the present study, we used the online version of the PTSOT (Friedman, Kohler, Gunalp, Boone & Hegarty, 2019). This test assesses participant's ability to imagine how a stimulus array will appear from another perspective. Participants were shown an array of seven 2-D objects (see Figure 1)..

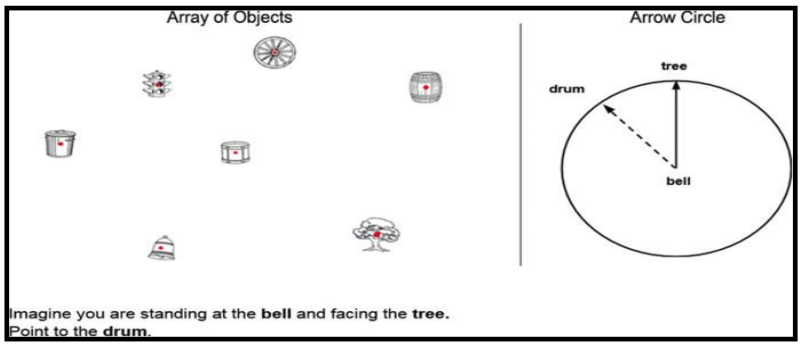


Figure 1. A sample question from the online PTSOT test.

Figure 1 represents a screenshot of a PTSOT question where an array of 2-D objects is shown on the left and on the right, an 'Arrow Circle' is represented. At the left bottom, one can see the question which asks the participant to imagine being positioned at a specific object ('bell' in the example), relative to a second object (the orienting cue; 'tree' in the example) and, the participant is further asked to predict the direction of the third object (target object; 'drum' in the example). In the arrow circle, the participant was always placed at the center, while the orienting cue was placed right in front of the imagined position (center). The task was to move the shown arrow, using mouse, in the circle so as to represent the direction of the 'target' object. The PTSOT had 12 questions which the participants had to complete within a time span of 5 minutes. These questions appeared in a randomized manner and responses were scored for angular error averaged across trials for each participant. We used PTSOT both as a pre-test and as a post-test for all the participants.

3. 3-D visualization and the AR tool- GeoSolvAR

The ability to mentally translate and visualize 2-D images into 3-D forms is a critical spatial ability, used in the learning of sciences (Srivastava & Ramadas, 2013), technology (Norman, 1994), engineering (Sorby, Drummer, Hungwe, Parolini, & Molzan, 2006) etc. It has been noted that students generally face difficulty with 2-D to 3-D transformations (Akasah, & Alias, 2010). Further, it has been shown that robust gender differences, favoring males, exist in perspective-taking spatial orientation abilities (Friedman, Kohler, Gunalp, Boone, & Hegarty, 2019).

Motivated by the last two arguments, and with support from literature on training for improving spatial skills, we studied an Augmented Reality (AR) tool- '*GeoSolvAR*' -which targets middle school students and helps them to solve problems related to visualization of 3-D geometric solids (Kaur, Pathan, Khwaja, & Murthy, 2018). The GeoSolvAR application is freely accessible on the web interface and can be installed on Android tablets/ phones with minimum Android OS version 4.0. This application

introduces basic concepts of 3-D views pertaining to directionality of objects, namely- top view, front view and side view. It has a set of six different activities related to different objects (see Figure 2).

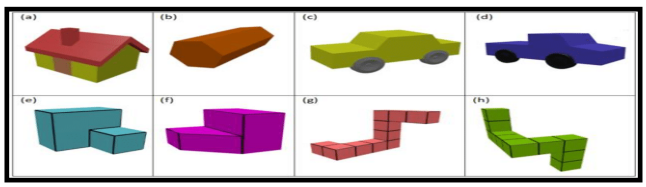


Figure 2. The 2-D images of models that were augmented via GeoSolvAR (Reproduced from Kaur, Pathan, Khwaja, Sarkar, Rathod, & Murthy, 2018)

AR tools function by superimposing computer-generated images onto an observer's view of the real world and, thus, have the potential to help with 3-D spatial skills (Kaufmann, & Schmalstieg, 2003). We found this opportunity with the GeoSolvAR application to see if the augmented 3-D view of the 2-D objects (in Figure 2) would help students to develop a changed spatial perspective for observing the objects in the environment. Specifically, AR activities, like in Figure 3, we hypothesized, can help with students' perspective-taking spatial orientation ability.

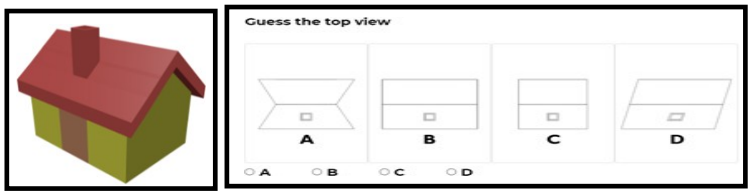


Figure 3. GeoSolvAR activity sample question. On the left, object 'house' is shown and, on the right, a question on predicting the 'top' view of the house is asked.

Further, since robust gender differences, favoring males, are known to exist on spatial abilities (Sorby, Drummer, Hungwe, Parolini, & Molzan, 2006), we began with the null hypothesis that no gender difference would exist on students' performance on the PTSOT with respect to the AR tool. To test these hypotheses, we conducted this pilot study where PTSOT was used as both pre- and a post-test and the GeoSolvAR 3-D visualization activities were used as an intervention.

4. Methods

4.1. Sample

16 middle school students (M=9; F=7; Average Age = 14.1 years; SD = 1.1) voluntarily responded to a general call for the pilot study. Their parents' consent and their assent were received before the conduct of the study. This study was approved by the IRB of the institute and the participants were free to leave the study at any point of time with no incurrent cost to them.

4.2. Data sources

The GeoSolvAR activities were conducted online on a web interface while the screen of the computer was being recorded. This provided us with data about the responses that students generated for each activity. Further, a student's performance on the computerized PTSOT was received at the backend of the experiment. This performance data carried information about the angular positions estimated by the individual for each target object for each trial in the set of 12 randomized questions. It also reported the angular error or the degree of deviation of the estimated angle from the actual orientation angle of the target object for all the trials and, finally, an average angular error was also computed across all the trials for the individual.

4.3. Task protocol

The study was divided into three phases, spread over a span of minimum 2 days and maximum 10 days, subject to the time-schedule of the students. In the first phase, the PTSOT pre-test was administered to a student. In the second phase, the student was introduced to the GeoSolvAR tool and was asked to perform all the six activities. In the third phase, the PTSOT post-test was administered. The first phase occurred on the first day, while the second and the third phases were conducted one after the other on the next day of student's arrival for the study. On an average, the gap between the first and the second day of the study was 5.3 days (SD = 3.59).

This gap was designed to reduce the possibility of simple recall of the pre-test questions and consequent biasing of the post-test data. It is to be noted here that students were not informed that they would be asked to do the PTSOT again on their other day of the study and, hence, it is likely that the students did not leave any mental markers to help them in the post-test questions.

5. Results

We analyzed the data to observe- a) if there was a change in student's performance on the spatial orientation task, b) student's performance on the GeoSolvAR activities, and c) if there was a correlation between student's performance on the PTSOT and the GeoSolvAR activities. We also performed a cohort-level analysis to find if gender played a role in spatial training through the GeoSolvAR activities. We share the findings from our study below.

5.1. Significant improvement in spatial perspective-taking ability

Overall, we found that the average angular error across the participants got significantly reduced in the post-test, or, the participants showed significant improvement in their perspective taking spatial orientation ability. A paired-samples t-test revealed that participants' average angular error degree in the pre-test (M = 74.91, SD = 19.16) was reduced by around 12 degrees in the post-test (M = 62.39, SD = 21.31), t (15) = 4.89, p = 0.00019. An extremely large effect size (Cohen's d = 1.26) was observed.

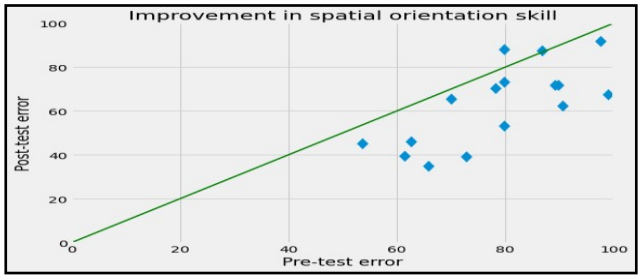


Figure 4. Scatter plot of pre and post angular errors for each participant (n=16)

Figure 4 shows a scatter plot comparing pre and post angular errors for all the 16 participants. The diagonal line depicts y=x, suggesting no change in the pre and the post scores. The students above the line show an increase in angular error and the ones below show decrease in angular error. Regardless of the small sample size, an average reduction of 16.7% in spatial orientation error is observed.

5.2. Students found GeoSolvAR activities challenging

A total of 14 predictive questions on the 3-D visualization activities were asked from the students as part of the GeoSolvAR tasks. For each correct response, students received score '1', else '0'. After solving activity questions, students were asked to use the GeoSolvAR application to view the augmented 3-D image of the 2-D objects shown in the activity question and they were free to change their responses, if they felt the need. Here, we report data from 10 students only and discard the rest where response to even one question was missing.

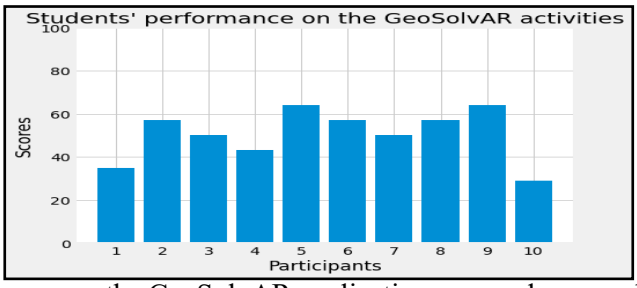


Figure 5. Students' scores on the GeoSolvAR application summed across all the 6 activities and presented on a scale of 100

Figure 5 shows a bar graph to depict students' scores on the GeoSolvAR activities on a scale of 100. The average score of the students is 50.6 (SD = 11.8), the maximum score attained was 64 and the minimum score was 28. This suggests that students found the GeoSolvAR activity challenging.

5.3. Positive correlation between PTSOT improvement and GeoSolvAR performance

We found a small positive but not significant correlation of the subset of students' perspective-taking ability with their performance on the GeoSolvAR activities, r (8) = .227, p = .53. The positive correlation discovered is encouraging for the use of the GeoSolvAR application for spatial skill training. However, given the small sample size, we cannot draw conclusions about the causal connection between the use of the application and its impact on students' perspective-taking spatial ability.

5.4. Large effect size gains in spatial orientation abilities of both males & females

Grouping the data into the two cohorts of male and female, we were able to analyze the role of gender on the perspective-taking spatial orientation ability. A paired samples t-test revealed that the average angular error in males in the pre-test (M = 78.29, SD = 11.78) was reduced in the post-test (M = 64.26, SD = 21.11), t (8) = 1.96, p = .08. The paired samples t-test in the females, similarly, revealed that the average angular error in the pre-test (M= 70.54, SD = 26.33) was reduced in the post-test (M= 59.9, SD= 23), t (6) = 1.69, p = 0.14. Notably, the effect size seen was large (Cohen's d = 0.69), and equal for both the males and the females.

5.6. Control

We understand that there's a possibility that students might be able to perform better on the post-test, just based on the repetition of the elements of the pre-test. However, this possibility is reduced because students are not provided with any feedback on their performance on the pre-test. Further, the average error scores were not shared with them after the test and, hence, students did not have any means to assess their performance. Thus, with no assessment baseline, students could not calibrate their actions in the post-test.

Nonetheless, for completeness, we administered the PTSOT task twice to two adults, one male (19-year-old) and one female (35-year-old). The individuals were not given the GeoSolvAR activity and were asked to repeat the test soon after they completed the first test. We found that the average error reduced by 0.91 degrees and 0.5 degrees in the post-test for the male and for the female respectively. This is an order of magnitude lower than the 12 degrees of average error reduction that we found in the 16 students who worked with the GeoSolvAR application.

6. Limitations & Way Forward

Our pilot study has revealed potential significance of the GeoSolvAR application as a tool to help improve the perspective-taking spatial orientation abilities of middle school students. We further found that the AR application led to large and equal effect size in both males and females (Cohen's d = 0.69).

However, given the small size of our pilot study, we cannot conclusively draw inferences about the role of the GeoSolvAR application in enhancing students' spatial ability.

Further, since students found the GeoSolvAR activities to be challenging, we might need to modify or improve the AR tool by including certain scaffolds, like reflection spots or well-placed cues, to help students navigate through the visuo-spatial activities (Kaur, Pathan, Khwaja, & Murthy, 2018). The limited data set also made it difficult to make sense of the low correlation between the GeosolvAR performance scores and the performance on the perspective-taking spatial ability.

For this work, we could not study the qualitative aspects of students' interaction with the AR tool and, hence, we are not able to pin-point specific underlying mechanisms that explains our result of high improvement on the spatial perspective-taking ability using the AR tool.

Nonetheless, our preliminary findings suggest that the GeosolvAR application has practical implications and that it can be readily used in a classroom setting. We look forward to build on this work and conduct more studies with larger groups of students. For future studies we also intend to analyze the video data for students' body movements, hand gestures and/or facial expressions to capture behavioral markers indicative of students' engagement with the mapping of the AR and the spatial tasks, like trying to orient one's body that can bring the participant closer in perspective to the imagined orientation. This would help us understand the different aspects of the AR tool which is helping students to improve their perspective taking spatial ability.

Acknowledgment

We are grateful to all the participants for their contribution to this study. We also thank Navneet Kaur, Rumana Pathan, Ulfa Khwaja and Sahana Murthy for illuminating discussions on the GeoSolvAR. Lastly, we thank the anonymous reviewers for their comments and suggestions.

References

Akasah, Z. A., & Alias, M. (2010). Bridging the Spatial Visualisation Skills Gap Through Engineering Drawing Using the Whole-To-Parts Approach. *Australasian Journal of Engineering Education*, 16(1), 81-86.

Bryant, D., & Tversky, B. (1999). Mental representations of perspective and spatial relations from diagrams and models. Journal of Experimental Psychology: Learning, Memory, and Cognition, 25, 137–156.

Francis, K., Rothschuh, S., Poscente, D., & Davis, B. (2021). Malleability of Spatial Reasoning with Short-Term and Long-Term Robotics Interventions. *Technology, Knowledge and Learning*, 1-30.

Friedman, A., Kohler, B., Gunalp, P., Boone, A. P., Hegarty, M. (2019). A Computerized Spatial Orientation Test. *Behavior Research Methods*.

Galinsky, A. D., Ku, G., & Wang, C. S. (2005). Perspective-taking and self-other overlap: Fostering social bonds and facilitating social coordination. *Group processes & intergroup relations*, *8*(2), 109-124.

Ishikawa, T., & Newcombe, N. S. (2021). Why spatial is special in education, learning, and everyday activities. *Cognitive Research: Principles and Implications*, 6, 20.

Kaufmann, H. and Schmalstieg, D. (2003). Mathematics and geometry education with collaborative augmented reality. *Computers & graphics*, Vol. 27, no. 3, pp. 339–345, 2003.

Kaur, N., Pathan, R., Khwaja, U., Sarkar, P., Rathod, B., & Murthy, S. (2018, December). GeoSolvAR: Augmented reality-based application for mental rotation. In *2018 IEEE Tenth International Conference on Technology for Education (T4E)* (pp. 45-52). IEEE.

Kaur, N., Pathan, R., Khwaja, U., & Murthy, S. (2018, July). GeoSolvAR: Augmented Reality based solution for visualizing 3D solids. In *2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT)* (pp. 372-376). IEEE.

Kozhevnikov, M. & Hegarty, M. (2001). A dissociation between object manipulation spatial ability and spatial orientation ability. *Memory & Cognition*, 29, 745-756.

Lowrie, T., & Diezmann, C. M. (2007). Solving graphics problems: Student performance in junior grades. *Journal of Educational Research*, 100(6), 369–378. doi:10.3200/JOER.100.6.369-378

Newcombe, N. (2017). Harnessing Spatial Thinking to Support Stem Learning. OECD Education Working Papers, No. 161, OECD Publishing, Paris.

Norman, K.L. (1994). Spatial visualization-A gateway to computer-based technology. *Journal of Special Educational Technology*, XII (3), 195-206.

Sorby, S. A., Drummer, T., Hungwe, K., Parolini, L., & Molzan, R. (2006). Preparing for engineering studies: Improving the 3-D spatial skills of K-12 students. *Science*, *88*(73), 65.

Srivastava, A., & Ramadas, J. (2013). Analogy and gesture for mental visualization of DNA structure. In *Multiple representations in biological education* (pp. 311-329). Springer, Dordrecht.

Thurstone, L. L. (1950). Some primary abilities in visual thinking. *Proceedings of the American Philosophical Society*, *94*(6), 517-521.

Tversky, B., & Hard, B. M. (2009). Embodied and disembodied cognition: Spatial perspective-taking. *Cognition*, *110*(1), 124-129.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology*, 101(4), 817–835. doi:10.1037/a0016127.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Noticing Critical Information for Productive Uncertainty Management During Engineering Design Activity

Navneet KAUR[a*], Chandan DASGUPTA[b]
[a]*IDP in Educational Technology, Indian Institute of Technology Bombay, India*
[b]*IDP in Educational Technology, Indian Institute of Technology Bombay, India*
*navneetkaur@iitb.ac.in

Abstract: The phenomenon of noticing information is crucial for experiencing uncertainties and influencing uncertainty management processes. This paper presents findings from an analysis of two sixth-grade teams engaged in a collaborative engineering design activity. We investigated how learners notice critical information during the problem-solving process and its role in resolving various uncertainties experienced by the team. Preliminary results show that three factors - mentors' invention, physical artifacts in the environment, and setbacks experienced during troubleshooting enabled crucial noticing processes that engaged learners in productive uncertainty management practices. Based on the findings, we then briefly discuss how to start strategizing scaffolds to help learners notice important information.

Keywords: Noticing, uncertainty, uncertainty management, engineering design, middle school learners

1. Introduction and Background

Wrestling with uncertainties provide opportunities for learners to meaningfully participate in the disciplinary practices (Kaur & Dasgupta, 2019; Manz, 2018). Uncertainties specifically play a crucial role in the context of engineering design. Due to the ill-structured nature of design tasks, engineers experience uncertainties about different aspects of the design problem all the time during the design process. For example, due to unclear objectives, analysis and evaluation of multiple solution alternatives, and troubleshooting issues (Dym, Agogino, Eris, Frey, & Leifer, 2005; M. E. Jordan & McDaniel, 2014).

Planning related to which uncertainties should learners must experience, and when and how should they experience these uncertainties can help learners engage in core disciplinary practices effectively (Manz, 2018). For experiencing such productive uncertainties that can engage learners in the disciplinary content and practices, it is first important that learners notice critical information relevant to the task that they are solving. This information could be certain concepts, scientific principles, or relationship between different variables involved in the task (Chase, Malkiewich, & S Kumar, 2019). Lobato et al. (2012) defines noticing as a process of actively selecting and interpreting relevant information. A given situation can contain enormous information to be perceived by the learner. Noticing affects how situations are perceived and interpreted and therefore shape learner's judgement and decisions (Chase et al., 2019).

This preliminary work focuses on investigating how learners notice critical information relevant to a given design task. Particularly, we seek to understand how noticing is related to learners' uncertainty experiences and the progress that they make during the design process while collaboratively resolving an engineering design problem. Thus, in this paper, we explore following research questions - (a) What factors trigger noticing processes when learners collaboratively solve an engineering design problem and (b) how does noticing affect teams' uncertainty management processes?

2. Methods

The data came from a study which was conducted with the broader aim of understanding middle school learners' uncertainty management processes in collaborative settings. We conducted a four-day study (approximately an hour each day) in a sixth grade classroom in a large public school in an Indian metropolitan city. We analyzed data of two groups of sixth grade learners. Group A consisted of four participants (three males and one female), and Group B also consisted of four participants (two males and two females). Teams were given a design task of building a balloon powered toy and asked to maximize distance and ensure that vehicle travels in a straight line. For details refer (Kaur & Dasgupta, 2022). Two primary researchers were present in the class as observers along with three teaching assistants (TAs) from the same department. TAs handled the audio and video data collection and also mentored groups from time to time

For data analysis, we first transcribed the video data of the teams and identified uncertainty episodes using a coding scheme developed by Jordan et al., (2014). We analyzed learners' discourse in these episodes to analyze how they notice critical aspects concerning the expressed uncertainty. For analyzing data, we followed inductive and interpretive approach and methods of interaction analysis (B. Jordan & Henderson, 1995). The unit of our analysis is "individual in a group" (Hogan & Fisherkeller, 2005). The authors then engaged in elaborate discussions to resolve the conflicts and reach consensus.

3. Findings

We will now describe how two prominent factors identified during our analysis helped learners in noticing critical information about the problem being solved and the role it played in their uncertainty management processes.

3.1 Role of mentor intervention

We observed that for Team B, Mentor (TA) played a key role in triggering the noticing processes among the team members. For most of the time during the design activity, this team was led by one of the team members named Anmol. He majorly took decisions on behalf of the whole team and other team members readily agreed with him. However, in one of the episodes he alone decided to use a 10mm straw as nozzle. The other team members had previously struggled with the uncertainty regarding which nozzle straw is appropriate for their design, however no body objected or questioned Anmols' decision.

At this point, the TA intervened and asked Anmol why he chose that specific straw, to which Anmol confidently replied, *"Because more air will come out of it [10mm straw] and if we use it, it [toy vehicle] will go for a longer distance"*. TA probed him further asking if *"more air"* or *"less air"* would help them in maximizing their vehicle's distance. This counter questioning by TA elicited doubt in Anmol's mind regarding whether his reasoning is correct or not and he started wondering about his decision. After this moment, other team members also started expressing their viewpoints. Opposite to Anmol's view, other team member, Anu, reasoned that to enable vehicle to travel longer distance, the 8mm straw would be better as, *"air will come out slowly slowly and it will go far."* Another participant, Krish, added to Anu's response saying that "speed" of the air coming out of the straw would also matter.

This episode clearly showed that TA's deliberate interruption and counter questioning helped Anmol notice that he does not have sufficient understanding for making the decision alone. He started accounting other team members' input which eventually initiated discussion among the teammembers. Since the team members had positioned Anmol as the decision maker of the team, they ignored the uncertainty they were struggling with before. However, after TA's intervention, they noticed that the person they positioned as a decision-maker also lacks understanding about which nozzle is better for their design and thereafter they all started to pitch in. They began expressing their viewpoints, posing questions, and engaging in argumentation, indicating that they acknowledged the need to address uncertainty as a collective unit and manage it through sense making processes like questioning, argumentation and logical thinking.

3.2 Role of physical artifacts/materials in the environment

Teams were provided with three different sized straws, out of which they had to choose one to use as nozzle for their design. We observed that both Team A and Team B questioned about the utility of the straws and wondered about why they were of different sizes while they explored the material kit provided to them in the beginning. For example, in case of Team A, seeing the straws kept on the table, one of the participants, Sehaj, asked about why different sized straws were given to them. The observable difference among the three straws triggered uncertainty about how the size of the nozzle mattered for their design. This was followed by elaborate group discussion regarding the expressed uncertainty. However, we observed that although materials in the environment can trigger important questions, they are not enough for sustaining productive uncertainty management actions like analytical discussions, critical evaluation or inquiry processes. Like in this case, another participant Jeet did not shared uncertainty expressed by Sehaj as he thought the size of the straws does not matter and although Sehaj persisted on discussing his uncertainty and even asked for mentor help, the entire team did not delve deeper for resolving his uncertainty.

4. Discussion and Conclusion

This preliminary exploration gave us insights about how mentors' careful and deliberate intervention at the appropriate time and presence of specific task-related materials like the three different sized nozzles triggered curiosity and desirable confusion related to important questions regarding the design problem.

However, it is important that learners are oriented towards the potentially productive information. The analysis showed that learners can perceive same information differently and they do not naturally notice relevant information by themselves. Therefore, there is a need scaffolds learners in a way that they not only notice important information but also find it meaningful. Factors like what material should be provided to the learners, and when and how should mentor/instructor intervene, have to be designed strategically. For our future work, we will dig deeper into the data to find out other triggers of *"noticing"* and more relevant ways of problematizing content for the learners so that they meaningfully notice relevant information and productively use it to manage uncertainties.

References

Chase, C. C., Malkiewich, L., & S Kumar, A. (2019). Learning to notice science concepts in engineering activities and transfer situations. *Science Education*, *103*(2), 440–471.

Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D., & Leifer, L. J. (2005). Engineering design thinking, teaching, and learning. *Journal of Engineering Education*, *94*(1), 103–120.

Hogan, K., & Fisherkeller, J. (2005). Dialogue as data: Assessing students' scientific reasoning with interactive protocols. In *Assessing Science Understanding* (pp. 95–127). Elsevier.

Jordan, B., & Henderson, A. (1995). Interaction analysis: Foundations and practice. *The Journal of the Learning Sciences*, *4*(1), 39–103.

Jordan, M. E., & McDaniel, R. R. (2014). Managing Uncertainty During Collaborative Problem Solving in Elementary School Teams: The Role of Peer Influence in Robotics Engineering Activity. *Journal of the Learning Sciences*, *23*(4), 490–536.

Kaur, N., & Dasgupta, C. (2019). Collaborative and Disciplinary Engagement Levels of the Teams While Managing Engineering Design Uncertainties. *2019 IEEE Tenth International Conference on Technology for Education (T4E)*, 54–60. IEEE.

Kaur, N., & Dasgupta, C. (2022). Investigating experts' and middle school learners' uncertainty management during collaborative engineering design tasks. *16th International Conference of the Learning Sciences (ICLS)*.

Lobato, J., Rhodehamel, B., & Hohensee, C. (2012). "Noticing" as an Alternative Transfer of Learning Process. *Journal of the Learning Sciences*, *21*(3), 433–482.

Manz, E. (2018). *Designing for and Analyzing Productive Uncertainty in Science Investigations*. International Society of the Learning Sciences, Inc.[ISLS].

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Investigating the Role of Students' Feedback-Seeking Actions in Harnessing and Adapting to the Distributed Competence

Narasimha SWAMY, Anveshna SRIVASTAVA & Chandan DASGUPTA
Interdisciplinary Programme in Educational Technology,
Indian Institute of Technology Bombay, India
klnswamy@iitb.ac.in

Abstract: Feedback-seeking actions are student-initiated dialogic actions directed at meeting their diverse needs concerning the teacher's instructional activities. Distributed competence refers to the notion that knowledge, skills, and perspectives needed to accomplish the goals of particular activities are distributed amongst individuals and tools of the community. The focus of this study is in contrast to the predominant educational practices related to feedback and competency development. Existing feedback practices and research primarily focus on what and how of providing feedback to students with only limited attention to facilitating student-initiated feedback-seeking. Besides, feedback interventions targeting competency development focus on improving individual performance by encouraging comparison or competition amongst each other. However, given the rapidly changing human activities where even well-regarded knowledge, skills, perspectives, and tools become obsolete, being proactive and cooperative in harnessing and adapting to the distributed competence takes the center stage. Therefore, we investigate the possible role of students' feedback-seeking actions in harnessing and adapting to the distributed competence. Our findings reveal the processes by which students' feedback-seeking actions contribute toward harnessing and adapting to the distributed competence in a particular instructional activity of significance. The study involved two groups of chemistry students working on a complex representational problem pertaining to the synthesis of an important medicinal drug. The findings are relevant for future research on designing sustainable classroom instructions and CSCL environments.

Keywords: Case study, undergraduate chemistry education, student agency, feedback as dialogue, feedback-seeking, distributed competence, cultural-historical activity theory, and microgenetic analysis.

1. Introduction

Prevailing educational practices related to feedback and competency development are largely informed by behaviorist and cognitivist perspectives (Hodges, 2018; Lipnevich & Smith, 2018). Within these practices, the focus is predominantly on providing unilateral feedback to students with only limited attention to facilitating student-initiated feedback-seeking. Effects of these feedback-provision-interventions are found to be inconsistent and at times even negative (Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Shute, 2008). Consequently, these feedback provision focused practices are considered unsustainable and detrimental to students' agency (Boud & Molloy, 2013). Moreover, the root cause behind these problems is traced to the lack of appropriate theory and gaps in key interconnected notions such as instruction, feedback, and competence. Existing ideas of these notions do not account for multiple cultural factors such as student diversity, teacher workload, and shortage, competence in the context of teams, evolving tools and workplace conditions, and the scale of educational challenges before governments. Besides, the interplay of cognition, emotion, motivation, and identity aspects of students also remains ignored. Against such a backdrop, we identified cultural-historical activity theory (CHAT) as a valuable theoretical lens.

CHAT is an interdisciplinary approach currently studied and applied in both education and workplace contexts (Engeström, 2018). It views any human activity as a goal-directed collective system mediated by cultural factors such as tools, representations, norms, community, and the nature of

cooperation. It recommends viewing any chosen human activity as situated within a web of interconnected human activities. If we consider instructional activity, then its practices have to be viewed in relation to other interconnected activities such as students' families, educational institutes, industries, and government. This is in contrast to standard research approaches viewing instructional activities in isolation by black-boxing many other interconnected societal issues. As a result, CHAT helped us consider and address multiple issues discussed previously. Given below is the reconceptualization of key notions followed by the discussion on how CHAT helped inform them.

- The instructional activity is a teacher's predetermined plan involving the organization of certain tasks, tools, representations, norms, individuals, and ways of cooperation (division of labor) estimated to support students' learning towards particular valuable outcomes.
- Feedback is any information concerning the given instructional activity that addresses students' needs tied to their knowing, doing, or being for accomplishing, modifying, or even avoiding the aspects of instruction. Students can derive feedback either by observing or working with tools, representations or having a dialogue with themselves and multiple others or using a combination of these modes.
- Students' feedback-seeking actions are those actions where decisions such as purpose, timing, mode, amount of feedback-seeking, choice of feedback source, and how feedback is used are all made by themselves.
- Competence is an evolving set of interconnected knowledge, skills, and perspectives needed to accomplish the goals of activity or cluster of activities grounded in relevant societal contexts. Aspects of competence are distributed within and beyond the elements of instructional activity.

In the above definitions, the idea of instruction acknowledges the complexity of any goal-directed human activity as informed by CHAT and does not reduce it to just tasks or talk like a set of problems or lectures. Further, instruction as an estimate and predetermined plan accounts for classroom realities such as student diversity, teacher workload, and large pupil-teacher ratio. Consequently, we also do not view teachers' role as orchestrating instruction and feedback on a moment-to-moment basis. Then the idea of feedback acknowledges that it can be any information that meets students' diverse cognitive, emotional, motivational, identity, and behavioral needs. Especially, viewing feedback's role as not always accomplishing the goals of instruction aligns with the notion of instruction as an estimate where it may or may not address students' diverse needs. This along with the idea of students' feedback-seeking extends the scope for students' agentic actions. That is, students can choose to seek feedback from multiple sources and through multiple modes for accomplishing, modifying, or even avoiding the aspects of instruction. Such conceptualization aligns with CHAT's perspective on viewing learning as a joint transformation of not just individuals but also activities (Engeström, 2018). This is also unlike many prevailing practices where teachers initiate and drive feedback events even when peers exchange feedback.

The focus on students' feedback-seeking aligns with Nicol's (2013) emphasis on getting students ready for life beyond university, like occupational settings where feedback seldom comes from others proactively. Instead one has to identify potential sources from the community and elicit feedback from them. Also, feedback received in such a manner can be vague, ambiguous, and even conflicting. Hence, feedback received has to be subjected to rigorous evaluation. Compare this with the formative feedback guidelines in education which require an external agent such as a teacher to continuously monitor student performance, simultaneously account for various student and task characteristics, and accordingly adapt the content, mode, timing, and frequency of the feedback messages to be provided (Shute, 2008). Lastly, the idea of competence acknowledges that it is not an isolated set of stable skills but one that can also become obsolete with the changes in particular activities within which it is situated. This is because as activities change by incorporating novel tools, representations, norms, and ways of cooperation; some of the prevailing knowledge, skills, and perspectives may become redundant. Moreover in simple, routine, and standard activities competence may appear as situated within an individual or tool, however, in complex activities, it is very often distributed amongst multiple individuals, tools, and artifacts (Engeström, 1992; Rogoff, 2003). Together, these notions help inform the investigation of our research question i.e., How do students' feedback-seeking actions help them harness and adapt to the distributed competence?

2. Methods

To answer the above research question, we identified and gave students a representational problem adapted from Wong, Sultana, & Vosburg, 2010. It pertained to the synthesis of an important medicinal drug 'Warfarin' (refer to *Figure 1*). Choice of the problem was informed by the CHAT's view that 'knowledge worth learning is continuously evolving'. While there were other methods of synthesis, the particular technique underlying the chosen problem is precise, cheap, fast, and environment-friendly. The underlying technique 'asymmetric organocatalysis' is well-recognized in chemistry for its significance in current and future pharmaceutical research and green chemistry.



Figure 1. Chemical reaction pertaining to warfarin synthesis and 3D printed molecular models

Solving the problem required students to work with various tools such as 3D printed molecular models, model construction kits, and a molecular visualization application. Students also had to employ major forms of representational reasoning like analogical reasoning and/or thought experimentation with representations (refer to *Figure 2*). Students had to interpret complex symbolic representations of chemical reaction steps, determine the spatial arrangement of the intermediate structure formed based on their choice of catalyst and then finally predict the expected drug's spatial arrangement by applying stereochemistry concepts. Here the expected drug's spatial arrangement is most crucial as its therapeutic effectiveness is tied to it. The nature and complexity of the problem meant that students would experience numerous challenges nudging them to go beyond just verbal exchange and use or build the molecular models and sketch multi-perspective diagrams as part of their feedback-seeking dialogue. Thus the given problem helped us generate rich analyzable data where we could examine how students' feedback-seeking actions helped harness as well adapt to the competence distributed across peers, tools, and other activities.

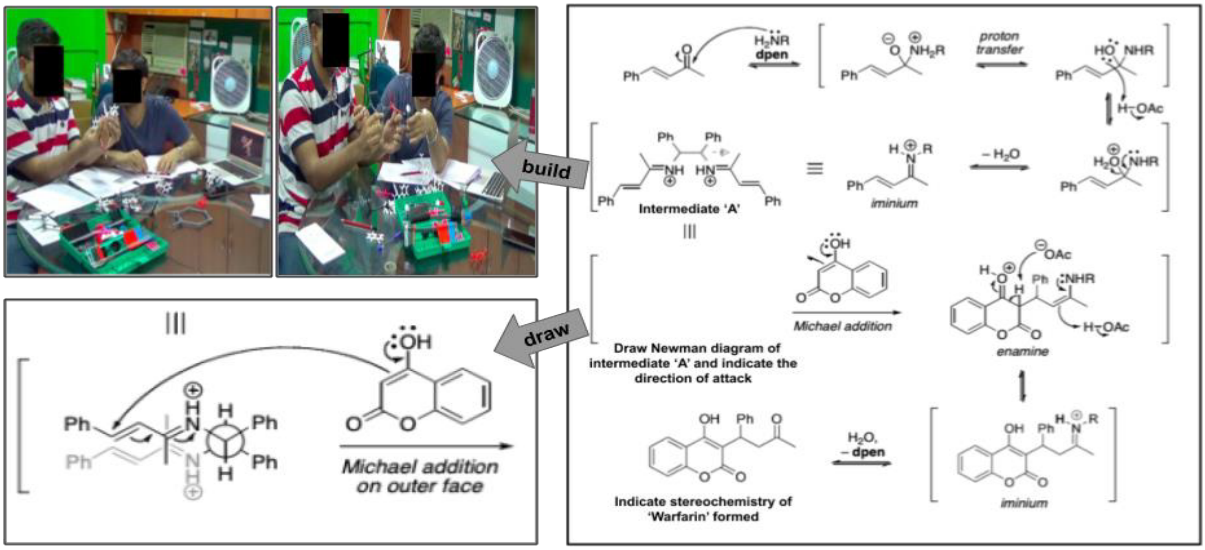


Figure 2. Reaction mechanism of warfarin synthesis adapted from Wong, Sultana, & Vosburg, 2010.

We adopted the case-study approach as it helps us answer our explanatory research question by providing rich and insightful descriptions (Yin, 2009). It affords going beyond exploration and diving

into the context-sensitive description of why things happened the way they happened. For data collection, we identified two groups of chemistry students using convenience sampling within our institute. One group had two male students and another had three male students. A prerequisite for selection was that they should have completed an introductory undergraduate stereochemistry course. The study was conducted with each group separately in a lab setting. In a post-activity interview, both groups reported that they had not come across any similar problems in their regular courses.

The given problem was an hour-long activity designed such that students could either work individually or choose to collaborate as and when needed. This was to ensure that there was no external compulsion to seek or provide feedback. The same was explicitly mentioned in the worksheets of instruction given to them. Besides, worksheets were minimal in design i.e., we intentionally did not include any prompts for self/peer assessment, hints, or questions for reflection. Worksheets given to each student had the following details and tasks in sequence. The first page included information on stereo chemicals as medicinal drugs in general and warfarin in particular, then on the need for selective synthesis of such medicines, the importance of underlying concepts and skills for their future careers, also task-specific expected goals and definitions, and a list of available tools along with the expected nature of their participation. The second and third pages were on identifying, naming, and translating 3D models of warfarin and dpen organocatalyst into 2D diagrams as per chemical conventions. The fourth and last page of the worksheet had a laboratory procedure underlying warfarin synthesis and its reaction mechanism with a missing step that students had to figure out.

Data collected as part of the study included video recordings of the students' interactions with each other and the researcher and their task-specific responses on worksheets such as sketches, highlights, and comments. The video was recorded using a single camera placed at a sufficient distance away from the round table where participants were seated. Participants could move around the table freely to assist each other if needed. The researcher's presence during students' problem solving was kept limited and occasional. Data analysis was done using a microgenetic method. It emphasizes tracing the emergence of particular student actions to momentary processes by making high-density observations spanning the timescale of the event concerned (Chinn & Sherin, 2014). The choice of method aligns with the CHAT framework's need to temporally examine how feedback-seeking emerges, evolves, and ends. Another reason was the duration of feedback-seeking episodes which varied between a few seconds to minutes. Observations were not limited to actions of a student seeking feedback, they also included what others were doing.

To arrive at key inferences, we employed competitive argumentation with colleagues where sample data was collaboratively analyzed to rule out possible alternate explanations. However, there were differences amongst us with respect to the emotions identified. So we reflected upon the sources of differences and found that they were mainly arising due to differences in our noticing of particular behavioral aspects and attribution. For example, some were attending to a student's facial expression while others were attending to their body posture. So we put together a list of parameters tied to emotions from literature such as behavior considered i.e., specific verbal and nonverbal aspects, the intensity of the emotion, the time dimension of emotion i.e., whether it was elicited in retrospective e.g. relief, or as prospective e.g. anxiety. These parameters helped us resolve differences.

The unit of analysis is an episode of students' proactive feedback-seeking where all decisions tied to it should have been made by the students themselves. One episode is distinguished from another by considering the change in the purpose for which the feedback is sought. We used empirical referents such as verbal questions addressing peers, moving closer to the feedback source while thinking aloud, pointing at and moving the artifact closer to the source while explaining, and drawing the attention of the source by tapping a pencil or raising a voice for identifying an episode's starting point. Similarly, the episode's endpoint was signified by referents such as nodding head in approval, verbal acknowledgment, moving away from the source, source getting back to his/her task, or taking back artifacts. For identifying factors comprising instructional activity which influenced students' different feedback-seeking actions, we examined aspects tied to task, artifacts, student, and feedback sources by going back and forth a few seconds before and after the episode.

3. Findings

3.1 Role of Feedback-Seeking Actions in Harnessing Distributed Competence

Throughout the instructional activity, students proactively sought feedback from each other for diverse purposes and engaged in varying amounts of dialogue depending on the characteristics of tasks and peers. They also used multiple languages and modes for feedback-seeking and subjected the feedback received to considerable evaluation. Particularly, in line with the notion of distributed competence, it was observed that even within a single task the role of feedback-seeker and feedback-provider kept shifting among the group members. For instance, consider the following observations from a 3 member student group while they were working on determining 3D warfarin molecules' stereochemistry and sketching their 2D diagrams as per the required chemical convention.

After a few moments into the above task, student S1 (redshirt in *Figure. 3*) was observed holding and comparing the models of two warfarin molecules in one hand and sketching their abbreviated forms in another hand. Besides, while sketching S1 was also thinking aloud "R1..R2..R3…so…1, 2, 3 and hence S". Here, by R1, R2, R3 student is referring to different parts of the molecule and then by 1, 2, 3 he is assigning a certain priority to those parts. 'S' indicates the stereochemistry of warfarin. Immediately after uttering those words, S1 taps on student S2's (yellow shirt in *Figure. 3*) arm and invites his attention. As S2 pays attention, S1 repeats his thinking while pointing at different parts of the molecule and asks "Is this fine?". S2 responds by giving confirmation. Later, S1 continues to seek feedback from S2 intermittently for different purposes i.e., once to confirm whether a particular atom in the molecule is oxygen and at another instance, he invites comments on his sketches of warfarin molecules. In relation to the same task a few moments later, S2 was observed to seek feedback from S1 for ensuring procedural accuracy needed to assign group priority within warfarin and S3 sought feedback from S2 on his 2D sketches.



Figure 3. Student S1 (red shirt) seeking feedback from student S2 (Yellow Shirt).

The above observations suggest that although S1 was able to recall the needed knowledge and procedures from his prior learning, he was still very unsure about it. For S2 to be in a position to confirm S1, he should have been able to successfully recall and be certain about it. Therefore, S1 here is harnessing S2's competence of being able to successfully recall the specific knowledge and procedure with a certain confidence that was needed at the moment. Next, when we investigated how S2 came to be in a position to confirm S1's other feedback-seeking instance i.e., on confirming whether a particular atom of warfarin was oxygen, we found that during the initial orientation each student had attended to different aspects of the information told by the researcher. Each student also asked different questions to the researcher regarding specific molecular models and their particular features. S2 had then raised questions on the warfarin model and its specific features including oxygen to which the researcher had responded. Consequently, it put him in a position to confidently confirm S1's question on oxygen. A similar observation was also made in the subsequent task where S1 acted as a feedback source to S3's question on catalyst molecules. This was again mapped back to S1's initial question on catalysts to the researcher during orientation. Thus it could be said that during orientation participants' attentional resources were distributed and later while working on the given tasks they were harnessing their results from each other. Then regarding S1's invitation to S2 and S3's invitation to S2 for checking their sketches, it could be said that it helped them both harness S2's monitoring capabilities. Additionally, S2's feedback-seeking from S1 by revealing his reasoning and eliciting S1's reasoning as a response

helped S2 rectify a crucial error in the priority assignment. Overall, the above feedback-seeking actions helped students to harness the distributed nature of each other's attention, memory, reasoning, and monitoring resources.

Similar to the distributed nature of competence amongst individual students, the competence needed to accomplish a task can also be distributed across multiple artifacts either provided by the instructor or those created by the students. Therefore, harnessing competence distributed across relevant artifacts becomes as crucial as harnessing competence distributed across individuals. Consider the following interactions from a 2 member undergraduate student group depicted in *Figure 4* below. The depiction lists the sequence of artifacts i.e., different molecular models and sketches referred by student S1 (feedback-seeker) and student S2 (feedback-provider) before and during feedback-seeking. These interactions emerged when both students were trying to make sense of the reaction mechanism behind warfarin synthesis.

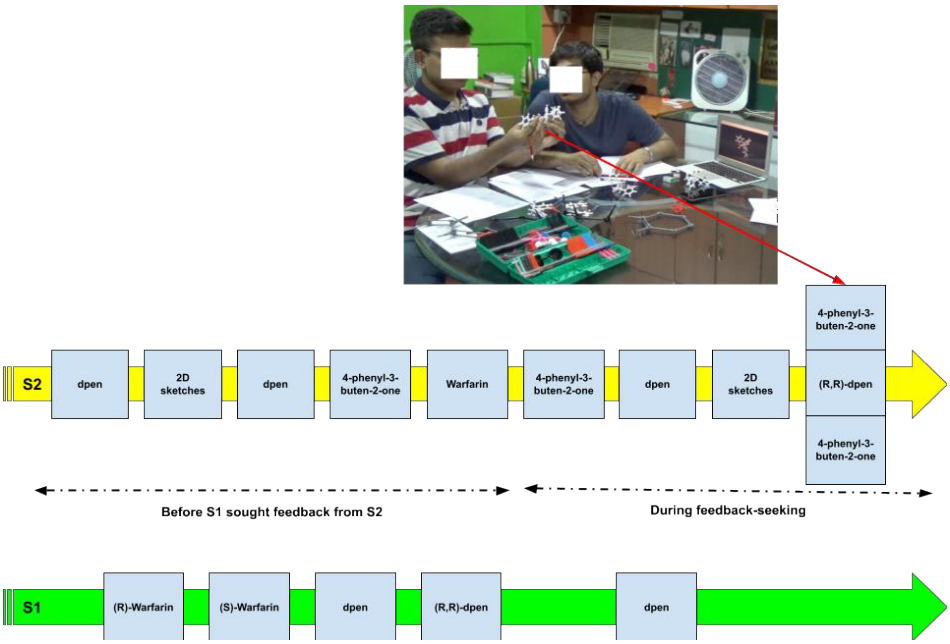


Figure 4. The sequence of artifacts referred by students S1 (blue shirt) and S2 (striped shirt).

The above depiction shows that by the time S1 sought feedback, S2 had already referred to different individual molecular models i.e., dpen and phenyl-buten-one which he would later use to build an intermediate structure required to answer S1's question on catalyst: What does the dpen do? and where does it attack? In comparison to S2, S1 was observed trying to make sense of the reaction mechanism by referring to just one model i.e., dpen tied to the intermediate structure. In the reaction mechanism, the intermediate structure was formed by two molecules of 4-phenyl-3-buten-2-one and one molecule of 'dpen' catalyst with a unique spatial arrangement. In the above depiction, letters (R) or (R, R) and (S) or (S, S) before individual molecular model names refer to the spatial arrangement or stereochemistry of atoms and groups of atoms within that molecule. Importantly, we also observed that just before feedback-seeking S1 was experiencing discomfort or difficulty as indicated by his facial expression and gestures, whereas S2 was observed uttering a few self-directed words followed by humming a tune of some song. S2's utterances and humming tune probably signaled feedback availability to S1. These observations suggest that the emergence of the feedback-availability state in S2 and the feedback-non availability state in S1 can be attributed to the nature in which students were harnessing the competence or the relevant information and affordances distributed or encoded across multiple models. Thus feedback-seeking at the self-level involving interactions with artifacts can be as crucial as feedback-seeking at an interpersonal level for harnessing distributed competence. This feedback-seeking episode was followed by a state of feedback non-availability amongst both participants for a few minutes. This was in relation to making sense of the subsequent part of the reaction mechanism. During this feedback non-availability state, S1 tried to sketch the missing reaction step whereas S2 spent time reflecting on possibilities in the next reaction step. However, the act of sketching led to the emergence of feedback availability in S1, and consequently, S2 was observed seeking

feedback from S1. Here again, S1 can be said to have harnessed the competence that emerged with the artifact he produced i.e, the sketch of the missing reaction step.

3.2 Role of Feedback-Seeking Actions in Adapting to the Distributed Competence

Within an episode of feedback-seeking dialogue, although one student largely assumes the role of feedback-seeker and other/s as feedback-provider/s, it was observed that the feedback-seeker at times points to errors or adds to the missing information in the feedback received or even verifies it by consulting alternate sources. Also, the student seeking feedback was observed to help the feedback-provider by doing certain work like sketching or finding a particular molecular model needed to generate the feedback. So feedback-seeking is not a simple one-way reception of answers from feedback providers. Instead, it involves both feedback-seeker and provider supplementing or at times complementing each other's competence i.e., adapting to the distributed competence amongst them. Table 1 presents a few excerpts from a feedback-seeking episode in support of our argument.

Table 1. Excerpts from a Feedback-Seeking Episode Reflecting Adaption to Distributed Competence

Sl. No	Excerpts
1	Feedback-seeker (S1) pointing to error in feedback received from S2 **S1**: attacking on what? **S2**: this one (pointing pencil alternately at reaction mechanism and model) so dpen will attack here on carbonyl directly.. carbonyl addition **S1**: where ..dpen? **S2**: this is dpen in my hand (turns the model in hand towards S1) **S1**: No dpen is this one (pointing at another model) **S2**: Haan (yes) sorry
2	Feedback-seeker (S1) adding to the missing information in feedback received from S2 **S2**: See this is intermediate..the attack is happening on this carbon **S1**: Intermediate? add the other one too. Looks for the missing molecular model in intermediate structure and gives to the S2
3	Feedback-seeker (S1) trying to verify the feedback provided by S2 with alternate source **S2**: Attack has to happen on this carbon from behind…pointing at the intermediate structure that he put together **S1**: Why will it attack from behind? **S1**: Moving the laptop in which the molecular visualization application was open..asks the researcher "Can we see the intermediate in this?"

In addition to the above instances, we also observed that students were revising their mode of feedback-provision and feedback-seeking within an episode to get their message across. For example in one episode, initially, the mode of feedback-seeking and the mode of feedback-provision involved just the verbal exchange but then the student seeking feedback remained unconvinced with the feedback received. So he kept probing the feedback-provider verbally. As a result, the peer providing feedback adapts one's mode of feedback-provision. He first uses the model to convey one's feedback and later sketches the 2D diagram to convey the same, recognizing that the former did not serve the purpose. In response, the student seeking feedback still unconvinced builds a model to convey one's reason and goes on to elicit more feedback. Importantly, these transitions were fraught with emotions such as irritation and frustration. Here students were trying multiple modes of feedback-seeking and provision unsuccessfully. Such unsuccessful attempts nudged them to adapt to each other's competence by complementing their mode of feedback-seeking and provision.

Other than adapting to distributed competence at the interpersonal level, students also learn to harness and adapt to the competence incorporated within particular tools. This is because tools like the model construction kit provided to students to an extent constraint student actions in a desirable

direction and help reduce errors in their prediction. However, a student in a 3 member group tries using the model construction kit for a while and then gives up saying that it's too complicated. Later in the post-activity interview, he reveals that particular software that translates 2D chemical structural diagrams to 3D molecular would have been much easier for the same task. So letting students consider and even try alternate tools during instructional activity can help them better harness and adapt to the competence distributed across multiple tools.

3.3 Barriers to Harnessing and Adapting to the Distributed Competence

Both the nature of student interactions during instructional activity and their post-activity interview responses made us aware of various barriers which students might experience in harnessing and adapting to distributed competence through feedback-seeking. These barriers included issues such as the language of communication, group composition, and size, nature of cooperation, access to alternate tools, opportunity to consult non-group members, and also the presence of an instructor.

For instance, consider the following interactions of a two-member student group which occurred after almost 7 minutes into the instructional activity. At this particular moment, S2 pointing at specifics of a molecular model and then looking at S1 asks him in Hindi (a native language) “this one here is tetrahedral center right?”. Until then the conversations were in English only. In response to S2’s question, S1 first looks at where S2 is pointing and then instead of answering him looks at the researcher and asks hesitantly “can we talk in Hindi right? Can we talk in Hindi? it’s….” Researcher responds “okay, yeah yeah” before S1 completes his sentence. But then a moment later asks them about their difficulty in speaking English. For which S2 responds by saying “no.. he is” suggesting that it is S1 who has the difficulty and S1 confirms by saying with an embarrassed smile “I am comfortable in Hindi so”. Finally, when the researcher tells them to speak in whichever language they are comfortable with, S1 responds with a relieved smile that he will go with Hindi.

In the above interactions, S1’s timing of the question on language and emotions reflects his heightened sensitivity towards the choice of language that he can or cannot use in instructional settings. Origins of such heightened sensitivity can be traced to the popular practices in educational institutions here where the use of native languages is often restricted and any non-adherence is treated with humiliation in order to push students to learn English as it is viewed as key to access better employment opportunities in future. However, compelling students to interact in a language that they are not comfortable with might significantly hinder their feedback-seeking and hence can act as a barrier to harnessing and adapting to the distributed competence. Similarly, in a 3 member group, students were observed to spontaneously shift between 3 different languages. Besides, in the post-activity interview, they also expressed that they wish to have the freedom to speak in any language of choice.

Closely related to the above language issue is the group composition. Students in the interview expressed that they would like to have the freedom to choose group members. In the context of this study, it translates into students being able to choose potential feedback sources. Underlying reasons included members not acknowledging the differences in language backgrounds i.e., within a group, few members would keep conversing in a particular language that other members do not understand. Besides, reasons also included issues tied to cooperation arising due to differences in motivation i.e, not all members will be ready to invest the time and effort equitably. As reported by one student, the difference in motivation itself arises because different student groups in the classroom are aiming for different future prospects with respect to academics and careers. So some might be more interested in particular topics than others. Additionally, one student expressed the need for having more members in the group in comparison to the current group size. Therefore, negotiating these issues will be crucial for facilitating students to harness and adapt to distributed competence.

Regarding the researcher’s questions on whether they should have been provided with the opportunity to consult non-group members, students expressed that such an opportunity would have been very helpful. Moreover, they revealed that they would have approached a few seniors working in an organic synthesis lab as they would have better knowledge about the topics such as the one given to them. Then with regard to questions on the presence of a researcher or teacher, all students expressed in agreement that the limited or occasional presence of the teacher or researcher is sufficient. They said that it helped them discuss with each other much more freely than if the researcher or teacher were to be present throughout the activity. Also, as previously discussed students expressed that they should be free to access and try alternate tools during the instructional activities.

Finally, whether a student perceives the given activity as a learning event or a testing event may have significant implications for feedback interactions within a group. For instance, we observed in one group that a student who perceived the instructional activity as a testing event was reluctant in providing feedback initially. He even expressed irritation when other members started discussing. This occurred due to the norm of instructional activity which stated that students can either choose to work individually or collaboratively. Hence this student perceived that the given activity might be a testing event. However, this observation does indicate that the current trend of merging learning events with testing events focused on evaluating individual competence might act as a hindrance to feedback interactions amongst students.

4. Discussion and Conclusion

This study provides preliminary evidence on the role of students' feedback-seeking in harnessing and adapting to the competence distributed across peers and tools in a particular instructional activity of significance. Findings included observations made at the intersection of student-student interactions and student-tool interactions. These observations helped reveal that distributed competence is the natural outcome of distributed nature of students' attention, memory, reasoning, and regulation. Similarly, observations also reveal that the competence needed to accomplish a task can be distributed across multiple tools and student-generated artifacts. This is again because tools are materials designed to embody certain human actions to constrain user actions and thinking in desirable directions. Findings also show that students' feedback-seeking is not a simple one-way reception of answers but it is dialogic and involves both feedback-seeker and provider supplementing or at times complementing each other's competence. Such dialogic interactions help nurture adaptation to the distributed competence. Besides, the findings highlight that feedback-seeking at the self-level or trying to generate the needed feedback by thinking or working with tools individually could be as important as feedback-seeking at an interpersonal level. Hence giving students the freedom to regulate shifting between feedback-seeking at the self-level and interpersonal level might be desirable.

Overall, when the above findings are considered together, it tells us that multiple factors such as tools, representations, norms, community, and the nature of cooperation within the immediate instructional activity as well as larger culture influences students' feedback-seeking actions. More importantly, these findings would not have been possible without addressing the gaps in the key interconnected notions such as instructional activity, feedback, students' feedback-seeking actions, and competence. Reconceptualization of these notions as per cultural-historical activity theory helped us with the much-needed perspective to observe and make sense of the student interactions. Also, note that the reconceptualization of these notions holds for both classroom instruction and CSCL environments. However, more research is needed to establish the role of feedback-seeking actions in not just harnessing or adapting but also for further nurturing distributed competence amongst students in actual classroom settings. This can potentially help inform student-agency-oriented sustainable alternatives to the prevailing largely teacher-driven approaches to feedback provision and students' competency development. In this regard, findings that particularly hold interesting insights for future research is that students' feedback-seeking can also occur due to dynamically emerging feedback availability cues amongst others working nearby. Also, recall students' intention to seek feedback from seniors during the interview. That is, if student groups of different academic years were working in a single classroom on activities of varying complexity and overlapping learning goals as shown in *Figure 5* below, then their various actions might help signal and elicit each other's feedback-seeking on much more diverse purposes. Similar affordances can also be built into CSCL environments connecting activities and students of different academic years or proficiencies. Here note that the current age-segregated classrooms were primarily informed by cognitive perspectives which viewed biological maturation as a precondition for learning. However, sociocultural approaches emphasize the developmental significance of being surrounded by activities far beyond one's capabilities along with the opportunities for naive participation in them (Rogoff et al., 2005).

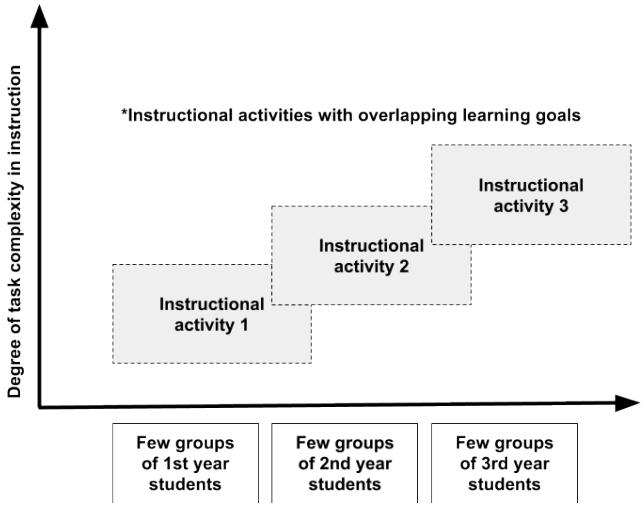


Figure 5. Classroom design for sustainably facilitating students' feedback-seeking actions

In classroom designs such as above, students can draw feedback from multiple sources and also through multiple modes. That is students can derive feedback either by observing multiple others or having a dialogue with them for various purposes. Finally, another important direction for future research is to investigate further and systematically address various barriers to students' feedback-seeking along their cognitive, emotional, motivational, and identity dimensions.

References

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10/gcphxw

Chinn, C. A., & Sherin, B. L. (2014). Microgenetic Methods. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (2nd ed., pp. 171–190). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.012

Engeström, Y. (1992). *Interactive expertise: Studies in distributed working intelligence*. Dept. of Education, University of Helsinki.

Engeström, Y. (2018). Expansive learning: Towards an activity-theoretical reconceptualization. In *Contemporary Theories of Learning* (2nd ed.). Routledge.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10/bf4d36

Hodges, B. D. (2018). The Shifting Discourses of Competence. In B. D. Hodges & L. Lingard (Eds.), *The Question of Competence* (pp. 14–41). Cornell University Press. https://doi.org/10.7591/9780801465802-005

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254–284. https://doi.org/10/gtw

Lingard, L. (2018). Rethinking Competence in The Context of Teamwork. In B. D. Hodges & L. Lingard (Eds.), *The Question of Competence* (pp. 42–69). Cornell University Press.

Nicol, D. (2013). Resituating feedback from the reactive to the proactive. *Feedback in Higher and Professional Education*, 34–49.

Rogoff, B., Correa-Chávez, M., & Cotuc, M. N. (2005). A Cultural/Historical View of Schooling in Human Development. In D. B. Pillemer & S. H. White (Eds.), *Developmental Psychology and Social Change* (1st ed., pp. 225–263). Cambridge University Press.

Rogoff, B. (2003). Distributed Cognition in the Use of Cultural Tools for Thinking. *The cultural nature of human development*. Oxford University Press.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10/bd5cgq

Wong, T. C., Sultana, C. M., & Vosburg, D. A. (2010). A Green, Enantioselective Synthesis of Warfarin for the Undergraduate Organic Laboratory. *Journal of Chemical Education*, *87*(2), 194–195. https://doi.org/10/fkd2r6

Yin, R. K. (2009). *Case study research: Design and methods* (4th ed). Sage Publications.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Understanding Learners' Negotiation Processes During Ill-Structured Engineering Estimation Problem-Solving

Ulfa KHWAJA*, Sahana MURTHY
Indian Institute of Technology Bombay, India,
*ulfa_khwaja@iitb.ac.in

Abstract: Estimation of physical quantities is an important practice in engineering. However, estimation problems are ill-structured and complex, and learners require support to solve them successfully. Modelling Based Estimation Learning Environment (MEttLE) is an open-ended learning environment (OELE) designed to support undergraduate engineering students in estimation problem-solving based on disciplinary model-based reasoning tasks and metacognitive scaffolds. When learners collaboratively solve estimation problems in MEttLE, its open-ended pedagogical design implies that different learners may have different approaches to reach a required common ground and learners' negotiation becomes crucial. This study aims to understand where, why, and how learners negotiate and how it influences estimation problem-solving. We analyze learners' problem-solving approach through their collaborative discourse and actions as they work on MEttLE to build and evaluate models to estimate a physical quantity. Our findings indicate that learners' negotiation strengthens the processes that underlie good estimation problem-solving.

Keywords: Negotiation, disciplinary model-based reasoning, open-ended learning environments, collaboration

1. Introduction

Estimation is an ill-structured problem involving determining approximate values for a physical quantity without access to complete information (Jonassen, Strobel & Lee, 2006). Making estimates is an important and frequent practice used by engineers and scientists when exact data and precise governing equations are unavailable (Mahajan, 2014). Estimation problems are challenging and hard to solve for learners as they involve applying knowledge from different topics to a real-world system, making assumptions, identifying the important physical quantities and those that can be neglected, and making judgments regarding numerical values (Jonassen et al., 2006). Prior research has shown that the processes underlying good estimation performance include model-building, model contextualization, evaluation, planning, monitoring, and revising the problem-solving approach (Kothiyal & Murthy, 2020). To learn estimation problem solving, learners need to be explicitly scaffolded through activities that allow them to engage in the above processes.

In this paper, we study learners' problem-solving processes when they collaboratively work on MEttLE, an open-ended learning environment (OELE) designed to support estimation problem-solving based on disciplinary model-based reasoning tasks and metacognitive scaffolds (Kothiyal & Murthy, 2020). MEttLE's pedagogical design requires learners to explore, grapple with incomplete information, make assumptions and decide the solution approach while they solve the estimation problem. The open-ended nature of these activities can be demanding for learners (Emara, 2018), and collaboration can aid learners in solving such tasks (Kirschner, Paas, & Kirschner, 2009). However, students' collaborative learning processes in approaching ill-structured estimation problems and their impact are not well understood. Learners may have different approaches or conflicting ideas while maintaining a common shared goal because of the open-ended nature of the tasks. Thus, learners' negotiation of which approach to choose, which concepts are relevant and why, and which resources to use become crucial in understanding their problem-solving processes. This study aims to understand learners' negotiation

processes and strategies as they solve an engineering estimation problem related to power in MEttLE and its impact.

2. Theoretical Framework

Negotiation is crucial in CSCL research (Carell & Herrmann, 2009) as effective collaboration requires negotiation (Dillenbourg & Baker 1996), especially when learners face differing opinions, ideas, meanings, and concepts from their partners. It is most relevant and valuable in complex problems or tasks that require collaborators' ideas, contributions, and critical examinations to find a suitable solution by pursuing convergence (Carell & Herrmann, 2009). Thus, making it a crucial part of any supported method for complex problem solving (Richardson & Anderson, 1995). Negotiation is defined in multiple ways, such as resolving conflicts (Dillenbourg & Baker, 1996), an attempt to agree on the goals, actions, and beliefs that comprise the planning and acting decisions of the collaboration (Sidner, 1994), and a critical examination of reasons for and against different points of view (Baker, 1994, pp.200). While negotiating, learners must put forth their points and defend their positions using strategies such as argue, justify, refine, standpat, etc. (Baker, 1994, pp.200). Negotiation on a common ground is essential to ensure that all have argued and agreed upon a common understanding. Understanding how this takes place is important in problem-solving (Beers, 2006).

To analyze learners' negotiation strategies and processes, we have applied codes from social modes of co-construction (Weinberger & Fisher, 2006). These processes include Externalization (EX), Elicitation (E), Quick consensus Building (Q), Integrated oriented consensus building (I), and Conflict-oriented consensus (C) building. Externalization is articulating thoughts without reference to others' contributions and making one's knowledge external. Elicitation is questioning the learning partner to seek help, verification, and clarification to attain the same understanding of the content the partner intended to share. Once this knowledge is understood, it is contested for acceptance or rejection, followed by negotiation of common ground. Quick consensus building is accepting the contributions to move ahead in the task. Here agreement is merely for coordinating the discourse move, and a true common ground may not have been reached. Integrated-oriented consensus building is integrating and applying the perspectives of the collaborators. In this process, there stands a chance when one may give up their initial belief based on a peer's contribution. Conflict-oriented consensus building is working on the reasonings of peers by disagreeing, modifying, or replacing their perspectives. During this, learners criticize the proposals and find better arguments for their position.

2.1.1 Modelling-Based Estimation Learning Environment (MEttLE)

Modelling-Based Estimation Learning Environment (MEttLE) (Figure 1) is an OELE for estimation problem solving that triggers a model-based estimation process. It is structured into five sub-goals (Figure 1a). The three-phased model building sub-goals - functional, qualitative, and quantitative- help create, contextualize and evaluate models of complex problems essential to solving estimation problems, one calculation, and one evaluation. It has metacognitive prompts for evaluation, planning, monitoring, and reflection of models and problem-solving processes, simulators (Figure 1b), hints, an info center, guide me, question prompts, and a problem map that affords the modelling process. Novices can choose any path and revisit any sub-goals at any time in MEttLE. The problem solving is complete only when they have evaluated the reasonableness of their calculated estimate of power (Figure 1c). If the evaluation fails, they may need to reflect on their estimation by revisiting previous sub-goals.

In this study, to solve the problem, the solver must estimate a physical quantity (power of car), understand the problem system (the car), analyze how it behaves under the given operating conditions, determine dominant parameters for these conditions, and create an equation involving the dominant parameters. He has to make assumptions and approximations to create and determine a simplified model of the problem context and its corresponding mathematical equation.

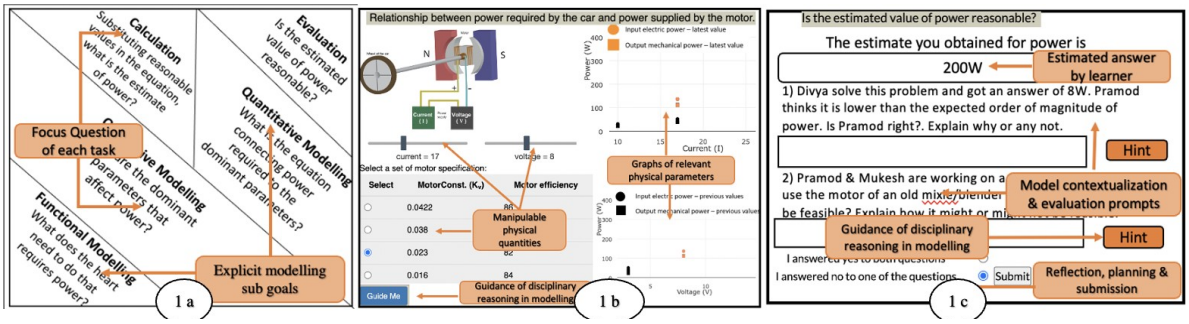


Figure 1. Overview of MEttLE Interface. (1a) Estimap with sub-goals; (1b) Simulator-4, and (1c) evaluation sub-goal with model contextualization prompts

3. Methods

3.1 Participants

5 females and 9 males of 3rd or 4th-year Mechanical and Electrical undergraduates from different engineering colleges in Mumbai participated in this study. Students from these disciplines were chosen because the problem in MEttLE had concepts that required knowledge specific to the disciplines. A google form to gather interested participants' demographic details and availability was circulated among students by their teachers. Only those who were double vaccinated were considered for this study. IRB approval from our institution was sought.

3.1.1 Procedure and Data collection

Seven groups of two students each visited the lab on different days for the study. The researcher explained the study's goal, the procedure, and the data to be collected, after which participants signed the consent form. The lab setup was of a single computer system with MEttLE. The researcher explained the collaborative script where one learner would be the "driver" and control the mouse and keyboard to interact with MEttLE. At the same time, the other would be the "navigator" and would question, argue, and encourage the driver to elicit his/her thought process. These roles would be switched every 20 minutes. They were asked to articulate their thoughts and views explicitly during collaboration. On average, the time taken by each pair to interact with MEttLE was about 90 minutes.

Data were collected using five sources (Screen recording, Video camera, Observational logs, Interviews, and paper-based tests). We used only two for this research. The primary source was the recording using the OBS studio (OBS, 2021), which ran on the computer containing MEttLE and captured learners' interaction with it, audio, and front video. The secondary source was the semi-structured interview where the researcher played the screen recording of the learners' interaction and asked questions to get a deeper insight into why learners did or said something during the intervention.

3.1.2 Data Analysis

Researchers viewed the recordings multiple times and used Baker's model of negotiation (Baker, 1994) to identify the segments of negotiation. These were transcribed to include the verbal and interaction data (such as entering text, clicking, changing values, etc.) with MEttLE and the facial, tone of voice, and learners' gestures wherever necessary. We then applied the codes from the framework for social modes co-construction (Weinberger & Fischer, 2006) to identify learners' negotiation strategies. Two researchers independently coded two episodes and attained inter-rater reliability of 0.945 Kappa, indicating almost perfect agreement.

4. Findings

For analysis, we chose only five groups out of the seven that participated in our study. The remaining were discarded as their data was either not recorded properly, or they were silent throughout the intervention. We report only two episodes of negotiation here in the order of the activities in MEttLE. Episode 1 is between two male friends, S3 and S4, from final-year Mechanical engineering at the same college. They had participated in many robotics contests, solved open robot design and building problems, and experienced projects that required understanding and applying power, energy, and related concepts. Episode 2 is between S1 (female) and S2 (male) pursuing degrees in Electronics & Telecommunication and Mechanical Engineering, respectively. They studied in different colleges and did not know each other. S1 was in her final year and had worked on a car designing project where she used MATLAB and Python programming. She had an understanding of power, energy, and related concepts. S2 had recently completed his graduation and had worked on designing and coding an AR simulation for a Lathe machine operation. All students were familiar with spoken and written English.

4.1 Episode 1

This episode begins when S3 (Driver) and S4 (Navigator) start working in MEttLE. S3 asked the researcher if they could converse in Hindi while solving the problem. They watched the video that discussed the example of estimating the power of a human heart in MEttLE and then moved to the explanation of the estimap related to it (Figure 1a). They took many turns negotiating the appropriate sequence of the sub-goals and the domain-specific details related to solving this problem.

Table 1. *Episode 1 between S3 and S4*

Learner	Verbal/Interaction with MEttLE	Time	WC
(L4) S3	Functional Model Qualitative Model [pointing the cursor to functional and the qualitative]	0:09:17	EX
(L5) S4	What does the car need to do that requires power? [Reads the question in Functional Model sub-goal]	0:09:20	
(L6) S3	what are the dominant parameters that affect power? [Reads the question in Qualitative Model] So this will come second	0:09:23	EX
(L7) S3	Substituting reasonable values in the equation, what is the estimate of power? [Reads the question in the Calculation sub-goal]	0:09:27	EX
(L8) S4	hey, this one is the second…once we have a value, we can go ahead and solve this	0:09:33	C
(L9) S3	this should come third [speaking about calculation mode] because [putting his cursor on the qualitative and then functional model sub-goal] you have to first figure out the parameters that you need in the equation to solve the equation.	0:09:38	C
(L10) S4	[interrupts S3 and points his finger to the qualitative model on the screen. S3 brings the cursor to the qualitative model] But those are not the power that you require to solve the equation; they are just the dominant parameters that you require to think	0:09:48	C
(L11) S4	… let's just see the others.	0:10:09	Q

(L4) S3 externalized verbally and pointed the cursor to state that functional modeling should be approached first, (L6) qualitative modeling second, and the (L7) calculation sub-goal third. (L8) S4 stated that the calculation sub-goal should be second in the sequence, thus contradicting S3's statement illustrating conflict-oriented consensus-building. (L9) S3 further contradicted S4 and justified his argument by explaining why the calculation sub-goal should be approached third. S3 defended his claim with a new argument, a trait seen while conflict-oriented consensus building. S3's reasoning pushed S4 to bring another perspective (L10). The discussion abruptly ended when S4 (L11) suggested they read other sub-goals. Since they could not build common ground, they decided to opt for a quick consensus-building approach and moved ahead in the problem-solving process.

After repeated negotiation via different processes, they finalize a sequence most favourable to solve the estimation problem in MEttLE. When the researcher probed S3 on why they discussed the estimap, he elaborated that attaining a common understanding at the planning stage was crucial as it set a stage for the entire problem-solving. He added that it also put light on domain-specific details considered and contested by both as important parameters for this problem. The learners applied their

understanding from their negotiation on the human heart problem to the actual car racing problem in MEttLE.

4.2 Episode 2

S1 (Navigator) and S2 (Driver) solved the problem and attained an estimate of 200W in the Calculation sub-goal. In the Evaluation sub-goal, they examined whether their calculated answer was reasonable against i) the order of magnitude and ii) how it compares with known values.

S2 found that the power requirement for the example 10-kg toy car was 20W in the Info Centre. Hence the discussion was about the estimated power value and the car's parameter (weight).

Table 2. *Episode 2 between S1 and S2*

Learner	Verbal/Interaction with MEttLE	Time	WC
(L1) S2	I think it is 20W	1:28:09	EX
(L2) S1	But it's a toy car [points her fingers on the screen and changes tone to emphasize the type of car] ...we weren't…	1:28:11	
(L3) S2	they didn't mention anything about toy car..they just say car	1:28:16	C
	S2 manipulates the parameters in Simulator 4	1:29:00	
(L4) S1	Is it only this much? [changes tone] ...current 10? It will increase a little, at least.	1:29:11	C
	S2 increases the value of the current	1:29:34	I

(L1) S2 externalized his view by announcing that their answer needed to be changed to 20W. (L2) S1 disagreed and argued that weight alone could not be the primary criterion to decide the electric car's power. She insisted on re-evaluating their solution approach for the electric racing car, which was a different context than the example in the Info Centre. (L3) S2 gave his reasonings, returned to the Calculation sub-goal, and clicked on Simulator-4 (Figure 1b). Here, he tried to reduce the parameter of Current to the minimum to acquire the power of 20W. (L4) S1 contended and insisted that the value of the current be raised to a reasonable number as it's an electric car. S2 increased the current to a value acceptable to both, depicting an integration-oriented consensus-building approach. S1 reasoned that the estimate of power for the racing car would always be greater than 20W even if the current were reduced to the minimum. Hence the answer should be the same, i.e., 200W. S1 used the simulator to facilitate the modelling process and tried questioning, explaining, and critiquing S2 (conflict-oriented consensus building), who stuck to a simplified model and tried to attain his desired answer of 20W (stand pat).

After many turns of conflicting views, they did not attain a common ground. They did new calculations and obtained an answer of 65W, which they agreed to retain and move to the next subsection. This was quick consensus building where the partners agreed merely to move ahead.

5. Discussion and conclusion

We saw evidence of social modes of co-construction of knowledge multiple times in the negotiation episodes. Neither partner was inclined to agree easily when students used a conflict-oriented consensus-building approach. The negotiation took multiple turns, each providing his/her arguments and counterarguments. When integration-oriented consensus building was used, an agreement was quickly reached and was robust. On the other hand, when quick consensus building was seen, learners did choose to move ahead; but circled back to their original standpoint later, indicating that the agreement may not have been complete or authentic. Whether learners agreed or disagreed at the final state, engaging in negotiation led to productive behaviors concerning their estimation problem-solving. In Episode 2, students could create, contextualize and refine their models and push to reason more as they attempted to reach a common ground. In Episode 1, students' negotiation at the problem-solving steps related to planning and domain concepts showed some impact on their metacognition and conceptual understanding, which was later applied to the actual car problem. This study reaffirmed that collaboration in TELEs encourages learners to critically reason, elaborate, and justify their claims more

than they would individually on ill-structured problems (Jeong & Hmelo-Silver, 2016). Learners' negotiation supported their performance in MEttLE.

In our study, we observed potentially influencing factors of the amount and quality of negotiation, such as learners' comfort in effectively conveying their viewpoint or critiquing their partner's idea in their native tongue Hindi (national language). Thus, language familiarity plays a vital role in negotiation (Lai, Lin, & Kersten, 2010). Further, dyads of friends were more productive in negotiations towards problem-solving compared to the non-friends' group, which was substantiated by S1 in the interview. Thus, relationship with collaborators plays a role in the negotiation process (Maldonado, Klemmer, & Pea, 2009). In addition, the collaborative script gave learners a structure to articulate their thoughts and negotiate effectively. Lastly, the negotiation was heavily guided by MEttLE's design, which supported the complexity of tasks, intermediate reflection spots, and the degree of flexibility offered to learners. This increased the latitude for negotiation (Dillenbourg & Baker, 1996) and fostered them to converge on an acceptable common ground to navigate the other steps.

We see the following limitations: Our findings are based on a small sample and for a short duration in lab conditions. These findings form the basis of our next steps: expanding the analysis to a larger group of participants, confirming the conjectures related to factors influencing the negotiation quality, correlating negotiation and performance, and designing collaboration strategies to better support students in learning estimation problem-solving. Overall, we found that the negotiation process between learners impacted their estimation of problem-solving. We believe that the processes that underlie good estimation, that is, applying model-based reasoning processes and paying attention to planning and evaluation, are strengthened when learners collaborate and negotiate.

References

Baker, M. (1994). A model for negotiation in teaching-learning dialogues. Journal of Interactive Learning Research, 5(2), 199.

Beers, P. J., Boshuizen, H. P., Kirschner, P. A., & Gijselaers, W. H. (2006). Common ground, complex problems, and decision-making. Group decision and negotiation, 15(6), 529-556.

Carell, A., & Herrmann, T. (2009). Negotiation-Tools in CSCL-Scenarios-Do they have a valid use?

Dillenbourg, P., & Baker, M. (1996, June). Negotiation spaces in human-computer collaborative learning. In Proceedings of the International Conference on Cooperative Systems (pp. 12-14).

Emara, M., Rajendran, R., Biswas, G., Okasha, M., & Elbanna, A. A. (2018). Do Students' Learning Behaviors Differ when they Collaborate in Open-Ended Learning Environments?. Proceedings of the ACM on Human-Computer Interaction, 2(CSCW), 1-19.

Fischer, F., Kollar, I., Haake, J. M., & Mandl, H. (2007). Perspectives on collaboration scripts. Scripting computer-supported communication of knowledge: cognitive, computational, and educational perspectives. Springer, New York, 1-10.

Jeong, H., & Hmelo-Silver, C. E. (2016). Seven affordances of computer-supported collaborative learning: How to support collaborative learning? How can technologies help? Educational Psychologist, 51(2), 247-265.

Jonassen, D., Strobel, J., & Lee, C. B. (2006). Everyday problem-solving in engineering: Lessons for engineering educators. Journal of engineering education, 95(2), 139-151.

Kothiyal, A., & Murthy, S. (2020). Disciplinary Model-Based Reasoning and Metacognition Underlies Good Estimation Performance by Engineering Undergraduates.

Kirschner, F., Paas, F., & Kirschner, P. A. (2009). A cognitive load approach to collaborative learning: United brains for complex tasks. Educational psychology review, 21(1), 31-42.

Lai, H., Lin, W. J., & Kersten, G. E. (2010). The importance of language familiarity in global business e-negotiation. Electronic Commerce Research and Applications, 9(6), 537-548.

Mahajan, S. (2014). The art of insight in science and engineering: mastering complexity (p. 408). The MIT Press.

Maldonado, H., Klemmer, S., & Pea, R. (2009). When is collaborating with friends a good idea? Insights from design education.

OBS (2021). Open Broadcasting Software. Retrieved from https://www.obsproject.org/

Richardson, G. P., & Andersen, D. F. (1995). Teamwork in group model building. System Dynamics Review, 11(2), 113-137.

Sidner, C. L. (1994, August). An artificial discourse language for collaborative negotiation. In AAAI (Vol. 94, pp. 814-819).

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. Computers & education, 46(1), 71-95

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Monitoring of Learners' Activities in Software Structure Design Exercises

Yasuhiro NOGUCHI[a*], Kanta INOUE[a], Satoru KOGURE[a], Koichi YAMSHITA[b] & Tatsuhiro KONISHI[a]
[a]*Faculty of Informatics, Shizuoka University, Japan*
[b]*Faculty of Business Administration, Tokoha University, Japan*
*noguchi@inf.shizuoka.ac.jp

Abstract: Learners acquire modeling skills from software design exercises using UML class diagrams by recognizing and solving errors by comparing their models with the requirement specifications. Teachers and teaching assistants must identify learners' impasse according to the learners' modeling activities and provide them appropriate feedback when they cannot solve their impasse. In this study, we constructed a system to observe learners' modeling activities during exercises. In an experiment, we also collected modeling activity data from learners and analyzed the patterns identifying learners' impasse based on their activities.

Keywords: Modeling exercises, software design, impasse detection, monitoring

1. Introduction

In recent years, as software has become larger and more complex, the importance of learning about software design has increased. A wide variety of learners, both university students and working professionals who are in computer science specialties, require to acquire software design skills (Hara et al., 2019; Krusche et al., 2020). In an exercise to acquire software design skills, learners create a model based on provided requirement specifications, and the learnes review their models each other. However, some learners experience an impasse in the process of creating their own models, and they have trouble reaching the milestone to be reviewed and refining their model based on the feedback. Before the exercise, knowledge of typical mistakes in learners' modeling process (Chren et al., 2019; Chourio et al., 2019) can be provided to learners. However, in the exercises, learners must improve their modeling skills to enable them to recognize and solve errors in their models by themselves. Therefore, teachers and teaching assistants (TAs) should follow learners' modeling activities, including their trial and error, and support them when they cannot break the impasse by themselves.

In such modeling exercises in actual classrooms, a small number of teachers/TAs often must support many learners, and it is difficult to provide one-on-one support for each learner's modeling activities. In addition, it is difficult for teachers to identify only from the learner's created model which part of the model caused an impasse and how the learner performed the trial and error. In this study, we constructed a system for monitoring the modeling activities of multiple learners during their exercises. In an experiment, we collected the modeling activity data of six learners and analyzed the pattern of learners' impasse based on their modeling activities.

2. Modeling Activity Monitoring System

Figure 1 shows an overview of our system. In the exercise, learners use the modeling tool Astah* (Change Vision, 2022) with our add-on software. The add-on software periodically records learners' class diagram in JSON format. The system aggregates these models, which are recorded periodically during the exercise via the network, and extracts the differences between recorded models to create a record of the learner's editing history in the modeling exercise.

The learner's editing history is categorized into three types of operations: create, modify, and delete. Additionally, these operations are classified by the elements to be edited: classes, attributes, methods, and relations (e.g., association, aggregation, composition, inheritance, dependency, etc.).

Moreover, each element has specific properties (e.g., an attribute has type, name, visibility, multiplicity, etc.). Figure 2 shows an example of a learner's modeling activities. In the extraction process, an element is categorized as “create” if it exists in the history at a particular moment but not in the previous point. The reverse case is categorized as “delete.” An element is classified as “modify” if the same location of the element is edited between one point and the previous point. For instance, in (2) of Figure 2, the role name “corners” was modified from “points” and the multiplicity “4” was created. The class “Circle” with an attribute “radius” and method “draw” were created. The association between “Point” and “Circle” was created with the multiplicity “1” and the role name “center.”

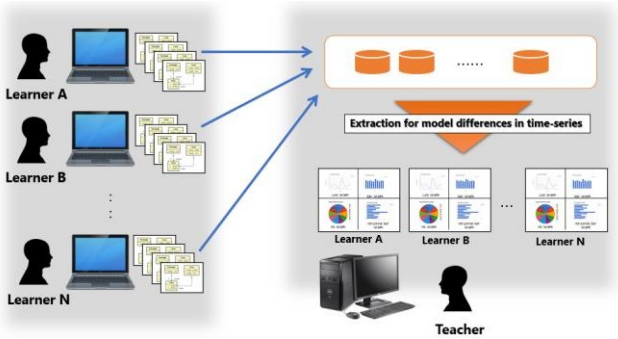


Figure 1. Overview of the monitoring system.

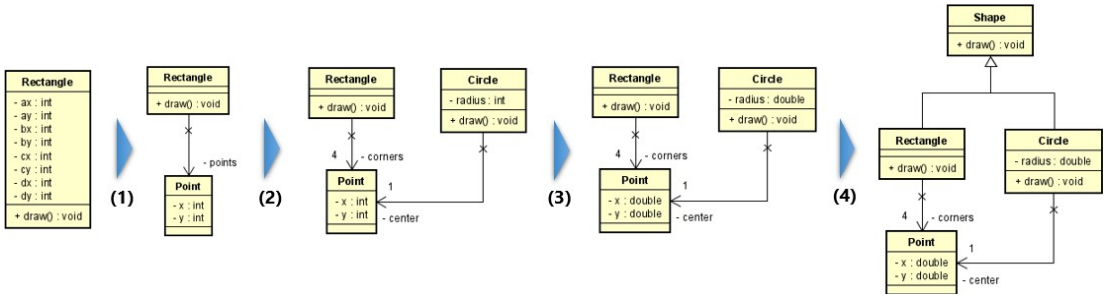


Figure 2. Learner’s modeling activity.

3. Experiment

In this experiment, we collected the modeling activity data of six learners in an exercise to create a model of a point-of-sale system in 60 minutes, and they recorded the time when they felt an impasse. We evaluated three impasse-detection rules based on the learners’ activity data that were customized from rules to detect programming exercise impasses (Yamashita et al., 2017):

- A learner did not change the model for more than 10 minutes.
- A learner deleted an element more than five times within 5 minutes.
- A learner deleted more than 15% of the diagram and it reverted to the previous model.

However, these rules did not adequately identify the learners’ impasse records. Figure 3 shows a learner's editing activities and number of elements in the diagram during the modeling exercise. The purple, red, and blue arrows show the time when a rule correctly identified the learner’s impasse record, the time only when the learner recorded an impasse, and the time when a rule was triggered at not impasse time, respectively.

Regarding the red arrowed time, we observed that the learner recorded the impasse; however, the learner focused on another area instead of refining the elements related to the impasse. Regarding the blue arrowed time, the learner deleted all the attributes in a class and replaced them. The extraction process should be improved to be categorized as “modify” even when the learner takes time to replace a chunk of elements. From the observed modeling activities, learners used various modeling methodologies (Torre et al., 2018; Störrle et al., 2018). One attempted to complete one class and then begin creating the next. Another attempted to create all the classes first and then detail them. Impasse-detection rules must be adjusted depending on the modeling methodology selected by the learner. In addition, some models created by the learners included major errors even when they did not

experience an impasse. The rules should be designed based on the learners not having tool-supported systematic verification methods such as “build,” “run,” and “trace” in programming exercises.

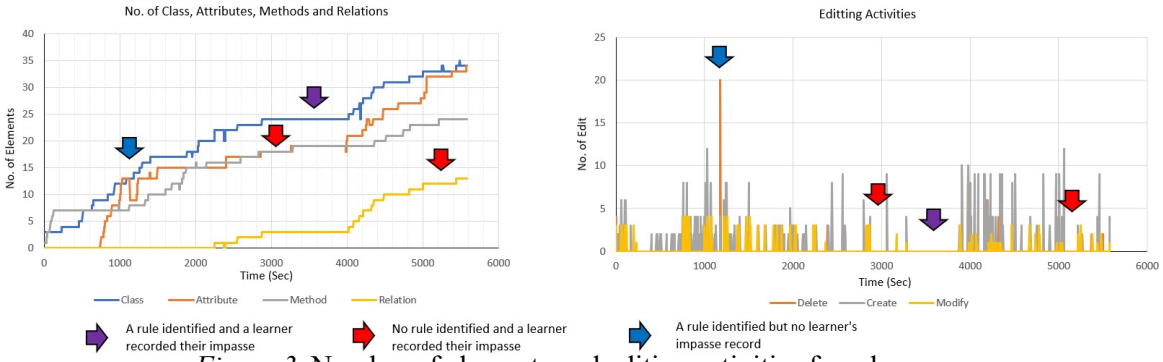


Figure 3. Number of elements and editing activities for a learner

4. Conclusion

In this study, we built a prototype to monitor and analyze learners' modeling activities. Using the prototype, we collected modeling activity data from six learners and analyzed learners' impasse based on their activities. From the learners’ activities, we could observe the characteristics of learners’ modeling activities and differences from programming exercises. In a future study, we will refine the impasse-detection mechanism based on these characteristics.

Acknowledgements

This work was supported by JSPS KAKENHI, Grant Numbers JP18K11566 and 22K12311.

References

Change Vision. Premier Diagramming, Modeling Software & Tools | Astah, https://astah.net, accessed on 2022/05/01.

Hara, S., Kayama., M., Nakao, T., Nagai., T, & Taguchi, N. (2019). A UML programming environment for ICT related subject at Junior High School. *In Proceedings of the 2019 The 3rd International Conference on Digital Technology in Education (ICDTE 2019)*. ACM, USA, 141–146. https://doi.org/10.1145/3369199.3369215

Krusche, S., Frankenberg, N., Reimer, L. M., & Bruegge., B. (2020). An interactive learning method to engage students in modeling. *In Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET '20)*. ACM, 12–22. https://doi.org/10.1145/3377814.3381701

Chren, S., Buhnova, B., Macak, M., Daubner, L., & Rossi., B. (2019). Mistakes in UML diagrams: analysis of student projects in a software engineering course. In Proceedings of the 41st International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET '19). IEEE Press, 100–109. https://doi.org/10.1109/ICSE-SEET.2019.00019

Chourio, P., Azevedo, R., Castro, A., & Gadelha, B. (2019). Most common errors in software modeling using UML. *In Proceedings of the XXXIII Brazilian Symposium on Software Engineering (SBES 2019)*. ACM, 244–253. https://doi.org/10.1145/3350768.3353820

Störrle, H., Baltsen, N., Christoffersen, H., & Maier, A. M. (2018). How do modelers read UML diagrams? Preliminary results from an eye-tracking study. *In Proceedings of the 40th International Conference on Software Engineering,* ACM, 396–397. https://doi.org/10.1145/3183440.3195025

Torre, D., Labiche, Y., Genero, M., Teresa., M. B., & Elaasar, M. (2018). UML diagram synthesis techniques: a systematic mapping study. *In Proceedings of the 10th International Workshop on Modelling in Software Engineering (MiSE '18)*. ACM, 33–40. https://doi.org/10.1145/3193954.3193957

Yamashita, K., Sugiyama, T., Kogure, S., Noguchi, Y., Konishi, T., Itoh, Y. (2017). An educational support system based on automatic impasse detection in programming exercises. *Proceedings of the 25th International Conference on Computers in Education*, 288-295.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Innovative and Effective Spreadsheet Tool for Learning Sentiment Analysis and Prediction

Michelle CHEONG[a*] & Jean CHEN[a]
[a]*School of Computing & Information Systems, Singapore Management University, Singapore*
*michcheong@smu.edu.sg

Abstract: In this paper, we describe a spreadsheet tool which was developed and implemented to teach sentiment analysis and prediction to university students who are less technically inclined. We adopted the "Use-Modify-Create" cycle to ensure that the students learn the computing concepts and practices in a progressive and active learning manner, and we evaluated the effectiveness of the tool using the Learning Object Evaluation Scale for Students (LOES-S). The evaluation shows that our tool performed well in terms of quality, learn and engagement constructs, and our paired T-test shows strong evidence that, on average, the course module with the tool does lead to improvement in the students' knowledge in sentiment analysis and prediction.

Keywords: Sentiment analysis and prediction, spreadsheet tool, less technically inclined students

1. Introduction

Teaching and learning technical topics such as sentiment analysis and prediction can be daunting for the instructors and students, especially for students who are less technically inclined. To achieve effective teaching and learning, digital education system needs to incorporate technologies and tools to assist the teaching and learning process (Williamson, 2016).

One such tool is the Text Sentiment Analysis Excel Add-in from the Azure Machine Learning Web Services available in the MS Office Store. However, such an Add-in will only function as a black box to the users and using it to teach students will not be able to achieve the intended learning outcomes such as, interpreting the theory and concepts of sentiment analysis and prediction, changing the parameters to improve the prediction results, analyzing the prediction results obtained, and deploying in new implementations. As such, the "Use-Modify-Create" cycle (Lytle, 2019) which is commonly used to learn computing concepts and practices in a progressive and active learning manner, cannot be implemented. Expecting the less technically inclined students to develop their own sentiment analysis and prediction tool using R or Python programming would be challenging, and the intended learning outcomes may be overshadowed by the technical difficulties met during coding, leading to students 'missing the forest for the trees'.

In this study, we explain how we used an innovative spreadsheet tool in a classroom for the learning of sentiment analysis and prediction, following the "Use-Modify-Create" cycle. In addition, we will provide a statistical analysis of the students' knowledge gain based on pre-test and post-test results, and share the students' quantitative and qualitative feedback on using the tool. The results of this study can inspire educators to consider using easy-to-use and effective tool to teach technical topics to students who are less technically inclined, to achieve the intended learning outcomes.

2. Literature Review

To prepare the next generation of learners to have a basic understanding of machine learning, many educators have designed tools to achieve this objective. von Wangenhelm et al. (2021) provided a ten-year systematic mapping of visual tools for teaching machine learning (ML) in K-12. 16 tools were identified, and most of them support both the development of ML models and the deployment as software artifacts. The deployment platforms are mainly block-based programming environments

including Scratch (Agassi et al., 2019; Alturayeif et al., 2020; Druga et al,, 2019; Zimmermann-Niefield et al., 2020), App Inventor (Tang et al., 2019; Zhu, 2019), and Snap! (Jatzlau et al., 2019; Kahn et al., 2018). The application areas are mainly in image recognition, speech recognition and synthesis, and object detection.

Batrinca & Treleaven (2015) surveyed techniques, tools and platforms for social media analytics. In the paper, they reviewed software tools and explained how to use them to scrape, cleanse and analyze social media data, mainly to benefit researchers to carry out their research works, rather than to use them for teaching and learning. Specifically for text analysis tools, they listed Python NLTK – Natural Language Toolkit (www.nltk.org) and GATE (http://gate.ac.uk/sentiment). While both NLTK and GATE have been used for teaching in a few institutions, they will require the students to have software engineering knowledge and programming skills in Python or Java.

R is another programming language that is often used to perform word cloud and sentiment analysis. Kabir, Ahmed & Karim (2020) described how they used R programming language and R Studio IDE to analyse Amazon earphones online reviews to know people's attitude, opinions, reviews and sentiments. They discussed their research steps in terms of how they read in the data set to build the corpus, perform data cleaning, create the term document matrix, create the bar plot and word clouds, obtain the sentiment score, and finally draw conclusions on the opinions and sentiments for the earphones. This work is a suitable reference for anyone to replicate what they have done but was not designed as a teaching material to be delivered in a classroom.

The work by Subramanian & Cote (2018) is the closest match to our study. In their paper, they described how a fully online graduate course in text mining using SAS Enterprise Miner and Twitter data were taught at the Quinnipiac University. They listed the topics taught in the course and shared the details of the student's project in terms of the data, deliverables expected, and grading rubrics. One selected student's report was presented as appendix. While they concluded that the course was a success, there was no data collected and analysed to evaluate how effective the course was, and no evaluation metric was used to assess whether SAS Enterprise Miner was indeed an effective tool to use in the teaching and learning.

To assess the effectiveness the use of software tools for learning, evaluation metrics for learning objects can be applied. IEEE Learning Technology Standards Committee (LTSC) defined learning object as "any entity, digital or non-digital, which can be used, re-used or referenced during technology-supported learning." (LTSC, 2007). Kay & Knaack (2008) proposed a multi-component model, known as the Learning Object Evaluation Metric (LOEM) with four constructs: interactivity, design, engagement, and usability, based on comprehensive review of literature. They tested the model with 1113 students, 33 teachers and 44 learning objects, and showed good internal and inter-rater reliability, and all four constructs correlated significantly with student learning performance. The same study also developed a reliable student-based evaluation tool for assessing learning objects, known as the Learning Object Evaluation Scale for Students (LOES-S) using three constructs: learning, quality, and engagement (Kay & Knaack, 2009).

This study aims to explain how the tool was implemented in a classroom for learning sentiment analysis and prediction following the "Use-Modify-Create" cycle, for less technically inclined university students. We provide a statistical analysis of the students' knowledge gain based on pre-test and post-test results. In addition, we adopted the LOES-S proposed by Kay & Knaack (2009) to perform the evaluation of tool effectiveness from the students' perspective.

3. Sentiment Analysis and Prediction Tool

3.1 Text Pre-Processing and Sentiment Prediction

To perform sentiment analysis and prediction, the first step is text pre-processing which includes tokenization and normalization, where normalization will involve stemming and lemmatization. In stemming, part of the word will be cut off, either at the beginning or the end of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word. For example, "studies" will be stemmed into "studi", while "studying" will be stemmed into "study". Such indiscriminate and crude way of cutting will not always result in words which are easily recognizable. Lemmatization attempts to overcome the shortcomings of stemming, by considering the full

morphological analysis to identify the lemma for each word. For example, "studies" and "studying" will both be linked to the lemma "study". To do so, it is necessary to have detailed dictionaries which the algorithm can look through to link the inflected word back to its lemma.

Due to the shortcomings in stemming and the complexity in lemmatization, the tool that was developed does not include both features. Instead, it will include stemming to remove unnecessary characters that will obscure the word. For example, "…apple" will be front stemmed into "apple" with "…" cut away, while "orange…" will be back stemmed into "orange" with "…" cut away. The objective is to ensure that complete words are kept in the dictionary for easy understanding and analysis.

The Naïve Bayes' Classifier is implemented in the tool to classify a new text into positive or negative sentiment. Neutral sentiment was not included in the prediction, and other classifiers can also be implemented easily in future versions of the tool.

3.2 The Tool

The tool was developed using Excel and VBA programming and the workbook consists of nine sheets. Excel was used as it is a tool most people are comfortable with, and the buttons with VBA programs assigned will improve usability tremendously. The first sheet **PasteHere** is the main user interface where the user will paste the labeled texts to train the machine learning algorithm, as well as to paste new texts for sentiment prediction. At the top of the **PasteHere** sheet are eight buttons, and each button has a VBA macro program assigned, for the user to perform automated tasks and computations. Figure 1 shows the user interface on **PasteHere** sheet with the eight buttons, and the arrows added between the buttons will guide the user how to use the tool in a sequential manner.

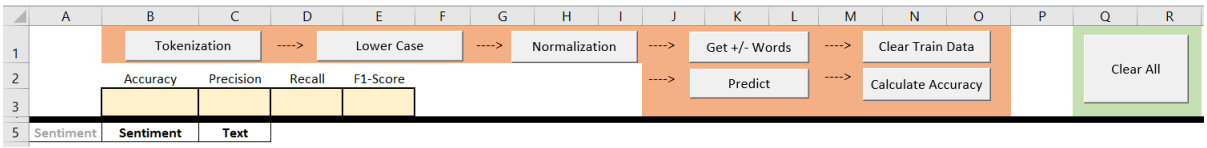


Figure 1. User Interface of PasteHere sheet.

The eight other sheets as described below:

- **Work** sheet – This is an empty sheet initially. It will be filled with useful words which are left over after the VBA program performs word removal and word stemming.
- **Positive** sheet – This is an empty sheet initially. It will be filled with words which are extracted from positive texts in the training data set. Each word will have its associated frequency count, log probability and conditional probability.
- **Negative** sheet – This is an empty sheet initially. It will be filled with words which are extracted from negative texts in the training data set. Each word will have its associated frequency count, log probability and conditional probability.
- **Stopwords** sheet – This sheet contains a list of 174 stop words (e.g. a, is, of, the) retrieved from http://www.ranks.nl/stopwords.
- **FrontRemove** sheet – This sheet contains a list of prefixes (e.g. #, @, http, RT) which the VBA program will remove words with these prefixes.
- **Otherwords** sheet – This sheet contains a list of words which are neither Stopwords nor FrontRemove words. It is useful for words which are specific to the data set and the user would like to have such words removed.
- **FrontStems** sheet – This sheet contains a list of special characters which the user would like to have them stemmed from the front of the words.
- **BackStems** sheet – This sheet contains a list of special characters which the user would like to have them stemmed from the back of the words.

The tool is flexible to accommodate any text for text pre-processing and sentiment prediction, as the user can edit the work lists in the **FrontRemove**, **Otherwords**, **FrontStems** and **BackStems** sheets.

4. Research Methodology and Implementation

4.1 Research Methodology

The tool was used in a course module "COMM629 Analytics" with 38 students pursuing the Master of Communications Management programme offered by the Lee Kong Chian School of Business at the Singapore Management University. The module was delivered over a 5-week period with a 3-hour class session each week. The topic on sentiment analytics was delivered in the third session. During this session, theory and concepts related to sentiment analytics were taught, before a hands-on laboratory session was conducted to guide the students to "Use" the tool for sentiment analysis and prediction for a set of Twitter data. In the "Modify" step, the students will edit the list of words in the **FrontRemove**, **Otherwords**, **FrontStems** and **BacksStems** sheets to learn how they can affect the creation of positive and negative word features, which will in turn affect the prediction accuracy. In the "Create" step, students are tasked to perform a new experiment using self-sourced data set, and apply what they have learned in the "Modify" step to obtain as high a prediction accuracy as possible for the new data set.

4.2 Classroom Exercise

The class exercise consists of three main parts and students only need to click the buttons in sequence as explained in section 3.3:

- Part 1 is to use the training data with labeled tweets to obtain the positive and negative words with their respective frequency count, log probability and conditional probability computed.
- Part 2 is to use the test data set as new data for the machine to predict the sentiment for each new tweet.
- Part 3 is to compute the accuracy, precision, recall and F1-score by comparing the machine predicted sentiment with the user labeled sentiment for the test data set.

4.3 Pre-Test and Post-Test

The students took the pre-test at the start of session 1, after they have given consent to take part in the study, and they took the post-test at the end of session 5. There were ten Multiple-Choice Questions and for each question, four answer options were provided where the last answer option is always "I do not know the answer". The students were advised to choose the last answer option "I do not know the answer" if they are not sure of the answer instead of making a wild guess. This is to prevent the situation where students managed to obtain the correct answer due to random selection, which will affect the computation of actual knowledge gain.

4.4 Feedback Survey and Evaluation Metric

As there is no control group (where students will not get to use the tool), a survey was conducted at the end of Session 5 with the students to collect quantitative and qualitative feedback on the effectiveness of the tool. The four survey questions match the three constructs: learn, quality, and engagement in the Learning Object Evaluation Scale for Students (LOES-S), proposed by Kay & Knaack (2009):

- Q1: The tool is easy to understand and use. (Quality)
- Q2: I learn how to use the tool to pre-process text to create word features for positive and negative sentiment prediction. (Learn)
- Q3: By using the tool, I can understand and apply the concepts and principles behind sentiment analytics better. (Learn and Engagement)
- Q4: By using the tool, I can easily conduct experiments on sentiment analytics. (Engagement)

5. Effectiveness of the Tool for Learning Sentiment Analysis and Prediction

5.1 Pre-Test and Post-Test Analysis

With the pre-test and post-test average scores, we conducted a paired T-test to test the null hypothesis (H0) that the true mean difference between the scores is zero, and obtained a p-value of <0.0001 (<

0.05, reject H0) and t-statistic of 13.092. Thus, there is strong evidence that, on average, the course module with the tool does lead to improvement in the students' knowledge in sentiment analysis and prediction. It is acknowledged that there is no control group to make the comparison in knowledge gain, as there is no alternative course that teaches the same topic without the use of the tool.

5.2 Quantitative Feedback Analysis

To gain more insights into the effectiveness of the tool in learning sentiment analysis and prediction, a survey with four questions, each with five options ("Strongly Disagree", "Disagree", "Neutral", "Agree", and "Strongly Agree") was conducted. Based on the percentage of students who rated "Agree" and "Strongly Agree", the feedback results in Table 1 show that the tool performed very well. The best construct was Learn (78.4%), followed by Learn and Engagement (75.0%), then Quality (67.6%). Engagement in terms of using the tool to conduct experiments scored the lowest (58.3%). A plausible reason could be that the students needed more practice before they could conduct new experiments independently. Recall that the topic on sentiment analytics was delivered only in one 3-hour session, which could be insufficient to cover the theory and concepts, and to conduct the hands-on laboratory on using the tool. Providing more time and more practice exercises will improve this construct in future.

Table 1. *Quantitative Feedback Results*

	% Rated Agree and Strongly Agree
Q1: The tool is easy to understand and use. (Quality)	67.6%
Q2: I learn how to use the tool to pre-process text to create word features for positive and negative sentiment prediction. (Learn)	78.4%
Q3: By using the tool, I can understand and apply the concepts and principles behind sentiment analytics better. (Learn and Engagement)	75.0%
Q4: By using the tool, I can easily conduct experiments on sentiment analytics. (Engagement)	58.3%

5.3 Qualitative Feedback Analysis

Out of 38 students, 15 students did not provide qualitative feedback. Of the 23 feedback comments, almost all of them are positive which include the following:

- "The tool enables us to approach data and its analysis in methodological way, without involving complex equations."
- "Through the simple-to-understand steps, the learning and pickup of skills are no-issue for any layman."
- "The tool strengths my understanding of the data analysis, the background and the meaning."
- "Easy to use and helps in understanding of certain fundamental steps necessary to be conducted before sentiment analysis can be done."

Other feedback comments include "Need more practice to understand the usefulness" and "Can it be enhanced to include neutral sentiment", which are the potential improvement areas for future implementations.

6. Conclusions

In the age of digital education, learning technologies will enhance students' learning immensely. The challenging task in teaching technical topics to the less technically inclined students will require tools which are easy-to-use and yet effective. Our spreadsheet tool for the learning of sentiment analysis and prediction was shown to perform very well in terms of quality, learn and engagement constructs, in a course module delivered to 38 business school university students. The statistical analysis of the pre-test and post-test scores shows strong evidence that, on average, the course module with the tool does lead to improvement in the students' knowledge in sentiment analysis and prediction. Using LOES-S to perform the evaluation of tool effectiveness from the students' perspective, the tool

performed very well in Learn and Engagement constructs. Future improvements include adding in neutral sentiment prediction, prediction using a few other classifiers for comparison, and allowing students more practice opportunities to enhance learning.

Acknowledgements

Due to the need to collect data on students who are participants in the study, IRB application was submitted and approved as "Category 1:Exempt from Further IRB Review" with IRB Exemption Number IEB-22-0430E023(422).

References

Agassi, A., Erel, H., Wald, I. Y., & Zuckerman, O. (2019). Scratch nodes ML: A playful system for children to create gesture recognition classifiers. *Proceedings of the Conference on Human Factors in Computing Systems*, ACM, 1–6. https://doi.org/10.1145/3290607.3312894

Alturayeif, N., Alturaief, N., & Alhathloul, Z. (2020). DeepScratch: Scratch programming language extension for deep learning education. *International Journal of Advanced Computer Science and Applications*, 11(7), 642–650. https://doi.org/10.14569/IJACSA.2020.0110777

Batrinca, B., & Treleaven, P.C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & Soc*, 30, 89–116. DOI 10.1007/s00146-014-0549-4

Druga, S. Vu, S. T., Likhith, E., & Qiu T. (2019). Inclusive AI literacy for kids around the world. *Proceedings of Fab Learning*, ACM 104–111. https://doi.org/10.1145/3311890.3311904

Jatzlau, S. Michaeli, T., Seegerer, S., & Romeike, R (2019). It's not Magic After All – Machine Learning in Snap! using Reinforcement Learning. *Proceedings of IEEE Blocks and Beyond Workshop*, Memphis, TN, USA, 37–41. https://doi.org/10.1109/BB48857.2019.8941208

Kabir, A.I., Ahmed, K., & Karim, R. (2020). Word Cloud and Sentiment Analysis of Amazon Earphones Reviews with R Programming Language. *Informatica Econo*mica, 20(4), 55–71

Kay, R. H., & Knaack, L. (2008). A multi-component model for assessing learning objects: The learning object evaluation metric (LOEM). *Australasian Journal of Educational Technology*, 24(5), 574–591

Kay, R. H., & Knaack, L. (2009). Assessing learning, quality and engagement in learning objects: the Learning Object Evaluation Scale for Students (LOES-S). *Education Technology Research and Development*, 57, 147–168. DOI 10.1007/s11423-008-9094-5

Kahn, K. M., Megasari, R., Piantari, E., & Junaeti, E. (2018). AI Programming by Children Using Snap! Block Programming in a Developing Country. In *Proceedings of the 13th European Conference on Technology Enhanced Learning*, Leeds, UK.

Learning Technology Standards Committee (LTSC). (2007). The Learning Object Metadata standard. IEEE.

Lytle, N., et al. (2019). Use, Modify, Create: Comparing Computational Thinking Lesson Progressions for STEM Classes. *Proceedings of the Conference on Innovation and Technology in Computer Science Education*, ACM, 395–401. https://doi.org/10.1145/3304221.3319786

Subramanian, R., & Cote, D. (2018). Using SAS™ Software to Enhance Pedagogy for Text Mining and Sentiment Analysis Using Social Media (Twitter™) Data. *Journal of International Technology and Information Management*, 27(2), 73–98.

Tang, D., Utsumi, Y., & Lao, N. (2019). PIC: A Personal Image Classification Webtool for High School Students. In *Proceedings of the IJCAI EduAI Workshop*, Macao, China.

von Wangenheim, C.G., Hauck, J.C.R., Pacheco, F.S., & Bueno, M.F.B. (2021). Visual tools for teaching machine learning in K-12:A ten-year systematic mapping. *Education and Information Technologies,* 26, 5733–5778. https://doi.org/10.1007/s10639-021-10570-8

Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and 'real-time' policy instruments. *Journal of Education Policy*, 31(2), 123–141.

Zhu, K. (2019). An Educational Approach to Machine Learning with Mobile Applications. M.Eng thesis, MIT, Cambridge, MA, USA.

Zimmermann-Niefield, A., Polson, S., Moreno, C., & Shapiro, R. B. (2020). Youth making machine learning models for gesture-controlled interactive media. In *Proceedings of the Interaction Design and Children Conference*. ACM, 63–74. https://doi.org/10.1145/3392063.3394438

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Topic-Based Representation of Learning Activities for New Learning Pattern Analytics

Jinghao WANG[a*], Tsubasa MINEMATSU[a], Yuta TANIGUCHI[b], Fumiya OKUBO[a] & Atsushi SHIMADA [a*]
[a]*Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan*
[b]*Research Institute for Information Technology, Kyushu University, Japan*
*wangjiha029@gmail.com, atsushi@limu.ait.kyushu-u.ac.jp

Abstract: In recent years, several kinds of e-learning systems, such as e-book and Learning Management System (LMS) have been widely used in the field of education. When students access these systems, their activities on the systems will be continuously and automatically recorded and stored as learning logs. As the learning logs are stored in association with students and indicate students' learning activities, most studies have been "student-based" learning log analyses focused on students and each student's learning behavior. However, the "student-based" learning log analysis focuses on each student's learning behavior during the entire lesson (for example, studied well or didn't study enough) and cannot show what they learned. Therefore, if there is a need to investigate students' learning behavior regarding each topic of the lesson, such as which topic is learned well and which not in order to optimize the syllabus, we cannot conduct "student-based" learning log analysis directly. Instead of "student-based" learning log analyses, this study describes a method of "learning-topic-based" learning log analysis. We will show how to convert a learning log associated with students into a learning-topic-associated one and shape the logs into a two-dimensional matrix of learning topics and learning activities. Then we apply Non-negative Matrix Factorization (NMF) to the matrix in order to extract the learning patterns by activity. In addition, we make a three-dimensional matrix (tensor) of students, learning topics, and learning activities by subdividing the learning activities of each learning topic by students. We then apply Non-negative Tensor Factorization (NTF) to the tensor to extract detailed learning patterns. The methods proposed in this study will help teachers to have a comprehensively view of students' learning behaviors towards each learning topic easily even if the learning log is in a large-scale, so teachers can adjust syllabus according to the attracted learning behaviors, which is helpful to increase learning efficiency.

Keywords: Educational big data, e-book, learning analytics, Non-negative Matrix Factorization (NMF), Non-negative Tensor Factorization (NTF)

1. Introduction

In recent years, in the field of education, with the introduction and practical application of various kinds of e-learning systems such as e-books, "learning logs" that show what kinds of learning activities students completed are automatically recorded and have become easy to use. Against this background, by analyzing the learning logs accumulated and feeding back the analysis results to teachers and students, it is possible to promote improvements in lesson plans and learning methods, increase learning efficiency, and reduce the burden on teachers and students. In the conventional learning environment, grasping and improving learning conditions is done through limited data such as exam results and teacher's daily class observation. However, by using the abundant learning logs on e-learning systems, we can expect to improve teaching and learning methods scientifically. The fields of application of learning logs are wide-ranging, including recommendation of teaching contents, prediction of grades and early detection of students with poor grades.

The different e-learning systems include Learning Management Systems (LMSs), e-book systems and learning diary collection system. The learning activities of each student are recorded in the database of each system. By making a request to the databases, we can access the accumulated learning logs. As learning activities in each system are conducted by students, the learning logs acquired in each database are accumulated in a form linked to each student. For example, from an e-book system, we can obtain learning logs such as "student A read page 10 of e-book for 100 seconds." Due to this

characteristic of learning logs storage, most studies have been based on students ("student-based") and focus on each student's learning activities.

As a result of the "student-based" learning log analysis of the existing studies, there have been studies examining the patterns of the learning conditions of students that revealed the relevance between grades and learning patterns[1]. However, the "student-based" learning log analysis focuses on the learning activities of each student in the entire lesson, through which it is not possible to identify what they have learned. For that reason, in case of educational support from teacher's perspective, such as when teachers want to investigate what kinds of lesson content are well learned or review the teaching plan, a "student-based" learning log analysis is not useful.

Therefore, in this study, we seek to analyzing learning logs in the "learning-topic-base" of learning activities conducted by "learning topics." Learning topics are keywords that appear in the lesson. For example, in the subject of "Digital Signal Processing," keywords like "Fourier Transform," or "Discrete Signal" can be chosen as learning topics. In existing studies, "learning topic-based" learning log analysis is not seen.

There are two reasons why there has been almost no "learning-topic-based" learning log analysis. First, it is difficult to extract learning topics from teaching contents. Learning topics are not explicitly defined in the teaching contents, and no existing study has proposed a method for extracting learning topics. Second, it is difficult to acquire and organize the learning logs of each learning topic. As mentioned above, learning logs are stored in association with each student in the databases of e-learning systems. The subject of existing studies was how to convert or reshape the "student-based" learning logs into "learning topic-based" ones. We will propose a method of extracting learning topics and acquiring and merging "learning topic-based" learning logs in Section 3.

We apply Non-negative Matrix Factorization (NMF) and Non-negative Tensor Factorization (NTF) to analyze learning logs in this study. NMF and NTF have been proposed as effective analytical methods for extracting data characteristics from two-dimensional matrices and three-dimensional or higher-order tensors, respectively. The two-dimensional matrix of learning topics and learning activities can be formatted by reshaping learning logs that have been converted to "learning-topic-based" ones. Depending on the lessons, the number of types of learning topics or learning activities can reach the tens or even hundreds. Therefore, in this study, we try to reduce the ranks of the matrices of learning logs by applying NMF to a matrix to make it easier to extract learning patterns. Furthermore, by extending the above matrix method to a three-dimensional tensor of students, learning topics and learning activities and applying NTF to the tensor, we can extract and interpret learning patterns considering students, learning topics and learning activities at the same time.

This study established the following three points:

- We propose a method to acquire and reshape learning logs into "learning-topic-based."
- We propose a method to format a three-dimensional tensor of students, learning topics and learning activities.
- We apply NMF and NTF to the learning log matrices and tensors and extract and interpret learning patterns.

2. Related Studies

2.1 Studies of learning log analysis

Jo et al.[2] analyzed the length, frequency, and interval of log-in time to investigate how working learners develop time management strategies in online class. Based on the results of the analysis, the authors reported that learners with regular login intervals show higher learning performance. Gitinabard et al.[3] used students' access log to e-books and learning forum to predict dropout. Prediction results showed that learners at risk of potentially dropping out could be identified early. These studies are all learning log analysis in "student-base", and there are no attempts of "learning-topic-based" learning log analysis.

2.2 Studies applying NMF and NTF

Hasumoto et al.[4] extracted latent customer purchasing patterns by applying NMF to store purchasing data. Based on the extracted results, the authors clarified what kind of purchasing activities have a strong influence on "Customer Lifetime Value," which reflects future corporate profits, and proposed a new way to formulate corporate marketing strategies.

NTF, a multidimensional extension of NMF, analyses data that cannot be expressed in two-dimensional matrix. Kuwano et al.[5] reshaped search logs from bus route and timetable search site into a four-dimensional tensor of place of departure, destination, day of the week, and time of the day. By applying NTF to the tensor, the authors extracted patterns of bus use characteristics considering all the four factors above.

As described above, NMF and NTF have been applied in a wide range of fields for pattern extraction. However, few studies have attempted to apply them in the field of Learning Analytics.

3. Extraction of learning logs based on keywords

Since we choose NMF and NTF as the analysis methods of this study, we reshape learning logs into a two-dimensional matrix and a three-dimensional matrix (tensor).

Subsection 3.1 will give a brief introduction of e-learning systems from which the learning logs are obtained, and details of learning logs. As mentioned in Section 1, original learning logs are stored in association with students, so we need to convert them into learning topic-shaped ones. Subsection 3.2 will show how to extract learning topics, convert learning logs into "learning topic-based" ones and shape them into a two-dimensional matrix. Subsection 3.3 will show how to extend the matrix into a three-dimensional tensor.

3.1 Introduction of e-learning systems and learning logs

> *In this study, we will obtain learning logs from three e-learning systems: an e-book system, Learning article management system and Reflection management system.*

3.1.1 e-book system

Students can view teaching contents everywhere using computer or smartphone via the e-book system. We collect learning logs from e-books as follows:

- The content name and page number that the student has read;
- the reading time of each page; and
- the history of learning support feature use. *

*As shown in Table 1, students can use following learning support features.

Table 1. *Learning support functions.*

Activity name	Description
Add marker	to highlight text
Add bookmark	to make a bookmark of contents on current page
Add memo	to take a memo of contents on current page
Getit	shows the student understood contents on current page
Notgetit	shows the student didn't understand contents on current page

3.1.2 Learning article management system

A learning article is an article that contains knowledge and learning methods generated by the student during the process of learning. Learning articles are written by learning topics and one article corresponds to one learning topic or multiple related learning topics. Students post the key points,

knowledge, and learning methods used for understanding the learning topics in one Learning article using sentences and figures. The Learning management system is a system that allows sharing of Learning articles among students, and provides functions to post, view, and "like" articles. We collect learning logs from the Learning article management system as follows:

- The number of articles posted by each student;
- the number of characters and figures in each post;
- the number of views and likes of each post; and
- the articles posted (text data)

3.1.3 Reflection management system

A Reflection management system supports a learning diary that describes the student's reflections on what they learned in the lesson that day. A Reflection diary includes what the student understood and failed to understand, impressions of the lesson, and questions to the teacher, which are subjectively described from the student's point of view. We collect learning logs from the Reflection management system as follows:

- The number of articles posted by each student;
- the number of characters in each post;
- the articles posted (text data)

3.2 Converting "student-based" learning logs to "learning-topic-based" logs

As shown in Table 2, a learning log shows when, where, and by whom the learning activity was conducted. By combining learning logs obtained directly from e-learning systems, we can make a timeline of each student's learning activity and extract his learning behavior. However, with original learning logs obtained from e-learning systems, it is impossible to extract specific information about what kinds of learning topics the students learned. As a result, we need to reshape the original "student-based" learning logs into "learning-topic-based" ones.

Table 2. *An Example of learning logs obtained from the e-book.*

Student ID	Time	Learning content name	Page number	Activity
001	4:02:13 pm ,1 Feb 2021	2. Fourier Series	4	Add Marker
002	3:22:12 pm ,1 Feb 2021	2. Fourier Series	5	Add Memo
003	6:53:45 pm ,1 Feb 2021	2. Fourier Series	7	Notgetit

First, as learning topics are not defined in advance, it is necessary to define several learning topics. There are three steps to define learning topics.

1. *Extract nouns from the text of e-book using the morphological analysis tool "MeCab."*
2. *Remove alphanumeric characters and words that contain only one Japanese character (as one-character words usually don't make sense).*
3. *Remove words that have little relevance to course content.*

After extracting the learning topics, we will merge the learning logs of each learning topic. There are three steps to merge learning logs.

1. Sum the total amount of learning activities of each page of e-book. For example, as shown in Table 3, page 5 of the e-book contains learning activities conducted by student 001, 002, and 003. Then, the total amount of learning activities conducted on page5 will be shown in the 4th row of Table 3.
2. Make a text search as to whether each learning topic is included as a substring in the e-book text (check whether the learning topic appears as a substring on each page of the e-book) and create a list P for each learning topic of the pages where the learning topic appears. For example, assume the learning topic "discrete time signal" appeared on the 3rd and 5th pages of the e-book. Then P(discrete time signal)={3,5}.

3. Merge the learning logs of the pages included in list P of each learning topic.

Table 3. *Learning activities on page 5 of the e-book.*

Student id	Number of markers	Number of memos	...
001	14	5	...
002	3	2	...
003	8	4	...
Total	25	11	...

After these steps, we obtain a two-dimensional matrix of learning topics and each topic's learning activity as in Table 4.

Table 4. *Two-dimensional matrix of learning topics and learning activities.*

Learning Topic	(e-book) Number of marker	(e-book) Number of memo	...	(Learning article) Number of graphs
Periodic Signal	10	8	...	1
Aperiodic Signal	6	5	...	0
...	...	...	...	...

3.3 Tensor of students, learning topics and learning activities

The "learning-topic-based" learning log matrix introduced in Subsection 3.2 is a two-dimensional matrix of learning topic and learning activity. By applying NMF to this, it is possible to extract the learning behavior and grasp the learning condition of each learning topic. However, the two-dimensional matrix contains the learning activities of each learning topic conducted by all the students and the activities are not divided by student, so it is difficult to analyze it while considering students, learning topics, and learning activities at the same time, making it difficult to fully grasp the learning pattern with the two-dimensional matrix.

Therefore, in this study, we aim to grasp more detailed learning patterns by constructing a three-dimensional tensor of student, learning topic and learning activity and applying NTF to the tensor. Formatting into a three-dimensional tensor can be realized by dividing the learning activity of each learning topic for each student using the two-dimensional matrix in Subsection 3.2. For example, as shown in Figure 1, assume students A, B, and C have drawn 3, 5, and 2 markers for the learning topic "periodic signal," respectively. In the matrix, the learning topic is on the x-axis and the learning activity is on the y-axis. A three-dimensional tensor can be generated by dividing the 10 markers related to "periodic signals" for each student as 3 for A, 5 for B, and 2 for C and reflecting them in the z-axis.

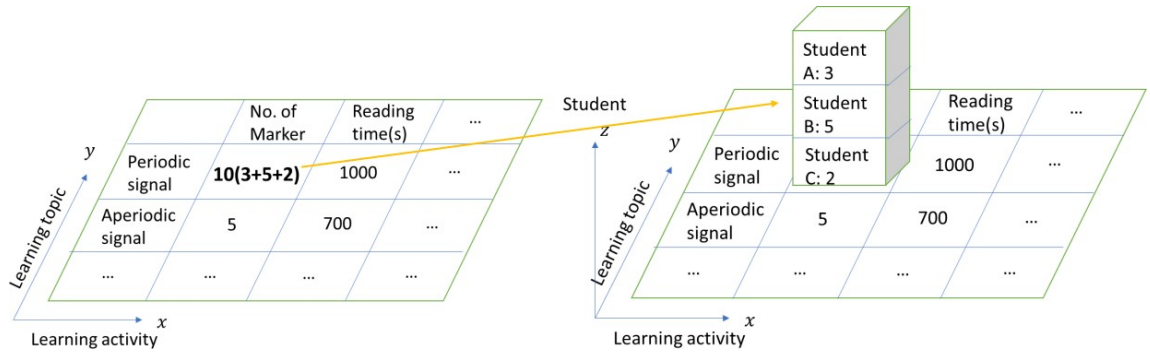


Figure 1. Formatting of the tensor.

4. Non-negative Matrix Factorization (NMF), Non-negative Tensor Factorization (NTF)

This Section outlines NMF and NTF, which are methods for analyzing the matrices and tensors of the learning logs created in Section 3. Subsection 4.1 introduces NMF, and Subsection 4.2 introduces NTF.

4.1 Overview of NMF

NMF[6] is one of the methods usually used for feature extraction from two-dimensional matrices. As mentioned in Section 2, NMF can be applied to any data that can be represented by a non-negative two-dimensional matrix and has been used in many studies in recent years. Since the learning log matrix created in Section 3 of this study is also a two-dimensional matrix composed of non-negative value data, NMF can be applied. NMF analysis yields several frequent patterns that represent the characteristics of the data. We will show the detailed results of the analysis applying NMF to the learning logs in Section 5.

A conceptual diagram of NMF is shown in Figure 2. The size of the original matrix X (the observation matrix, which means we will observe patterns from this matrix) is $m * n$. The result of applying NMF to X is to decompose it into the product of the matrix $T$of size $m * r$ and the matrix V of size $r * n$. Here, r is a parameter set in the initial stage of NMF called the "factor." By setting the factor to a value smaller than m and n, X can be decomposed into low-ranked matrices T and V, making it easier to observe patterns in the data. T and V represent the features of the vertical and horizontal axes of X, respectively, and are called feature matrices.

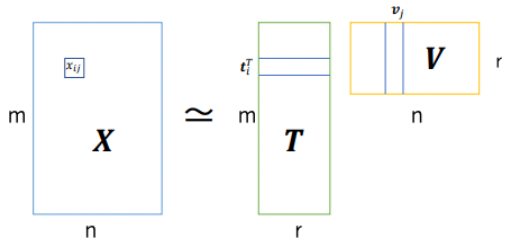


Figure 2. Conceptual diagram of NMF.

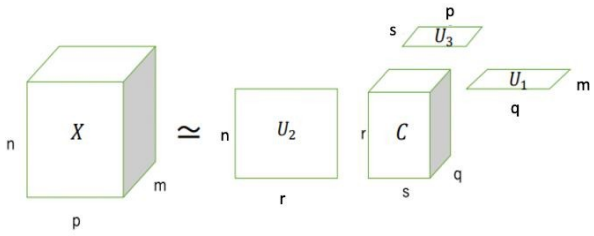


Figure 3. Conceptual diagram of NTF.

4.2 Overview of NTF

NTF[7] is an extension to higher dimensional matrices of NMF, which can only be applied to 2D matrices. A matrix of three or more dimensions is also called a tensor, and NTF is a method of extracting data characteristics from a tensor. NTF is formulated as $X \cong C \times U_1 \times U_2 \times U_3$. *A conceptual diagram of NTF is shown in Figure 3. If the size of the original tensor X (observation tensor) is* $m * n * p$, *we can set* q, r, s *smaller than* m, n, p *and decompose* X *into the feature matrices* U_1 ($m * q$), U_2 ($n * r$), U_3 ($p * s$) *and a tensor* C *of size* $q * r * s$, *which is called a core tensor. The characteristics of the data can be grasped from the feature matrices and the core tensor. Specifically, data characteristics of three axes of* X *can be represented by* U_1, U_2 *and* U_3 *respectively. For example, in this study,* X *is a tensor of student, learning topic and learning activity, so* U_1, U_2 *and* U_3 *are feature matrices of students, learning topics and learning activities, respectively, obtained by NTF.*
The core tensor can be regarded as a compressed observation tensor, and the larger the value of an element of the core tensor, the more the combination of factors corresponding to that element appears. Specific examples will be described in detail in Section 5.

5. Experimental results of NMF and NTF

In this section, we applied NMF and NTF to the two-dimensional matrix formatted in Subsection 3.2 and the three-dimensional tensor formatted in Subsection 3.3, with consideration to the results of attempting to interpret learning patterns. Subsection 5.1 shows the results and considerations of

applying NMF to the matrix of learning topics and learning activities and Subsection 5.2 shows the results and considerations of applying NTF to the tensor of students, learning topics, and learning activities.

In this study, we analyzed the learning logs acquired in the lesson "Digital Signal Processing" taken by 82 third-year students in the Department of Electrical and Information Engineering in the Faculty of Engineering of our university in the spring semester of 2020.

We extracted 47 types of learning topics and obtained learning logs of 12 kinds of learning activities consisting of the following items.

- e-book: number of markers, number of memos, number of bookmarks, number of Getits, number of Notgetits, reading time
- Reflection: number of posts, number of characters of each post
- Learning articles: number of posts, number of characters of each post, number of figures of each post, number of likes of each post

The data widths of different types of learning activities are not the same, as the number of markers ranges from several to several tens, while the reading time ranges from several tens to several thousand seconds. Therefore, we performed normalization of dividing each type of learning logs by the maximum value of that type so that the values for all kinds of learning logs fell within the range 0 - 1.

5.1 Experimental results of NMF

As a result of reshaping in Subsection 3.2, we obtained a matrix of learning topics (47 types) and learning activities (12 types) as Table 5.

Table 5. *Finally reshaped matrix of learning topics and learning activities.*

Learning topic	(e-book) Number of marker	(e-book) Number of memo	...	(Learning article) Number of likes
Periodic Signal	0.35396	0.33333	...	0.32955
Aperiodic Signal	0.05244	0.03333	...	0.14015
...	...	...	...	...
Window Function	0.05191	0.06667	...	0.12121

*We applied NMF with the number of factors set to 4, generating a feature matrix of learning activities (12*4) and a feature matrix of learning topics (47*4) as shown in Tables 6 and 7, respectively. We set the number of factors considering 2 points: (1) the error of NMF needs to be reduced. We conducted matrix multiplication to the two feature matrices to reproduce a matrix as the same size of that in Table 5 (47*12). Then we calculated the error between the two 47*12 matrices. We found that as the number of factors get larger, the error between two matrices reduced. (2) the ease of interpretation. Although setting the number of factors bigger helps to reduce error, learning behaviors can't be summarized efficiently and it's difficult to interpret these factors. So, we made a tradeoff decision between (1) and (2) in our experiments utilizing NMF and NTF.*

Table 6. *Feature Matrix of learning activity. B: e-book, R: Reflection, L: Learning article, F: Factor*

Learning Activity	F 1	F 2	F 3	F 4
(B) Marker	0.001	0.065	0.124	0.544
(B) Memo	0.026	0.000	0.000	0.051
(B) Bookmark	0.000	0.002	0.000	0.283
(B) Getit	0.010	0.003	0.417	0.000
(B) Notgetit	0.119	0.000	0.003	0.021
(B) Reading time	0.253	0.004	0.041	0.214

Learning Activity	F 1	F 2	F 3	F 4
(R) Number of articles	0.313	0.323	0.002	0.001
(R) Number of characters	0.144	0.417	0.000	0.000
(L) Number of articles	0.185	0.003	0.015	0.112
(L) Number of characters	0.227	0.000	0.001	0.127
(L) Number of graphs	0.040	0.000	0.018	0.005
(L) Number of likes	0.054	0.004	0.053	0.000

According to Table 6, Factor1 is a factor specific to the learning activities of Reflection and Learning articles. Factors 2,3,4 reflect the learning condition in e-book as Factor2 is specialized for Marker and Bookmark, Factor3 is specialized for Memo, and Factor4 is specialized for Getit. Based on these considerations, the following can be inferred from the feature matrix of learning topics (Table 7).

Table 7. *Feature Matrix of learning topic. F: Factor*

Learning topic	F 1	F 2	F 3	F 4
Periodic signal	0.246	0.207	0.214	0.374
Aperiodic signal	0.176	0.010	0.018	0.019
Fourier series	0.410	0.385	0.221	0.003
Fourier transform	1.309	0.294	0.318	0.610
Continuous time signal	0.167	0.583	0.121	0.212
Sampling	0.544	0.337	0.039	0.348
Discrete-time signal	0.260	0.722	0.229	0.387
Z-transform	0.791	0.122	0.091	0.410
Inverse Z-transform	0.254	0.013	0.156	0.067
Discrete time system	0.308	0.046	0.134	0.211
Linear time-invariant system	0.163	0.106	0.206	0.418
Difference equation	0.155	0.038	0.151	0.304
Frequency characteristic	0.249	0.069	0.018	0.151
Discrete Fourier transform	0.376	0.092	0.000	0.038
Fast Fourier transform	0.308	0.002	0.000	0.076
Digital filter	0.097	0.006	0.000	0.000
Impulse response	0.283	0.029	0.123	0.575
System function	0.083	0.032	0.000	0.062
Analog signal	0.143	0.663	0.189	0.080
Sampling cycle	0.008	0.170	0.015	0.286
Quantization	0.113	0.151	0.023	0.130
Image recognition	0.000	0.000	0.000	0.025
DC component	0.000	0.230	0.080	0.000
Euler's formula	0.017	0.075	0.045	0.087
Fourier coefficient	0.000	0.365	0.162	0.166
Amplitude spectrum	0.000	0.000	0.063	0.143
Phase spectrum	0.000	0.000	0.063	0.143
Impulse function	0.084	0.011	0.141	0.127
Step function	0.000	0.009	0.043	0.046
Aliasing	0.081	0.017	0.000	0.013
Nyquist frequency	0.154	0.064	0.000	0.064
Laplace transform	0.157	0.061	0.000	0.013
Transfer function	0.306	0.082	0.054	0.279
Adder	0.049	0.042	0.003	0.069
Delayer	0.043	0.039	0.025	0.111
Phase response	0.090	0.009	0.000	0.091
Amplitude response	0.096	0.012	0.000	0.088
Discrete-time Fourier transform	0.249	0.133	0.002	0.043
Convolution	0.136	0.044	0.131	0.149
Conjugated	0.046	0.009	0.000	0.015
Conversion pair	0.006	0.146	0.179	0.234
Symmetry	0.006	0.036	0.544	0.000
Linearity	0.125	0.098	0.629	0.032
Complex	0.342	0.200	0.110	0.638
Filter	0.358	0.127	0.034	0.255
Linear phase	0.042	0.027	0.019	0.090
Window function	0.066	0.000	0.045	0.050

- Due to the high values of Factor 1 of “Fourier Transform” and “Z-Transform,” many Reflection and Learning articles on these two topics were posted.
- Since Factor 2 of “discrete time signal,” “analog signal” and “continuous time signal” were high, many markers and bookmarks were recorded on the pages of the e-book containing these learning topics.
- Since Factor 3 was high for “Symmetry” and “Linearity,” many memos were written on the pages of the e-book containing these learning topics.
- Since Factor 4 was high for “Complex,” “Fourier Transform,” and “Impulse Response,” many students pressed “Getit” button on these topics, which means these topics were understood well.

It can be assumed that the learning topics such as “Fourier Transform,” for which the values of all four Factors were at a high level, were well understood as a result of the students studying them well. On the other hand, learning topics such as “Fourier Series,” of which the values of other factors were high but only Factor 4, specialized in “Getit,” was low, had many learning activities but students’ level of understanding was low. This suggests that the teacher's explanations of these learning topics were imperfect. However, learning topics with low values of all four Factors, such as “Aliasing,” might have been considered unimportant by teacher originally, so students hardly learned about it. Even though the value of Factor 4 was low, it is not necessary to enhance the explanation of these learning topics.

Table 8. *Difference of Factor 1-4.*

	Factor 1	Factor 2	Factor 3	Factor 4
Topics of the first half	0.198	0.179	0.137	0.172
Topics of the second half	0.177	0.047	0.043	0.154

In addition, there are different tendencies in the learning condition of the learning topics that appeared in the first half and the second half of the lesson. Specifically, as shown in Table 8, we plotted the average values of Factors 1 to 4 of learning topics of the first half and the second half of the lesson. From Table 8 it can be observed that the learning activities of the learning topics that appeared in the first half of the lesson were more active than in the second half. As an interpretation of these learning patterns, in the first half of the lesson, students were encouraged by teacher to actively post and draw markers, etc., but in the second half, students became bored or accustomed to the lesson, and their learning activities decreased due to the loss of motivation.

Feeding back the above results to the teacher may facilitate teachers' adjusting and improving the lesson content and lesson plan.

5.2 Experimental results of NTF

Although we interpreted several learning patterns using the two-dimensional matrix of learning topics and learning activities, we could not consider students, learning topics, and learning activities at the same time. Therefore, in this section, we applied NTF to the three-dimensional tensor in Subsection 3.3 and interpreted the learning conditions of the three axes of the tensor.

As a result of formatting, tensor X of students (82 people), learning topics (47 types), and learning activities (12 types) was created. By taking the strategy of balancing the error reduction and ease of interpretation as mentioned in Subsection 5.1, the size of the core tensor was set as 4 factors for learning activities, 4 factors for students and 3 factors for learning topics. NTF yielded feature matrices of learning activities, students, and learning topics of sizes 12*4 (Table 9), 82*4 and 47*4, respectively.

In the feature matrix of learning activities (Table 9),

Table 9. *Feature Matrix of learning activities. B: e-book, R: Reflection, L: Learning article, F: Factor*

Learning Activity	F 1	F 2	F 3	F 4
(B) Marker	0.001	0.065	0.124	0.544
(B) Memo	0.026	0.000	0.000	0.051
(B) Bookmark	0.000	0.002	0.000	0.283
(B) Getit	0.010	0.003	0.417	0.000
(B) Notgetit	0.119	0.000	0.003	0.021
(B) Reading time	0.253	0.004	0.041	0.214

Learning Activity	F 1	F 2	F 3	F 4
(R) Number of articles	0.313	0.323	0.002	0.001
(R) Number of characters	0.144	0.417	0.000	0.000
(L) Number of articles	0.185	0.003	0.015	0.112
(L) Number of characters	0.227	0.000	0.001	0.127
(L) Number of graphs	0.040	0.000	0.018	0.005
(L) Number of likes	0.054	0.004	0.053	0.000

- Factor 1 specializes in learning activities of Reflection management system and Learning article management system and reading time of e-book.
- Factor 2 specializes in leaning activities of Reflection.
- Factor 3 specializes in times "Getit" was pressed in the e-book.
- Factor 4 specializes in Markers and Bookmarks added in the e-book.

In the feature matrix of students, (The top 5 students were selected for each factor.)

- Factor 1 specializes in learning activities conducted by students with IDs 060, 038, 067, 057, 055.
- Factor 2 specializes in learning activities conducted by students with IDs 051, 041, 050, 049, 079.
- Factor 3 specializes in learning activities conducted by students with IDs 020, 008, 075, 040, 063.
- Factor 4 specializes in learning activities conducted by students with IDs 015, 063, 011, 031, 062.

In the feature matrix of learning topics, (The top 2 learning topics were selected for each factor.)

- Factor 1 specializes in learning activities conducted for "Fourier Transform" and "Z-transform."
- Factor 2 specializes in learning activities conducted to "Analog Signal" and "Sampling."
- Factor 3 specializes in learning activities conducted to "Discrete Time Signal" and "Complex."

Elements of the core tensor show the weights of the combinations of the factors of the feature matrices along the three axes of the original tensor. After extracting the meaning of the factors of each feature matrix as above, we examined the values of every element of the core tensor and revealed what kinds of combinations of learning activities, students, and learning topics appeared most.

In this study, the core tensor is a 4*4*3 three-dimensional tensor. So, there are 48 elements in total. The three-dimensional coordinates represent the factors of learning activity, student, and learning topic, in that order. For example, (1, 2, 3) is a combination of Factor 1 for learning activity, Factor 2 for student and Factor 3 for learning topic. Each element has a weight, and the larger the weight, the more often the combination appeared. We pick the top 2 of the 48 elements and explain their meaning.

- Element (1, 1, 1) Weight=10.971

This is the most frequently appearing combination, where students with IDs 060, 038, 067, 057, 055 conducted learning activities including writing Reflection, Learning articles, and reading e-books on the learning topics of "Fourier Transform" and "Z-Transform."

- Element (2, 3, 1) Weight=6.477

The second most frequent combination was students with IDs 020, 008, 075, 040, 063 writing Reflection articles about "Fourier Transform" and "Z-Transform."

In this way, it is found that NTF can realize learning pattern analysis that simultaneously considers the learning condition of three perspectives of students, learning topics and learning activities, As shown in Subsections 5.1 and 5.2, unlike "student-based" learning log analysis, the "learning-topic-based" learning log analysis shows how to investigate each learning topic's learningconditions and provides a new method of extracting learning patterns, which can help teachers to adjustand improve their teaching plans. Furthermore, by formatting the tensor of students, learning topics, and learning activities, it is easy for us to observe detailed learning behaviors.

6. Conclusion

In this paper, we proposed a method of converting original "student-based" learning logs into "learning-topic-based" logs, and then reshaped them into two-dimensional matrices of learning topics and learning activities and three-dimensional tensors of students, learning topics, and learning activities. Last, we applied NMF to the matrix and NTF to the tensor and extracted learning patterns. The results of analysis

clarified that the characteristics of the huge amount of learning logs can be extracted and interpreted relatively easily using NMF and NTF.

These proposed methods will contribute to improving the learning efficiency of students and enhancing future lesson plans for teachers.

Future tasks include improving the extraction method of learning topics and the way to decide the number of factors before applying NMF and NTF. In this study, the number of nouns extracted from the "Digital Signal Processing" e-book was about 20000, the number of words remaining after duplicate nouns, alphanumeric characters, and single-character words were automatically removed was about 700. However, we removed those that were not related to the lesson content manually to obtain a final 47 learning topics. Since the learning topic selection criteria include subjective standards, the setting of quantitative criteria should be improved in future studies.

In addition, not only in this study but also in all studies applying NMF and NTF, it is necessary to determine the number of factors considering the ease of interpretation of the characteristics of the data while suppressing the error. However, quantitative criteria for determining the number of factors have not been formulated, which remains a task for future studies.

Acknowledgements

This work was supported by JST AIP Grant Number JPMJCR19U1, JSPS KAKENHI Grand Number JP18H04125 and JP22H00551, Japan.

References

Hiroaki Ogata, Chengjiu Yin, Kousuke Mouri, Misato Oi, Atsushi Shimada, Fumiya Okubo, Masanori Yamada, and Kentaro Kojima. Accumulation and analysis of learning logs for utilization of educational big data. *Journal of Japanese Society for Information and Systems in Education*, 33(2):58–66, 2016.

Il-Hyun Jo, Dongho Kim, and Meehyun Yoon. Analyzing the log patterns of adult learners in lms using learning analytics. *LAK '14: Proceedings of the Fourth International Conference on Learning Analytics And Knowledge*, pages 183–187, 2014.

Niki Gitinabard, et al. Your actions or your associates? predicting certification and dropout in moocs with behavioral and social features. *EDM2018*, 2018.

Kyosuke Hasumoto, Gendo Kumoi, et al. Predicting customer lifetime value in platform businesses using non-negative matrix factorization. *Journal of Information Processing*, 60(7):1283–1293, 2019.

Masashi Kuwano, Reo Kinoshita, Kei Fukuyama, Keishi Tanimoto, Kazunori Sugahara, and Takao Kawamura. Analysis of bus usage characteristics from search log data using non-negative tensor factorization. *Journal of Japan Society of Civil Engineers D3*, 73(5):I_569–I_578, 2017.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Amnon Shashua et al. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799, 2005.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Assessment of At-Risk Students' Predictions From e-Book Activities Representations in Practical Applications

Erwin D. LOPEZ Z.[a*], Tsubasa MINEMATSU[a], Yuta TANIGUCHI[a], Fumiya OKUBO[a] & Atsushi SHIMADA[a]
[a]*Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan*
*lopez.zapata.erwin.242@s.kyushu-u.ac.jp

Abstract: The use of e-book reading systems such as Bookroll, and their ability to record readers' activities allows the design of predictive models capable of identifying at-risk students from their reading characteristics. Even though previous works have obtained promising results in this task, these results may not evidence the expected prediction performance in practical applications due to their selected assessment methods. Accordingly, in this paper, we assess this performance in two practical scenarios. The first is when we keep stored data from previous years of our course which can be used to train our model, and the second is when we only have data from a different course to use in this training process. In order to obtain a more accurate assessment, we collected 92,574 samples of predictive performances from different models under the above-mentioned conditions. We also considered different feature representations along with variational latent representations, which can leverage our previous data to automatically design general hidden features. From our results, we understand that in the first condition we can expect a relatively good predictive performance, especially when using variational latent representations. However, in the second condition we found that even when using them, the predictive performances are very limited resulting in an impractical solution.

Keywords: e-book data, at-risk students' prediction, model generalizability, VAE

1. Introduction

The adoption of digital textbook reading systems in learning environments has opened the possibility to analyze the students' behaviors and predict their learning outcomes based on the action logs they leave in these systems (Flanagan & Ogata, 2017). An example of these reading systems is Bookroll, an application used in university classrooms for distributing class reading materials. This e-book application also includes features that provide students with a variety of digital interactions, such as adding memos, jumping to an arbitrary page, highlighting text, etc. (Flanagan & Ogata, 2017; Ogata et al., 2017)

From previous works, we know that Bookroll users can be modeled as a set of features extracted from their set of action logs. These features are usually defined as some actions' frequencies (e.g., the number of times that a student uses the highlight function of the application), counted from a certain predefined subset of the total data (e.g., the subset of in-class activities), and selected after conducting a features engineering process. By using their own designed feature representations, different works have attempted to use Machine Learning (ML) supervised models to predict at-risk students (defined as students who get a low score) and obtained promising results (Akçapinar et al., 2019; Hasnine et al., 2018; Huang et al., 2020; Murata et al., 2021; Okubo et al., 2018; Yin et al., 2019). However, apart from Huang et al. (2020), these works limited their evaluation scope to a similar dataset (e.g., subsets of data from the same course) without considering test subsets from different datasets (low variance in the data).

While all these developments are useful in the design of strategies for identifying at-risk students with predictive models, in practice, we are limited to using past data to train our models, which means that our data may preserve some differences. Considering that a usual methodological issue in Learning Analytics (LA) is the presence of biases (Pelanek, 2020), we need to verify whether the reported predictive performance in the literature applies to practical implementation. This assessment is

also in line with the idea of designing a model robust enough to still work with a dataset generated from a different educational environment proposed by Baker (2019) in a LAK 2019 keynote, and by Flanagan et al. (2018) in a workshop from ICCE2018 on predictive models based on Bookroll data.

Consequently, in the present work, we employ an improvement of the best assessment method found in the literature to explore the generalizability and practical performance (predictive performance in a practical condition) of different models and existing sets of features. Additionally, since previously designed feature sets do not necessarily represent the best choices, based on the papers (Le et al., 2018, Lopez et al., 2022) we implement Variational Autoencoders (VAE) and Semi-supervised VAE (Semi-VAE) to generate latent representations and assess the general performance achieved by them.

On that account, our research questions are as follows.

RQ1: In a university course supported by the Bookroll application and in a practical scenario where we use past data from the same course, to what extent can we predict the presence of at-risk students?

RQ2: In a university course supported by the Bookroll application, to what extent can we use past data collected from a course with different characteristics to predict the presence of at-risk students?

RQ3: Can VAE latent representations improve the predictive performance in practical scenarios?

2. Previous Work

2.1 Feature representations of Bookroll data

As mentioned in the previous section, there is found a trend of works that employed ML supervised models to map their designed set of features to the probability of being an at-risk student. We identify a total of 52 possible features through these works with 4 prevalent sets of features, which are designated as Feature Sets 1 to 4 ("FS1", "FS2", "FS3", and "FS4") in the present paper. Specifically, Hasnine et al. (2018) and Akçapinar et al. (2019) make use of similar features related to the engagement (e.g., number of sessions, short events, etc.), included in "FS1". Yin et al. (2019) present a unique set that considers hidden features (e.g., the ratio between the number of PREVs and NEXTs) included in "FS2"; while Huang et al. (2020) and Cheng et al. (2021) employ larger sets of basic features (e.g., number of PREV, NEXT, ADD_BOOKMARK, etc.) considered in "FS3". The "FS4" includes the features used by Okubo et al. (2018) and Murata et al. (2021), which consists of the number of ADD_BOOKMARKs, ADD_MEMOs, read slides, and the total number of events. We should point out that these last works also use data from sources other than Bookroll. However, we include their set of features because only these works employ time-aware models such as RNNs, and therefore, their designed features can be important when assessing these models' performances. Finally, in order to implement VAE models, we also consider all the existing features in the "ALL" Features Set.

2.2 Variational Autoencoders

Since each course has particular characteristics, the same feature may have different meanings in different datasets. Accordingly, the best feature selection may depend on the dataset characteristics.

On that account, we also explore the automatic feature selection with the use of VAE. This model is a special type of autoencoder where the encoder maps the input vector to a probability distribution of latent representations instead of a fixed one. The term "Variational" comes from the Variational Bayesian approach employed to define the optimization equations required to learn the parameters of this probabilistic model without facing intractability problems (Kingma & Welling, 2014). The VAE can be trained just as an autoencoder that learns the mean and standard deviation of their latent probabilistic distribution (Figure 1). The VAE Loss function is shown in equation 1, where D_{KL} is the Kullback-Leibler divergence that decreases when $q_\varphi(z|x)$ is near to be a normal distribution, and β is a hyperparameter introduced by Higgins et al. (2017) to avoid entanglement in the latent space.

$$LOSS = MSE(X, X") + \beta D_{KL}\left(q_\varphi(z|x)||p(z)\right) \quad (1)$$

Although our previous work was the first approximation to the use of VAE models for modeling Bookroll users' data (Lopez et al., 2022), different works in the LA field have used these models earlier. One example is that of Du et al. (2020), who implement a VAE for modeling at-risk students' latent representations and from them generate new samples of at-risk students to address the data imbalance issue presented in the K-12 datasets. Also, Ding et al. (2019) model MOOCs users with the use of a modified LSTM-AE to avoid the challenging process of designing handcrafted features that are effective for their prediction task. In this case, the authors use the model to substitute the feature representations with latent representations when training a predictive model, a method that allows an improvement of its prediction accuracy and reduces overfitting. Similar to this work, Bosch (2017) uses a VAE to extract the latent representations and feed them into a predictive model to detect the students' boredom. Its results also show a performance improvement when compared to predictive models that used manually designed features.

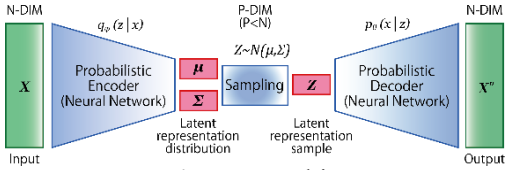


Figure 1. VAE architecture.

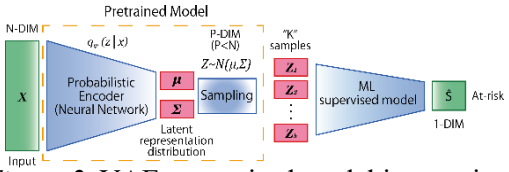


Figure 2. VAE-supervised model integration.

The data pipeline implemented by Ding et al. (2019), Bosch (2017) with educational data, and also by Lopez et al. (2022) with Bookroll data is shown in Figure 2. It can be described as the integration of two models, the first consisting of a trained VAE encoder and the second consisting of an ML-supervised model. Here the trained VAE encoder will apply non-linear transformations to the input vector to obtain its latent representation, and the ML supervised model may learn the relations between this representation and the probability of being an at-risk student. Since the VAE encoder is first pretrained on the task of reconstructing the data, it can encode any input data into general latent features.

2.3 Predictive models and methods

The works from Section 2.1 have examined the predictive performance of different ML supervised models to predict at-risk students from their Bookroll data traces (Akçapinar et al., 2019; Hasnine et al., 2018; Huang et al., 2020; Murata et al., 2021; Okubo et al., 2018; Yin et al., 2019). In Table 1 the reader can observe the models present in the literature and the number of times that have been considered by different works. These frequency values are based not only on a review of the mentioned works but also on the summary presented by Flanagan et al. (2018) from a workshop on learning performance prediction with Bookroll data.

Table 1. *ML supervised models in the literature.*

Model	Frequency
BartMachine	1
CN2Ruler Inducer	1
Generalized Linear Model	1
Rule-Based Classifier	1
AdaBoost	2
Gated Recurrent Unit	2
k-Nearest Neighbors	4
Logistic Regression	5
XGBoost	6
Decision Tree	6
Naïve Bayes	6
Neural Networks	8
Random Forest	8
Support Vector Machine	10

Since these works have adopted different predictive performance metrics and used different datasets, their results cannot be compared. However, from the variety of optimal models found in them, we can infer that even when delimiting our problem as an at-risk students' prediction using Bookroll data, no one model is the best. For this reason, is also difficult to assert if any of the possible set of features in the literature is better at encoding the information about the students' behaviors.

Furthermore, as regarded by the bias-variance trade-off (Kohavi & Wolpert, 1996), there is a chance that some features or models are good at making predictions in datasets with the same characteristics as their training one, but struggle to work with slightly different datasets. This causes assessment methodologies based on two datasets from the same course, and on two datasets from different terms of the same course to lead to different performance results. With this in mind, we should also review and analyze the assessment methodologies present in the literature.

First, Hasnine et al. (2018), Akçapinar et al. (2019), and Murata et al. (2021) only have data from a single course, and consequently, employ a cross-validation method to obtain their results. This methodology avoids a possible bias generated when selecting the train and test splits. However, it is difficult to account for the predictive performance gain due to the lack of variance in data, meaning that their results may be different from those obtained when using past data for training.

Although Huang et al. (2020) do not detail their methodology, we can infer from their results that they assessed an already trained model across 17 different datasets. Some of their performance results suggest that the LA bias problem prevents models to work with different datasets.

Finally, Okubo et al. (2018) and Cheng et al. (2021) implement a methodology based on using data from different sections of the same course. Specifically, the first work uses 15 different datasets and applies cross-validation for each course. In other words, for each validation loop, they define the data from 14 different sections as a train set and the remaining one as a test set. In contrast, the second prepared two datasets, each one corresponding to a course section, and applies 3-fold cross-validation with one of them to generate their train-test subsets, while using the other to assess them. This method is the best designed to avoid possible biases since it considers an additional test set taken from another environment (we will refer to it as assess_test). However, similar to the first kind of work, their results do not address the case when its assess_test contains past data or data with different characteristics.

3. Dataset

We used four datasets previously collected by our laboratory, of which the first one is managed internally, and the other three have been released to the public in the LAK22 Data Challenge Workshop. The datasets details are provided below.

DS1: A dataset containing 5,377,001 logs of unlabeled data from 2054 Kyushu University students enrolled in 16 different courses for the years 2019 and 2020. The considered courses have different characteristics, such as their duration, the professor in charge, their corresponding academic department, etc. The dataset does not preserve these details but contains information about their class schedule.

A-2019: A dataset containing 129,358 logs of labeled data from 50 Kyushu University students enrolled in the course “Programming Theory” in the year 2019.

A-2020: A dataset containing 147,452 logs of labeled data from 62 Kyushu University students enrolled in the course “Programming Theory” in the year 2020.

B-2020: A dataset containing 197,593 logs of labeled data from 90 Kyushu University students enrolled in the course “Cybersecurity Basic Theory” in the year 2020.

We detail some additional characteristics regarding these datasets in Table 2.

Table 2. *Datasets additional characteristics*

Characteristics	A-2019	A-2020	B-2020
# Students	50	62	90
# At-risk students	19	16	15
# Materials	21	12	9
Weeks duration	8	8	7
Professor ID	SA	SA	MT

4. Method

4.1 Experiment Preparations

The labels of our labeled datasets consisted of the students' final grades. We should mention that this grade is specific to its corresponding course. Similar to previous works, we considered the students with low scores (D and F) as at-risk students and prepared our datasets based on these new labels. Then, we prepared feature representations of the data considering FS1, FS2, FS3, FS4, and ALL sets. Also, we trained VAE and Semi-VAE models to generate data latent representations.

4.1.1 VAE latent representations

We first obtained the feature representations of dataset DS1 and then trained a VAE for each set. Each VAE training process was based on two steps. In the first, we made an approximation of the optimal architecture, the number of latent dimensions, and the values of the hyperparameters with an explorative studio with the support of the optimization framework Optuna (Akiba et al., 2019). In the second step, we manually set these values with a visual inspection of the training and validation loss values. For these two steps, we applied z-score normalization across the feature columns and randomly split the 2054 samples into train and validation datasets to avoid overfitting.

4.1.2 Semi-Supervised VAE representations

As the reader may note, the VAE models can eliminate information relevant to the at-risk students' prediction task in their search for better generalizability. The Supervised VAE proposed by Lopez et al. addresses this problem with the use of labels in the whole dataset. However, since we cannot access the labels of DS1 (unlabeled data), we need a model able to use only the existent labels. On that account, we propose a Semi-Supervised VAE model, which has the same architecture as a Supervised VAE (See Figure 3) but defines its prediction loss as shown in Equation 2.

$$LOSS = MSE(X, X'') + \beta D_{KL}\big(q_{\varphi}(z|x)||p(z)\big) + \gamma \sum_{i=1}^{k, s_i \neq -1} s_i \log \widehat{s_i} \qquad (2)$$

Therefore, if we label with a masking value of −1 all the unlabeled samples, the model will be able to only consider the samples with at-risk probability labels to calculate its prediction loss and hence prevent itself to eliminate the information of features related to the prediction performance.

Accordingly, we considered DS1 as our unlabeled dataset and A-2019 as our labeled dataset. We generated the masking labels to DS1 and applied the same methodology described in the previous section to obtain 10 additional Variational models.

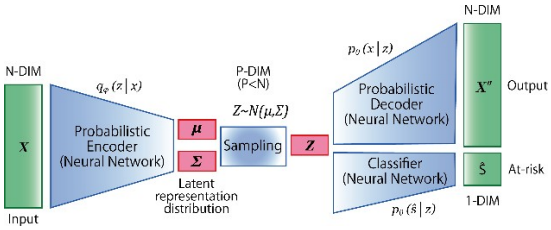


Figure 3. Semi-Supervised VAE architecture.

4.2 Models

Since no one model is the best for our application, we selected 7 models from Table 1. Our selections are the Support Vector Machine (SVM), Naive Bayes (NB), Logistic Regression (LR), Decision Trees (DT), XGBoost, Neural Networks (DNN), and Gated Recurrent Unit (GRU). We selected the XGBoost over the Random Forest since they share the same architecture but the first usually outperforms the second. Also, we included the GRU because is the only one that considers the time dimension.

4.3 Experiment

We conducted two similar experiments. In the first experiment, we select one Feature Set (e.g., FS1) to represent our different datasets. Then, similar to the best-found assessment method of Cheng et al. (2021), we use 3-fold cross-validation to split the A-2019 dataset into train-test subsets. Next, we train one selected model (e.g., DNN) with these subsets and collect its test performance results. After that, we use the A-2020 dataset as an assess_test subset (Section 2.3) to collect a new predictive performance of the same model. We finally repeat this process across all our selected models, Feature Sets, and latent representations of our Feature Sets (generated by our VAE models, Section 4.1.1 and 4.1.2). In the second experiment, we conduct the same process using the B-2020 dataset as the assessment dataset. The reader may note that the first experiment is intended to answer RQ1 since we use past data to make predictions in a current course, while the second can answer RQ2 since the A-2019 and B-2020 courses exhibit different characteristics (especially, the professor in charge and delivered content).

5. Results and discussion

5.1 Results of models trained on feature representations

From the two described experiments, we collect 92,574 samples of predictive performance results from different models trained on different feature and latent representations. To address RQ1 and RQ2 avoiding a bias due to our proposed latent representations, we consider only the results from models trained on the different Feature Sets. In Figure 4, the reader can see the distribution of the assessment predictive performance (F-score) samples collected in experiment 1 (N = 10,363) and experiment 2 (N = 10,363). We also plot the test predictive performance samples distribution (N = 10,363) to allow verification that our results are in line with the test performances reported in the literature.

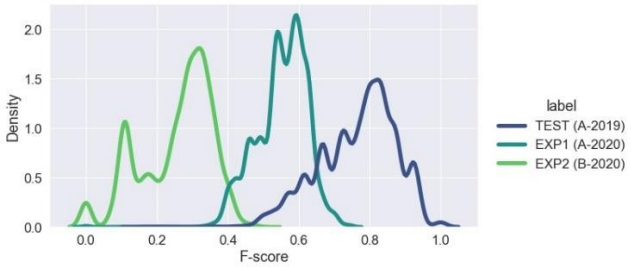


Figure 4. Test and assess_test subsets predictive performances.

From the experiment 1 results, we understand that when using past data to train our models, even if it is from the same course, we should expect a significant decrement (around 0.2 F-score in this case) in the predictive performance in contrast to our k-fold cross-validated test results. However, the mean of the experiment 1 distribution is 0.552 F-score, which is higher than the expected F-score of a non-predictor model (model that always predicts that a student is not at risk, 0.41 F-score for experiment 1), meaning that our designed models are still useful.

On the other hand, the experiment 2 results suggest that using past data from a different course may not be of use for making predictions, since their mean distribution of 0.257 F-score is lower than the expected F-score of a non-predictor model (0.286 for experiment 2).

Finally, although the literature suggests there is no optimal model, we account for the differences between models in our results. To this end, we compared the distributions from different models (N = 300) when using a fixed Feature Set and validated this comparison with a T-Test ($p<0.05$). We then ranked these models (The 1st order means the best model, the largest order the worst, and the same order indicates there is no statistical difference) as shown in Figure 5.

We observe that only the GRU model gets better results in all sets, meaning that we can rely on improving the test performance of our GRU models to also improve their performance in a real application. This result can be explained by the time-awareness of the models, meaning that the data changes in time preserve similar information in any course. Additionally, we also note that the simple models LR and SVM get better performances in the A-2020 assessment set, while there is no best model in the B-2020 assessment set. Finally, we can select only LR, SVM, and GRU to get a more accurate expected performance when using past data from the same course to train our models. Through this process, we get that their corresponding means are 0.582, 0.572, and 0.575 F-scores respectively.

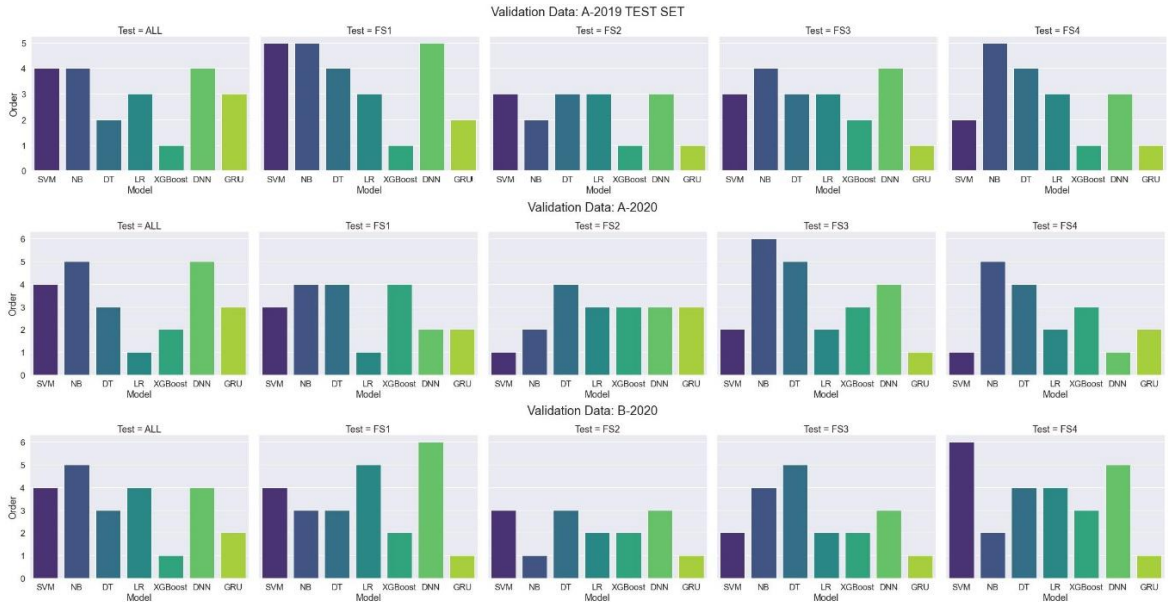


Figure 5. Models' performance rank in the test and assessment subsets.

5.2 Results of models trained on latent representations

To address RQ3, we now use also the VAE latent representations performance results. In this case, we fix a Feature Set and model to compare the results distributions of this model without latent representations (N = 300), the same model with the use of VAE (N = 300), and with the use of Semi-VAE (N = 300). We validated this comparison with a T-Test ($p<0.05$) and ranked the distributions from 1st to 3rd, assigning the same order when there is no statistical difference. The summarized results are shown in Figure 6.

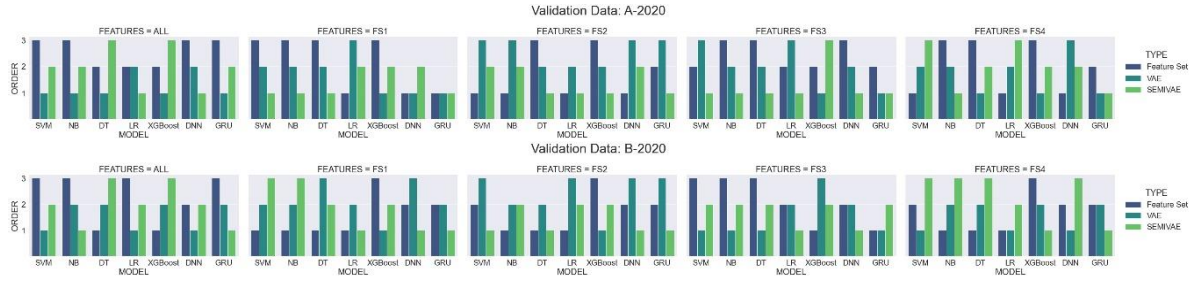


Figure 6. Feature Set, VAE, and Semi-VAE performance rank in the assessment subsets.

We observe that in general the VAE and Semi-VAE latent representations lead to models with better results in the A-2020 assessment data, however, in dataset B-2020 is difficult to determine which representations are better. If we consider each model mean as a sample and compare the distributions (N=7), in the A-2020 assessment data we get that the VAE and Semi-VAE performance are better than the non-latent representation ($p=0.0362$, $p=0.0191$ respectively). However, if we conduct the same process in the B-2020 data we get no statistical difference ($p=0.5506$). In addition, we see in Figure 6 (A-2020 data) that Semi-VAE tends to get better results more often, however, we do not have the statistical validation to claim that it is better than VAE ($p=0.3418$).

Finally, we report in Table 3 some of the better results obtained by VAE and Semi-VAE representations in the A-2020 assessment data. With also these last results in consideration, in response to RQ1, we conclude that in a university course supported by the Bookroll application and in a practical scenario where we use past data from the same course, we can still make predictions about the presence of at-risk students and expect an F-score 0.2 higher than a non-predictor model. In response to RQ2, we conclude that considering the current state of the art, it is not possible to predict the presence of at-risk students when using past data collected from a course with different characteristics (specific differences

in this study: delivered year, course content, and professor in charge). Moreover, in response to RQ3, we conclude that VAE latent representations can improve the predictive performance in the practical scenario when we use past data of the same course. However, we cannot assert whether VAE or Semi-VAE latent representations are better for this task.

Table 3. *Some of the highest performance distributions means (F-score)*

Model	Feature Set	Non-latent	VAE	Semi-VAE
DT	ALL	0.5529	0.6257	0.5334
GRU	ALL	0.5462	0.6215	0.6113
SVM	FS1	0.5898	0.6158	0.6185
NB	FS1	0.5806	0.6197	0.6333
GRU	FS2	0.5689	0.5595	0.6212
NB	FS3	0.4458	0.5740	0.6193
XGBM	FS4	0.5249	0.6294	0.5971
GRU	FS4	0.5872	0.6079	0.6070

5.3 Latent representations

The importance of the VAE models is that they can automatically select some features and how to use them to design their latent representations. In this regard, we additionally conduct a feature importance test (permutation feature importance algorithm proposed by Bisher, Rudin, & Dominici, 2019) for our variational models. The results for the VAE and Semi-VAE models are detailed in Figure 7. Though, these results are specific to our datasets and may not apply to any Bookroll dataset.

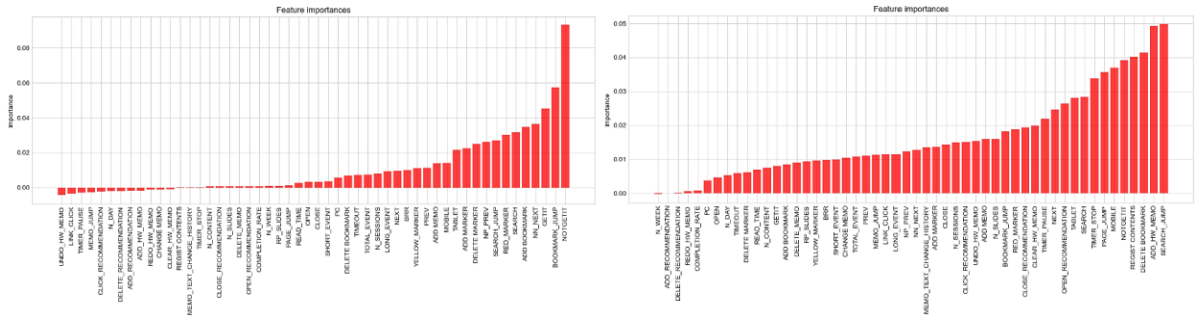


Figure 7. Left. VAE features importance. Right. Semi-VAE features importance.

We observe that the order of importance differs from the VAE to the Semi-VAE model. Since the two models are based on the same architecture, we may attribute these differences to the incorporation of prediction labels in the Semi-VAE model. To better understand these differences, we first should remember that the VAE model attempts to find general latent features that can reconstruct the data. This can be translated into two types of importance attributions. The first is based on the non-linear relationship that a feature has with other features, for example, the number of NEXT events and NN_NEXT (the number of NEXT events longer than 3 seconds) may be strongly related and the model should decide to attribute high importance to only one of these features. The second is based on the frequency of non-zero values of the feature across the students. For example, the feature ADD_HW_MEMO (adding a handwritten note) is not often used by the students and consequently, the VAE model may not attribute high importance to it. Accordingly, we may conclude, for example, that the NOT_GET_IT feature (number of times that a student pressed the button "did not understand this slide") is probably not very related to other features and relatively frequent (relatively to other independent features).

However, in the case of Semi-VAE, the model should also consider how important the feature is to make a prediction. This means that in our previous example the model may not consider that the

NOT_GET_IT feature is really important when estimating the probability of being an at-risk student. Moreover, the reader may note that the second most important feature of our Semi-VAE model is the previously mentioned ADD_HW_MEMO, which received low importance from the VAE model. This probably means that despite its infrequency, when a student uses this feature, our model can make a better prediction.

Besides the differences, we can find some similarities. One example is that redundant features such as N_WEEK and N_DAY have low importance in both models. The reason is probably that other reader's activities features are related to these features in all courses and the variational models do not need to preserve their information since it can be inferred (e.g., a student with a high number of notes, highlights, and search activities probably has a greater number of active days and weeks).

Finally, we would like to note that a Semi-VAE has more difficulty learning its mappings to the latent space since it has to deal with a trade-off between the importance of the features that have unique information and the importance of the ones that are better predictors. In this context, a features engineering process may be required to design better Semi-VAE latent representations, which in turn could lead to a statistically validated improvement of the VAE results. This idea is also supported by the results in Table 3, where the best results obtained by Semi-VAE representations take place when encoding a designed Feature Set (FS1 and FS2), in contrast to VAE, which works better encoding all possible features (ALL Feature Set).

6. Limitations

The experiment results reported in the previous section should be considered in light of some limitations. In this section, we will describe the two which we consider the most important.

The first is the optimality of the hyperparameters selected for our models. Even though we conducted several explorative trials to select the best hyperparameters, ideally, we should sample over different values for each model to avoid any bias from our selections. However, the computation time for our experimentation was more than 100 hours, which means that looping across hyperparameters would require months of computation.

The second is the collection time range of the DS1 unlabeled dataset used for training the variational models. As the reader may note, this dataset was collected in the same years of our datasets A-2019, A-2020, and B-2020. This means that our variational models were able to learn some singular characteristics present only in these years (e.g. The online activities due to the pandemic). While this ability is a good characteristic of the variational models, it also possibly introduces a bias that we cannot account for from our test and validation results. We should note that our first idea to avoid this possible bias was to use a dataset collected from 2018 and 2019. However, the datasets containing 2018 course data did not preserve their class schedule information.

7. Conclusions

From the inclusion of digital textbook reading systems in classrooms, several works have contributed to the idea of using the readers' data to identify at-risk students. The present work complements these contributions. Specifically, our results reinforce the idea that readers' interactions data preserve information related to their academic performance, but also evidence that these relationships are to a certain extent dependent on the course's characteristics. For this reason, it is difficult to use a model to make predictions across different courses. Nevertheless, we proved that the VAE latent representations and the Time-aware predictive models tend to manifest better generalizability characteristics, allowing an improvement of the predictive performance in a practical implementation.

Additionally, this paper introduces the Semi-supervised VAE to generate latent representations able to encode not only the feature relations present in different courses but also the course-specific features and their relationship with the academic performance. Since results do not provide evidence to prove or disprove that the Semi-VAE representations are better than the VAE ones, we conclude that further research is required to understand their differences. Considering that the Semi-VAE training is more complicated and easier to find a local minimum, a starting point can be to study techniques to improve their training. Some ways are to conduct a features engineering process or to design "warm-up" functions and apply them to certain loss components.

Acknowledgments

This work was supported by JST AIP Grant Number JPMJCR19U1, JSPS KAKENHI Grant Number JP18H04125, and JP22H00551, Japan.

References

Akçapinar, G., Hasnine, M. N., Majumdar, R., Flanagan, B., & Ogata, H. (2019). Developing an early-warning system for spotting at-risk students by using ebook interaction logs. *Smart Learning Environments*, 6(1):4.

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019) Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25rd ACM SIGKDD*, doi:10.1145/3292500.3330701.

Baker, R. S. (2019) Challenges for the future of educational data mining: The baker learning analytics prizes. *Journal of Educational Data Mining*, 11(1):1–17.

Bosch, N. (2017) Unsupervised deep autoencoders for feature extraction with educational data. *Deep Learning with Educational Data Workshop at the 10th International Conference Educational Data Mining*.

Chen, C.H., Yang, S. J. H., Weng, J.X., Ogata, H., & Su, C.Y. (2021) Predicting at-risk university students based on their e-book reading behaviours by using machine learning classifiers. *Australasian Journal of Educational Technology*, 37(4):130–144.

Ding, M., Yang, K., Yeung, D.Y., & Pong, T.C. (2019) Effective feature learning with unsupervised learning for improving the predictive models in massive open online courses. *Proceedings of the 9th International Conference on Learning Analytics Knowledge*.

Du, X., Yang, J., & Hung, J.L. (2020) An integrated framework based on latent variational autoencoder for providing early warning of at-risk students. *IEEE Access*, 8:10110–10122.

Fisher, A.J., Rudin, C., & Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. Journal of machine learning research : JMLR, 20.

Flanagan, B., Chen, W., & Ogata, H. (2018) Joint activity on learner performance prediction using the bookroll dataset. *Workshop Proceedings 26th International Conference on Computers in Education*, 487–492.

Flanagan, B., & Ogata, H. (2017) Integration of learning analytics research and production systems while protecting privacy. *Workshop Proceedings 25th International Conference on Computers in Education*, 355–360.

Hasnine, M., Akçapınar, G., Flanagan, B., Majumdar, R., Mori, K., & Ogata, H. (2018) Towards final scores prediction over clickstream using machine learning methods. *Proceedings of the 26th International Conference on Computers in Education*, 11.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., … Lerchner, A. (2017) beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*.

Huang, A.Y.Q., Lu, O.H.T., & S. J. H. Yang. (2020) Evaluation of classification algorithms for predicting students' learning performance based on bookroll reading logs. *Cognitive Cities*, 262–272.

Kingma, D. P., & Welling, M. (2014) Auto-encoding variational bayes. *CoRR,* abs/1312.6114.

Kohavi, R., & Wolpert, D.H. (1996) Bias plus variance decomposition for zero-one loss functions. *ICML*.

Le, L., Patterson, A., & White, M. (2018) Supervised autoencoders: Improving generalization performance with unsupervised regularizers. Advances in Neural Information Processing Systems, volume 31.

Lopez, E. D., Minematsu, T., Taniguchi, Y., Okubo, F., & Shimada, A. (2022) Exploring the use of probabilistic latent representations to encode the students' reading characteristics. *CEUR Workshop Proceedings*, 3120.

Murata, R., Minematsu, T., & Shimada, A. (2021) Early detection of at-risk students based on knowledge distillation rnn models. *Proceedings of the 14th International Conference on Educational Data Mining*, 699–703.

Ogata, H., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., … Hirokawa, S. (2017) Learning analytics for E-book-based educational big data in higher education. *Smart Sensors at the IoT Frontier*, 327–350.

Okubo, F., Yamashita, T., Shimada, A., Taniguchi, Y., and Shin'ichi, K. (2018) On the prediction of students' quiz score by recurrent neural network. CEUR Workshop Proceedings, 2163.

Pelanek, R. (2020) Learning analytics challenges: trade-offs, methodology, scalability. Proceedings of the 10th International Conference on Learning Analytics & Knowledge.

Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima K., & Ogata H. (2019) Exploring the relationships between reading behavior patterns and learning outcomes based on log data from e-books: A human factor approach. International Journal of Human–Computer Interaction, 35(4-5):313–322.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Topic-Level Social Network and Language Correlation in Course Discussion Forums

Ezekiel Adriel LAGMAY[a*] & Maria Mercedes RODRIGO[a]
[a]*Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, The Philippines*
*ezekiel.lagmay@obf.ateneo.edu

Abstract: In this paper, we characterize student contributions in online discussion forums and examine the relationship between these characteristics and students' peer-to-peer relationships. We use Coh-Metrix Language and Discourse Analysis metrics to predict the Weighted Degree and Closeness Centrality of student contributions. We performed the analysis on all students in the population under study and on a subset consisting of active students only. We found that students who have more direct connections with other students tend to have abstract word choices whereas active students also tend to be more expository and informational, have shallow ideas, and use simpler construction. We also found that all students who can easily share their thoughts to the entire class tend to have posts that are more informational, with deep and connected thoughts and ideas. Active students who belong in this group further exhibit simpler construction and more abstract word choices.

Keywords: Online Course Discussion Forums, Social Network Analysis, Language and Discourse Analysis, Linear Mixed-Effects Models

1. Introduction

In this study, we characterize student contributions in online discussion forums and examine the relationship between these characteristics and students' peer-to-peer relationships. We use Coh-Metrix Language and Discourse Analysis metrics to predict the Weighted Degree and Closeness Centrality of a student's participation in a course discussion forum (Dowell et al., 2015). We use a Linear Mixed-Effects Modelling approach that also takes into consideration the differing nature of each course discussion forum topic within that same course. Our research questions for this study are as follows:

1. For each Social Network Analysis (SNA) metric (Weighted Degree or Closeness Centrality) and learner group (all students or active students only), which of the characteristics of language best predict that SNA metric of the student in a course discussion forum topic?
2. For each SNA metric (Weighted Degree or Closeness Centrality) and learner group (all students or active students only), does its correlation with the characteristics of language remain constant regardless of the method of measuring language and discourse features (percentile or z-score)?
3. What insights can we derive from these metrics about the quality of students' contributions and their relationships with their peers?

This study is based on a previous paper by Dowell et al. (2015). In this work, the proponents first extracted language and discourse scores from each student's course discussion forum post contents using Coh-Metrix and correlated these with that student's SNA metrics Degree Centrality, Closeness Centrality, and Betweenness Centrality. The Coh-Metrix scores used as independent variables include Narrativity, Deep Cohesion, Referential Cohesion, Syntactic Simplicity, and Word Concreteness (Graesser, McNamara, & Kulikowich, 2011; Graesser & McNamara, 2011; Kintsch, 1998; Snow, 2002). Dowell et al. (2015) generated two datasets: one containing all the students in the course and the other containing only the active students (those who have made 4 or more posts) (De Laat, Lally, Lipponen, & Simons, 2007; Gillani, Yasseri, Eynon, & Hjorth, 2014). A Linear Mixed-Effects Modelling approach was used with the Coh-Metrix scores as the independent variables, the SNA metrics as dependent variables, and learner (User ID) and word count as random effects (or data groupings) (Dowell et al., 2015; Knowles, 2013). The results showed that "more narrative style discourse with less overlap between words and ideas, simpler syntactic structures and abstract words"

positively increases the significance and centrality of a student in the social network (Dowell et al., 2015).

This study aimed to replicate the methodology used in the paper, but with the following changes:

- Usage of Topic ID as one of the random effects, and as well basing the definition of active students as those who participated in all discussion forum topics in a course, rather than basing it merely from the number of posts made by the student.
- Use Weighted Degree instead of Degree Centrality as one of the SNA metrics to be correlated.
- Experiment with the usage of either the percentile or z-score methods of language and discourse metrics to determine if their correlation with the SNA metrics still hold for both scoring methods.

> *The study made use of the discussion forum of Course A from a university in Metro Manila, Philippines which ran from August 26, 2021 to December 18, 2021. It had 20 enrolled students who made a total of 206 discussion forum posts, each grouped into seven distinct graded topics. 10 of the students have made at least one post in all seven discussion forum topics; these students are hence classified as "Active Students". Prior to the analysis, the course discussion forum dataset of Course A is split according to topic. Thus, each topic has a total of five datasets – one for post contents (which are then analyzed using Coh-Metrix), two for social network graphs (All Students and Active Students; these are then used for extracting the Weighted Degree and Closeness Centrality metrics via Gephi 0.9.3), and two for discussion topic participants (All Students and Active Students). Then prior to the analysis proper, the SNA and Coh-Metrix results for all topics are joined together into two datasets, one containing all students and another containing the active students only. In total, 6 models were generated, one for each possible Weighted Degree/Closeness Centrality-All Students/Active Students-Percentile/Z-Score combination.*

2. Results and Discussion

Weighted degree. For the All Students models, Narrativity and Word Concreteness significantly had a positive correlation and negative correlation with Weighted Degree, respectively. On the other hand, for the Active Students models, Deep Cohesion, Word Concreteness, and Narrativity had a negative correlation, while Syntactic Simplicity had a positive correlation. This could mean that students who had more direct connections with their fellow classmates in discussion forums in general created posts that had more abstract words. Students in the Active Students subgroup tended to be more expository and informational in their posts, had shallow ideas, and used simpler construction. Students in the All Students subgroup who tended to be more narrative were also more likely to have more direct connections with others.

Closeness centrality. For all students, Narrativity had a negative correlation while Deep Cohesion and Referential Cohesion both had a positive correlation with Closeness Centrality. On the other hand, for active students, the correlation rules were the same, with the addition of the negative correlation with Word Concreteness and positive correlation with Syntactic Simplicity. This meant that students who could quickly disseminate responses to others tended to create posts that were more informational, with deep and connected thoughts, ideas, and words, while active students had the additional feature of using more abstract words and simpler construction.

Percentile vs. Z-Score. Narrativity, Referential Cohesion, and Deep Cohesion still held the same correlation rules whether percentile or z-score for both Weighted Degree and Closeness Centrality for the All Students models, and Deep Cohesion and Syntactic Simplicity for Active Students models. Hence, only Deep Cohesion still held the same correlation rules regardless whether percentile or z-score is used, regardless of learner group or SNA metric. The percentile-based models also had a higher percentage of the contribution of the independent variables to the predictable variance than the z-score-based models (Dowell et al., 2015).

3. Conclusion and Further Studies

In summary, some language and discourse characteristics had the same nature of correlation for all students in general and active students in particular, on a per SNA metric basis. However, determining the set of language and discourse metrics that could best predict both Weighted Degree and Closeness Centrality is only possible for the active students. Given the small sample size and the limitation of this study to a single class, future studies may want to determine if this finding continues to hold true for larger populations and across different subject areas.

This study also shows that, with the exception of Deep Cohesion, the nature and significance of the correlation between the SNA metrics and language and discourse features would differ in general depending on whether percentile or z-score is used as basis. The significance tests showed that, in this instance, models using the percentile-based scoring system are more predictive than those using z-score-based scores as they have a higher percentage of the contribution of the independent variables to the predictable variance.

Finally, we found that students who have more direct connections with other students tend to have abstract word choices whereas active students also tend to be more expository and informational, have shallow ideas, and use simpler construction. We also found that all students who can easily share their thoughts to the entire class tend to have posts that are more informational, with deep and connected thoughts and ideas. Active students who belong in this group further exhibit simpler construction and more abstract word choices.

Acknowledgements

The authors would like to thank the Ateneo Laboratory for the Learning Sciences, Ateneo Research Institute for Science and Engineering (ARISE), and Accenture for the funding and support needed for this research. The authors would also like to thank Deni Jaramillo and Miguel Saavedra for their assistance in setting up the necessary servers for the collection of Canvas data. The authors would also like to thank Natalie Newton and Danielle McNamara for providing the Coh-Metrix desktop application for the purposes of this study.

References

De Laat, M., Lally, V., Lipponen, L., & Simons, R.-J. (2007). Investigating Patterns of Interaction in Networked Learning and Computer-Supported Collaborative Learning: A Role for Social Network Analysis. *International Journal of Computer-Supported Collaborative Learning*, *2*(1), 87-103.

Dowell, N. M., Graesser, A. C., Hennis, T. A., Skrypnyk, O., Dawson, S., de Vries, P., … Kovanović, V. (2015). Modeling Learners' Social Centrality and Performance Through Language and Discourse. In *Proceedings of the 8th International Conference on Educational Data Mining* (pp. 250-257). Madrid, Spain.

Gillani, N., Yasseri, T., Eynon, R., & Hjorth, I. (2014). Structural Limitations of Learning in a Crowd: Communication Vulnerability and Information Diffusion in MOOCs. *Scientific Reports*, *4*.

Graesser, A. C., & McNamara, D. S. (2011). Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science*, *3*(2), 371-398.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *Educational Researcher*, *40*(5), 223-234.

Kintsch, W. (1998). *Comprehension: A Paradigm for Cognition*. Cambridge, U.K.: Cambridge University Press.

Knowles, J. (2013). Getting Started with Mixed Effect Models in R. Retrieved from: https://www.jaredknowles.com/journal/2013/11/25/getting-started-with-mixed-effect-models-in-r

Snow, C. E. (2002). *Reading for Understanding: Toward a Research and Development Program in Reading Comprehension*. Santa Monica, CA: Rand Corporation.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Assessment of Students' Feedback Behavior in A Game-Based Automated Feedback System – A Cross-Cultural Replication Study

Annika SILVERVARG [a*], Kristen BLAIR[b], Maria CUTUMISU[c] and Agneta GULZ[d]
[a]*Dept of Computer and Information Science, Linköping University, Linköping, 58183, Sweden*
[b]*Graduate School of Education, Stanford University, Stanford, CA, USA*
[c]*Dept of Educational Psychology, University of Alberta, Edmonton, Canada*
[d]*Lund University Cognitive Science, Lund, 22100, Sweden*
*annika.silvervarg@liu.se

Abstract: In this paper, we argue for the importance of conducting replication studies over various schools and countries when addressing topics about learning and instruction and propose educational technology to be a tool for this endeavor. We present an example of a cross-cultural replication study that makes use of educational technology in the form of a digital game-based automated feedback system. The study addresses feedback related behavior in 11–15-year-old students in US and Swedish classrooms, investigating students' choices to seek confirmatory (i.e., positive) or critical (i.e., negative) feedback, as well as their subsequent choices to revise their work based on this feedback. Comparisons of the data collected at several schools in the US and Sweden showed similar patterns of relationships among students' feedback-seeking behavior, their tendency to revise their work, and their learning outcomes in and outside the assessment environment. Overall, the findings revealed that this assessment approach seems to be generalizable from a North American to a European population. However, the findings showed both a significant difference between Sweden and the US regarding the preference for critical feedback and between different schools within each country. Thus, it is possible that the difference between countries reflects school differences rather than cultural differences.

Keywords: feedback, self-regulated learning, assessment, educational technology, cross-cultural replication study

1. Introduction

Although research on instruction and learning using educational technology has gained momentum around the world, it is not clear how relevant, applicable, and transferrable the results are across countries. Thus, there is a need for research examining the cross-cultural validity of these findings. In psychology, researchers debated the applicability and generalizability of research conducted in WEIRD (Western, Educated, Industrialized, Rich and Democratic) populations/countries/cultures to other non-WEIRD populations/countries/cultures (Arnett, 2008; Henrich et al., 2010). A study reviewing papers published in artificial intelligence in education (AIED) and intelligent tutoring systems (ITS) conferences during 2002-2103 found that most of first authors and samples for empirical studies originated from the US (50% and 62%) followed by English-speaking countries (23% and 22%) and European countries (17% and 11%). However, non-WEIRD continents (Asia, Latin America, Africa) only accounted for 10% of the authors and 6% of the samples. Moreover, there are noted cultural differences between the American and other WEIRD societies, particularly with regards to individualism (Henrich et al., 2010).

Thus, there is a need for replication or semi-replication studies across countries to enable discoveries of patterns and contextual factors (e.g., culture) that influence study outcomes. However, such studies are rarely being conducted within educational science and AIED. One exception is Ogan et al.'s (2014) study that investigated help-seeking behavior across cultures in the US, Philippines, and Costa Rica. The researchers found that their models for effective help seeking based on data logging in a digital learning environment transferred to some degree between the US and Philippines, but not

between these countries and Costa Rica. They discuss this finding in relation to cultural dimensions (Hofstede et al., 2010) such as Individualism and Power Distance, and how differences in these were found in the interaction between students and the teacher in the classroom. Differences in Power Distance have also been raised by teachers in Sweden regarding the (limited) applicability of research done in the UK to Swedish schools (Buljubasic, 2018). In Sweden, the governmental organization (Skolverket) recommends international research as a basis for teachers' work with formative evaluation and feedback, such as work by Dylan Wiliam (cf. Black & Wiliam, 2005) and Carol Dweck (cf. Dweck 2012). However, in a blog post, the Swedish teacher Johan Kant (Kant, 2021) discusses how the grading system and relation between the teacher and the student differs between countries and how that affects the way the results from studies on feedback and assessment conducted in the UK can or cannot be applied in Swedish schools.

In this paper we present a cross-cultural study on feedback behavior in an educational game used by 11–15-year-old students in US and Swedish classrooms, which is a replication of previous studies conducted in the US. We investigate students' choices of positive (i.e., confirmatory) versus negative (i.e., critical) feedback, as well as their inclination to revise the task on which they have received feedback. Then, we examine the correlations between (1) these behaviors and learning outcomes within the educational game used in the study and (2) these behaviors and an independent measure of student success (e.g., standardized test results). All studies were based on the curriculum-independent tool, Posterlet (Cutumisu, Blair, Chin, & Schwartz, 2015), a digital assessment game in which students design posters and learn graphical design principles. We pose the following research questions:

- Do learning choices (seeking critical feedback and revising posters) vary with country and/or school?
- Do the relations between learning choices and learning outcomes within the learning environment (i.e., Posterlet) vary with country?
- Do the relations between learning choices and broader learning outcomes outside the learning environment (i.e., academic achievement, such as grades or standardized tests) vary with country?

2. Background

In this section we provide background on the value of critical feedback, and how feedback-related behavior can differ between countries and cultures. We also describe the educational game used as the research instrument in our studies.

2.1 Critical Feedback

Feedback is crucial for effective learning as evidenced by a large body of scientific literature on feedback within the learning sciences (Fyfe & Rittle-Johnson, 2015). Despite this, little is known about how learners in different contexts pay attention to and process critical informative feedback. Evidence suggests that critical informative feedback is a form of feedback that can be especially beneficial for learning (Kluger & DeNisi, 1998). One explanation for why critical or negative feedback can be more effective than positive feedback for continued performance is that positive feedback indicates that one has done enough, whereas negative feedback indicates the need for a change (Cutumisu et. al, 2015). On the other hand, negative feedback runs the risk of triggering ego threat issues that lead people to ignore or neglect the feedback (Hattie & Timperley, 2007, Cutumisu et al., 2015; Tärning et al., 2020).

2.2 Culture and feedback related behavior

To our knowledge, no studies have directly addressed cultural differences in feedback behavior in educational contexts regarding either of: (i) students' inclination to seek critical feedback, (ii) their preferences when given the choice between critical or confirmative feedback, and (iii) their inclination to revise a task on which they have received critical feedback. However, Suzuki et al. (2008) conducted

a study with multi-ethnic youths, investigating their responses to praise and criticism. The participating 36 teenagers (16-year-olds on average) were girls in 4 sports teams at US schools. All sports teams had members from three or four ethnic groups: African Americans, Asian Americans, European Americans, or Latinas. The study focused on whether there would be cultural differences in the participants' reactions to (i) being praised (for doing a job well) and (ii) being criticized (for making a mistake). The hypotheses that potential differences in responses to critical versus confirmative feedback would correspond to the degree of collectivism versus individualism in the different cultural backgrounds of the participants were confirmed by the data. Participants from more interdependent (collectivistic) cultures (Asian Americans and Latinos) were more inclined to accept negative feedback or criticism and more inclined to use it to self-correct. A reason behind this may be that members in these cultures often are socialized to receive negative feedback and use it to self-correct in order to achieve normative behavior (Greenfield et al., 2000; Kitayama et al., 1995). On the other hand, participants from the more individualistic cultures (European and African Americans) were more inclined to deny responsibility for a mistake, and more likely to assert their own point of view despite having someone else bring criticism or critical feedback.

On a broader level feedback behavior is an important aspect of self-regulated learning (SRL). Purdie et al. (1996) found both differences and similarities in beliefs about learning and use of SRL strategies for Japanese and Australian students. Both groups used the "environmental structuring" and "self-evaluating" strategies, where the latter typically includes checking, revising, and redoing one's work, using self-questioning through quizzes, or being asked questions by other people. However, the strategy least used was "reviewing tests and other work", which is closely related to getting and responding to feedback. The authors speculate that this may be due to the students not getting informative feedback on test or work. As stated earlier, positive feedback does not necessary motivate students, and negative feedback that comes without an indication of what and how work should be improved may not help students learn either.

McInerney (2008) also explores how motivation and SRL can be related to cultural differences and cultural identity and concludes that even though the construct of self-regulation appears to be universal, how it is actually realized in different cultures seems to vary (e.g., with regards to help-seeking, fear of failure, and parents' expectations). Thus, given that the way students seek and use feedback relates to several SRL constructs, it is reasonable to assume that these behaviors or choices are influenced by culture.

2.3 The research instrument – The educational game Posterlet

Posterlet is a game-based assessment that enables students to design three posters for their fictitious school's Fun Fair (Cutumisu, Blair, Chin, & Schwartz, 2015). For each poster, the environment offers students the choice of either confirmatory (e.g., "It's good you told them what day the fair is.") or critical (e.g., "People need to be able to read it. Some of your words are too small.") feedback from three animal characters (Step 4, Fig. 1) to help them learn about graphical design principles. It also measures whether students choose to revise their work after feedback (Step5-6). Posterlet is designed so that positive ("It's nice that the poster says how much the booth costs") and negative ("You didn't say how much the booth was") feedback provide equivalent informational value. The game assesses 21 graphical design principles (Cutumisu et al., 2015).

Posterlet has many similarities with other digital educational resources that collect and process data. However, there are some important differences. One is that the topic addressed by the game is curriculum independent. It is not expected that students will have encountered the specific graphical design principles previously, but rather that they will learn them through feedback while interacting with the game. The focus of measurement or assessment is feedback-related behavior, behavior that can occur and be relevant in all subjects (as well as in contexts outside of school). It is designed to be a stand-alone tool and, given the many demands on classroom time, a complete session can be as short as 10-15 minutes. A main advantage of conducting a cross-cultural replication study based on Posterlet is that it is a curriculum independent, stand-alone tool, since curricula - in all subjects - tend to vary strongly between different countries.

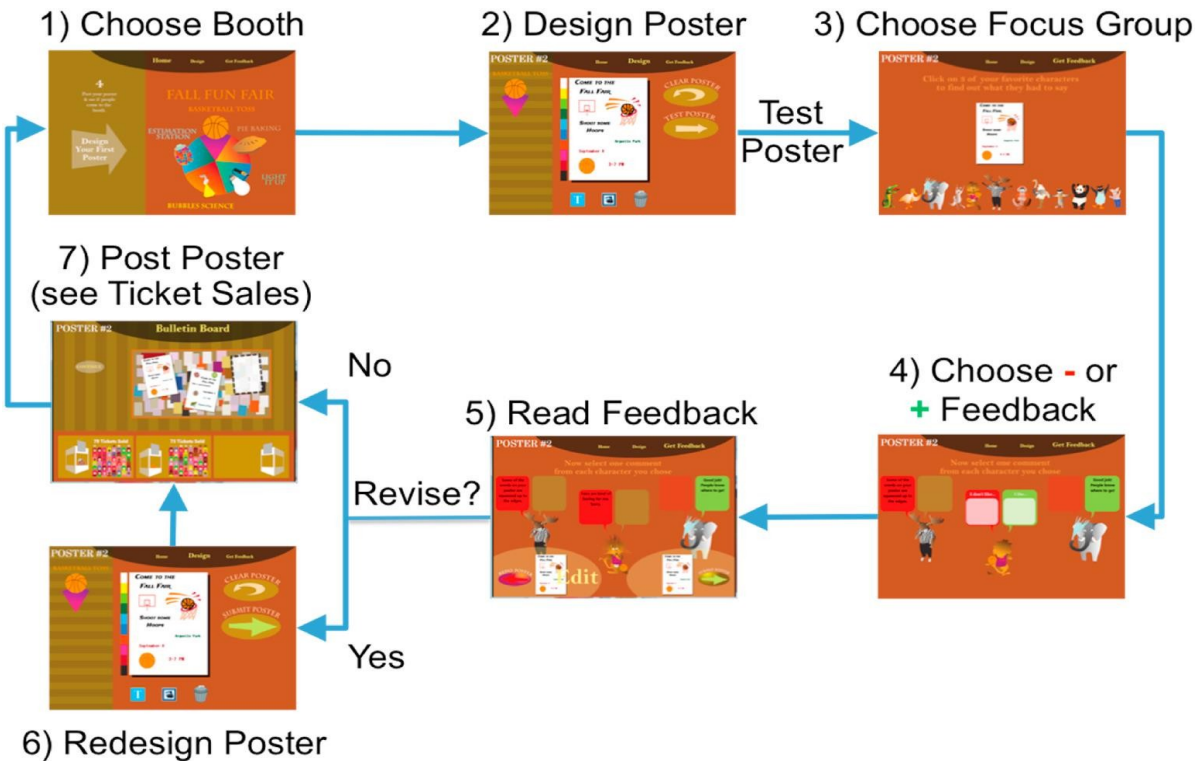


Figure 1. The steps involved in playing the Posterlet game. (Reprinted from Cutumisu et al., 2015 under Creative Commons License CC BY-NC-ND 3.0.).

2.4 Previous studies

Multiple studies using Posterlet in the US (Chin et al., 2019; Cutumisu et al., 2015, 2017; Cutumisu & Schwartz, 2021) have found consistent results across various data sets regarding the (i) relations between feedback and revising behavior (choice of negative feedback and choice to revise) and learning outcome measured in the application and knowledge of the design principles, and (ii) relations between feedback choices in the game and broader learning outcomes measured via standardized tests. These recurring patterns are that the more negative feedback students chose, the better they performed on the post-test and on the overall poster quality, even though both positive and negative feedback were equally informative. There is also a significant positive correlation between the choice to revise and post-test scores, which do not depend on the quality of the posters that students produced (Cutumisu et al., 2015, Chin et al., 2019). Secondly, the tendency to seek critical feedback also correlated with learning out-comes outside the game, such as standardized test in Mathematics and ELA (Cutumisu et al., 2015), and grades in science and math (Chin et al., 2019).

3. Method

As described above multiple studies using Posterlet in the US between 2013 and 2021 have found consistent results across various data sets. Moreover, the studies have included a variety of populations with both high and low socio-economic status, from schools with different profiles and from different geographical areas. There have also been comparisons between different ages of the students (Cutumisu & Schwartz, 2021). Thus, for our replication study we aimed for the same type of variation in recruitment of participating schools, except for age, where we limited our comparison to students in Grades 6-9 (i.e., 11–15-year-old). We strived to keep the instrument and procedure as similar as possible, described in more details below.

3.1 Participants and procedure

This study samples $N = 291$ Swedish students from five public middle/high schools in three different cities, with variation in socio-economic status. The data was analyzed and compared to data from three public middle schools from the states of California, Illinois, and New York ($N = 764$) previously reported (Chin et al., 2019; Cutumisu et al., 2015, 2017). Table 1 shows the school and participant information for the students included in the study.

Data collection took place in the participants' classrooms. Researchers provided instructions and students played the Posterlet game, followed by an online posttest on graphical design principles. Data from the game and the posttest were collected automatically online. The time spent on playing Posterlet was comparable over all studies. Swedish students played Posterlet for an average of $M = 14.8$ minutes ($SD = 4.1$), whereas USSchool2 students played Posterlet for an average of $M = 12.2$ minutes ($SD = 5.9$), USSchool1 students played Posterlet for an average of $M = 14.9$ minutes ($SD = 6.2$), and USSchool3 students played Posterlet for an average of $M = 14.9$ minutes ($SD = 4.07$).

Not all students completed the Posttest and learning outcomes such as grades and standardized test were not collected for all schools in the US (Table 1). To maximize the available data for each analysis, we used the subset of students who had complete data for each specific research question.

Table1. *Participant information*

School/ Country	**SES**	***N***	**Grade**	**Age Mean (Interval)**	**Posttest *N***	**Stand Test/ Grade *N***
SweSchool1	Low	39	7	13.23 (13-14)	39	39
SweSchool2	High	45	7	13.22 (13-14)	26	45
SweSchool3	Mid	34	8	15.18 (15-16)	22	34
SweSchool4	Mid	116	7	13.47 (13-15)	71	116
SweSchool5	Mid	52	7	13.9 (13-14)	52	52
Swe All		**286**	**7-8**	**13.68 (13-14)**	**200**	**286**
USSchool1	Low	172	6-8	12.2 (11-14)	163	65
USSchool2	Mid	272	6-9	12.1 (11-15)	226	116
USSchool3	High	89	7-8	(13-14)	-	75/57
US All		**533**	**6-9**	**11-15**	**389**	**256/57**

3.2 Measures

Three types of measures were employed in the study: (1) learning choices within the education game: choosing critical feedback and choosing to revise; (2) learning outcomes within the game: poster quality and posttest of graphical design principles; and (3) academic achievement/learning outcomes outside the digital learning environment (grades or standardized test results).

3

3.2.1 Learning choices

4

Critical Feedback (CF) measures the number of times students choose critical feedback (Step 4 of Fig. 1), and ranges from 0 to 9 (three posters and three feedback choices per poster). Revision is the number of times students chose to revise their posters, and ranges from 0 to 3 (one opportunity per poster).

3.2.2 Learning outcomes

5

Poster Quality is an in-game performance measure that scores how well the student used the 21 graphical design principles embedded in the game (-1 was assigned for an incorrectly applied principle; 1 was assigned for a correctly applied principle). Thus, a score ranging from -21 to 21 is possible for each poster, and the total measure for three posters can range from -63 to 63. Posttest measures the in-game learning of the design principles through four questions. The first is an open-response

questions where students are asked to list mistakes beginners who design a poster might make. For the other three questions students are provided a model poster that they are asked to write feedback on, either in free form (question 2) or by ticking of good or bad things about the poster relating to the design principles (question 3 and 4). The two first questions have maximum score of 21 (since there are 21 design principles) and on the last two one can score at most 5 since there are five correct answers. The total Posttest score is computed by summarizing the normalized scores (Z-scores) for each of the four questions.

1.1.3 Academic achievement

For a broader measure of learning outcomes outside the digital learning environment, data on students results on state standardized tests for Mathematics and English Language Art (for all US schools) or Math grades (for all Swedish schools and one school in US) and Swedish Language grades (for Sweden) were gathered when available.

4. Results

4.1 Do learning choices (seeking critical feedback and revising posters) vary with country and/or school?

6

We investigated whether students seek critical feedback and choose to revise to a similar or different extent. Table 2 shows the mean for the measures Critical Feedback (CF) and Revision for Sweden and for US overall, as well as for the different schools in each country.

Table 2. *Mean and standard deviation M(SD) for learning choices by Country and School*

School/Country	Critical Feedback	Revision
SweSchool1	3.6 (2.4)	0.9 (1.2)
SweSchool2	4.2 (2.2)	1.1 (1.1)
SweSchool3	4.4 (2.1)	1.4 (1.0)
SweSchool4	4.8 (2.0)	1.3 (1.2)
SweSchool5	5.1 (1.9)	1.5 (1.1)
Swe All	**4.6 (2.1)**	**1.3 (1.2)**
USSchool1	3.2 (2.1)	1.2 (1.1)
USSchool2	4.0 (2.5)	1.1 (1.1)
USSchool3	5.8 (2.1)	2.0 (1.0)
US All	**4.0 (2.5)**	**1.3 (1.1)**

A t-test comparing Sweden and US revealed a significant difference ($t(817) = 3.4566$, $p < .001$) in CF, with a small effect size (Cohen's $d = .2599$). No significant difference was found for Revision.

As previously reported (Cutumisu et al., 2015), a t-test for USSchool1 and USS-chool2 showed a significant difference ($t(402) = -3.2$, $p < .01$) for CF, but not a significant difference for Revision.

A one-way ANOVA revealed that there was a statistically significant difference in CF between at least two Swedish schools ($F(4, 285) = 3.515$, $p = .008$). Bonferroni post-hoc tests showed that the mean value of CF was significantly different ($p = .011$) between SweSchool1 ($M = 3.6$) and SweSchool5 ($M = 5.1$); it was also significantly different ($p = .025$) between SweSchool1 ($M = 3.6$) and SweSchool4 ($M = 4.8$); but it showed no significant differences between the other schools. A one-way ANOVA revealed that there was not a statistically significant difference in Revision between the Swedish schools ($F(4, 285) = 1.952$, $p = .102$).

Thus, we found differences between US and Sweden in students' critical feedback-seeking behavior. However, it is worth noting that the US range of critical feedback-seeking (3.2 to 5.8) is

greater than the range of Sweden (3.6 to 5.1). Thus, even though there are significant differences between the mean of Sweden and the US, this may relate to relative sample sizes across the different schools rather than culture, which is supported by the fact that there are significant differences between the schools within the countries.

No differences were found regarding students' choice to revise after they received feedback between the US and Sweden, nor were there differences between the Swedish schools or the USSchool1 and USSchool2.

4.2 Do the relations between learning choices and learning outcomes within the learning environment (i.e., Posterlet) vary with country?

7

To answer this question, correlations for the Swedish data were similar to those reported in the US studies (Table 3 and Table 4), such as the correlations between critical feedback-seeking (CF), Revision, Poster Quality and Posttest. The correlation coefficients for the Swedish study are presented in Table 5.

Table 3. *Correlations for USschool1 and USSchool2 study*

	N	**Revision**	**Poster Quality**	**Posttest**
CF	473	.47**	.28**	.23**
Revision	473		.34**	.24**
Poster Quality	473			.39**

Table 4. *Correlations for USSchool3 study*

	N	**Revision**	**Poster Quality**
CF	89	.40**	.37**
Revision	89		.26**

Table 5. *Correlations for Swedish study*

	N	**Revision**	**Poster Quality**	**Posttest**
CF	286	.48**	.22**	.19**
Revision	286		.21**	.23**
Poster Quality	286			.41**

To compare the correlations between the Swedish and US studies, a Fisher's *Z* test was performed. Table 6 includes *z*-scores and *p*-values for the Swedish and the US studies in USSchool2and USSchool1. It shows that the only significant difference in the correlations is between Revision and Poster Quality, which is higher for the US study. Overall, the relations between learning choices and learning outcomes were similar, regardless of culture.

Table 6. *Comparison of correlations between Swedish and US studies for USSchool1+2*

	CF-Rev	**CF-Poster quality**	**CF-Posttest**	**Rev-Poster Quality**	**Rev-Posttest**	**Poster Quality-Posttest**
Sweden	.48	.22	.19	.21*	.23	.41
US (School1+2)	.47	.28	.23	.34*	.24	.39
z-score, p	.21, .42	-.81, .21	-.05, .30	-1.82, .04	-.16, .44	.22, .41

The strongest correlation was found between CF and Revision. Thus, multiple regressions were performed to determine if CF and Revision were independent predictors of learning from the game. One regression used the Posttest as the dependent measure and one used Poster Quality.

As reported previously (Cutumisu et al., 2015), for USSchool1+2 both CF ($t(470) = 3.2$, $p = .002$, $\beta = .15$) and Revision ($t(470) = 5.4$, $p < .001$, $\beta = .25$) were predictors ($F(2,470) = 36.07$, $p < .001$, $R^2 = .13$) of Poster Quality score. Both CF ($t(411) = 2.7$, $p = .006$, $\beta = .15$) and Revision ($t(411) = 2.8$, $p = .005$, $\beta = .16$) were significant predictors ($F(2,411) = 16.11$, $p < .001$, $R^2 = .07$) of Posttest score. For USSchool3, CF ($t(88) = 3.40$, $p = .001$, $\beta = .37$) was a significant predictor of Poster Quality ($F(2,86) = 9.13$, $p < .001$, $R^2 = .17$; Chin et al., 2019).

The same regressions were conducted on the Swedish data, which showed that both CF ($t(283) = 2.39$, $p = .018$, $\beta = .16$) and Revision ($t(283) = 2.11$, $p = .036$, $\beta = .14$) were significant predictors ($F(2,283) = 9.74$, $p < .0001$, $R^2 = .064$) of Poster Quality. Revision ($t(197) = 2.01$, $p = .02$, $\beta = .18$) was a significant predictor for Posttest score ($F(2,197) = 6.27$, $p = .002$, $R^2 = .05$)

Thus, CF was a predictor of Poster Quality for all studies in both countries, and Revision predicted both Poster Quality and Posttest scores for both Sweden and the US study in USSschool1 and USSchool2.

4.3 Do the relations between learning choices and broader learning outcomes outside the learning environment (i.e., academic achievement, such as grades or standardized tests) vary with country?

We calculated the correlations between CF and Revision with Math and Swedish language grades for the Swedish students and compared these with similar correlations for the US studies. For USSchool1 and USSchool2 results for Standardized English Language Arts and Mathematics achievement tests were used, while both standardized tests (Math-CST and ELA-CST) as well as math grade were used in the USSchool3 study.

Table 7. *Correlation between learning choices and learning outcomes in Math and English/Swedish (grades or standardized test scores)*

	Math	**CF**	**Revision**	**Language**	**CF**	**Revision**
Sweden	Math grade	0.20**	0.24**	Swedish grade	0.16**	0.20**
USSchool1	ISAT	0.33**	0.21	ISTA	0.41**	0,31*
USSchool2	NYSTP Math	0.39**	0.28**	NYSTP ELA	0.33**	0.08
USSchool3	Math grade	0.21	0.30*	ELA-CST	0.23	0.18
USSchool3	Math-CST	0.29*	0.19			

As shown in Table 7, there were significant positive correlations in the Swedish data set between CF as well as Revision and grades in Math as well as Language. These are consistent with the previous results from the US studies, with several significant correlations between learning behaviors in the educational game and outcomes outside the educational game, which are positive and weak to moderate.

5. Discussion, Limitations, and Conclusions

We set out to conduct a cross-cultural replication study by replicating US studies in Sweden using the same method and instrument, hoping to make two different contributions: on the topic of feedback behavior and on the topic of conducting replication studies in another country.

Regarding feedback behavior, we found that a correlation between seeking out critical feedback and revising ones' work after feedback is present in both countries, with negative feedback but not

positive feedback used as an incentive to improve. This confirms previous results showing that praise or positive feedback is not used to revise nor necessarily leads to improved learning.

We also found significant correlations between seeking critical feedback and learning outcomes for all studies, which confirms previous theories that negative feedback overall is more fruitful than positive feedback (Kluger & DeNisi, 1998). This is especially interesting in this study because positive and negative feedback in Posterlet is designed to be equally informative.

Finally, even though the correlation between the amount of critical feedback the students chose, and their academic achievement was modest in our study, seeking critical feedback seems to be a productive behavior in academic settings in both countries.

The question remains whether this pattern of results is generalizable to other countries and cultures that are more different, for example more collectivistic cultures. Even though there are similarities between the US and Sweden, they are both part of WEIRD societies. Thus, it would be useful to expand the replication study to a greater diversity of countries, particularly those that fall outside the WEIRD categorization.

Regarding cross-cultural replication studies, one lesson learned is that replicating studies in different countries can inform us of what results are generalizable, but that it is hard to draw conclusions about differences. For example, we found a significant difference between Sweden and the US regarding the tendency to choose critical feedback, but we also found significant differences between different schools within each country, indicating that the difference between countries in our case likely reflects the sampling for different schools rather than country or culture. Furthermore, socio-economic status of schools is likely a relevant factor to consider for this kind of comparisons (Ewijk & Sleegers, 2010). We did not include this in the analysis, but there is a trend that students overall seek less critical feedback and revise to a lower degree in Schools with low socio-economic status (Table 2).

Also, school profiles and classroom practices – sometimes referred to as 'classroom cultures' – likely play a role (Roll & Wylie, 2016) and need to be taken into account. This is especially important to consider when using a digital tool for data collection that is considered "neutral" but can be integrated in the classroom in different ways. Teachers can be anywhere from absent from the classroom to collaborating and helping students. Ogan et al. (2014) noted that even if students were instructed to work silently and alone by their computer, in some classrooms, students may choose to collaborate informally, which influences how much help they seek in the digital learning environment. Such contextual differences are not captured by the data logs. Thus, if using educational technology for replication studies, the conditions in the classroom must either be carefully replicated or considered during analysis. As Roll and Wylie (2016) point out, it is important that such information also is included in articles.

Overall, when conducting replication studies and comparisons, using an instrument such as Posterlet is a strength in that it is curriculum independent and digitally delivered. Thus, some potential confounders in how a data collection of this kind would be contextualized and performed in different countries, are reduced.

References

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, 63(7)

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5-31.

Blanchard, E. G. (2012, June). On the WEIRD nature of ITS/AIED conferences. *International Conference on Intelligent Tutoring Systems* (pp. 280-285). Springer, Berlin, Heidelberg.

Buljubasic, A. (2018). Formativ bedömning i praktiken – En kvalitativ intervjustudie om lärares användning av formativ bedömning. (eng. Formative assessment in practice – A qualitative interview study about teachers) Master Thesis. Göteborgs Universitet.

Chin, D. B., Blair, K. P., Wolf, R. C., Conlin, L. D., Cutumisu, M., Pfaffman, J., & Schwartz, D. L. (2019). Educating and measuring choice: A test of the transfer of design thinking in problem solving and learning. *Journal of the Learning Sciences*, 28(3), 337-380.

Cutumisu, M., Blair, K. P., Chin, D. B., & Schwartz, D. L. (2015). Posterlet: A game-based assessment of children's choices to seek feedback and to revise. *Journal of Learning Analytics*, 2(1), 49-71.

Cutumisu, M., Blair, K. P., Chin, D. B., & Schwartz, D. L. (2017). Assessing whether students seek constructive criticism: The design of an automated feedback system for a graphic design task. *International Journal of Artificial Intelligence in Education*, 27(3), 419-447.

Cutumisu, M., & Schwartz, D. L. (2021). Feedback choices and their relations to learning are age-invariant starting in middle school: A secondary data analysis. *Computers & Education*, 171, 104215.

Dweck, C. (2012). *Mindset: How You Can Fulfil Your Potential*. New York: Constable & Robinson.

Van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational research review*, 5(2), 134-150.

Fyfe, E. R., & Rittle-Johnson, B. (2016). Feedback both helps and hinders learning: The causal role of prior knowledge. *Journal of Educational Psychology*, 108(1), 82.

Greenfield, P. M., Quiroz, B., & Raeff, C. (2000). Cross-cultural conflict and harmony in the social construction of the child. *New directions for child and adolescent development*, 2000(87), 93-108.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, 466(7302), 29-29.

Hofstede, G., Hofstede, G. J., & Minkov, M. (2010). Cultures and organizations: Software of the mind, third edition (3rd ed.). McGraw-Hill Professional.

Kant, J. "Några invändningar mot Dylan William". Johan Kants blogg (blog), September 28, 2011, https://johankant.wordpress.com/2011/09/28/nagra-invandningar-mot-dylan-wiliam/

Kitayama, S., Markus, H. R., & Matsumoto, H. (1995). Culture, self, and emotion: A cultural perspective on "self-conscious" emotions.

Kluger, A. *N*., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current directions in psychological science*, 7(3), 67-72.

McInerney, D. M. (2008). The motivational roles of cultural differences and cultural identity in self-regulated learning. *Motivation and self-regulated learning: Theory, research, and applications*, 369-400.

Ogan, A., Walker, E., Baker, R., Rodrigo, M., Mercedes, T., Soriano, J. C., & Castro, M. J. (2015). Towards understanding how to assess help-seeking behavior across cultures. *International Journal of Artificial Intelligence in Education*, 25(2), 229-248.

Purdie, *N*., Hattie, J., & Douglas, G. (1996). Student conceptions of learning and their use of self-regulated learning strategies: A cross-cultural comparison. *Journal of educational psychology*, 88(1), 87.

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, 26(2), 582-599.

Suzuki, L. K., Davis, H. M., & Greenfield, P. M. (2008). Self-enhancement and self-effacement in reaction to praise and criticism: The case of multiethnic youth. *Ethos*, 36(1), 78-97.

Tärning, B., Lee, Y. J., Andersson, R., Månsson, K., Gulz, A., & Haake, M. (2020). Assessing the black box of feedback neglect in a digital educational game for elementary school. *Journal of the Learning Sciences*, 29(4-5), 511-549.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Construction of a Role-Play Style Japanese Learning Support System that Allows Teachers to Edit Dialogue Situations

Satoru KOGURE[a*], Yudai KAMATA[b], Yasuhiro NOGUCHI[a], Koichi YAMASHITA[c], Tatsuhiro KONISHI[a], & Makoto KONDO[a]
[a]*Faculty of Informatics, Shizuoka University, Japan*
[b]*Graduate School of Integrated Science and Technology, Shizuoka University, Japan*
[c]*Faculty of Business Administration, Tokoha University, Japan*
*kogure@inf.shizuoka.ac.jp

Abstract: In this study, we propose a Japanese language learning support system that enables role-playing dialogue experiences. The system allows the teacher to specify dialogue situations and learning methods in detail. We have built a prototype system. We presented the system and the method of registering contents to the system, to Japanese language teachers teaching Japanese at universities, and conducted a survey.

Keywords: Japanese learning support system, role-playing, learning contents registration, cooperative learner agent

1. Introduction

According to the "Overview of Japanese-Language Education in Japan," (Japanese Language Division, Agency for Cultural Affairs, 2021), the number of learners of Japanese has basically been increasing every year, although the number of learners has been decreasing in recent years due to the recent world situation. The increase in the number of Japanese language teachers has not kept pace with the increase in the number of learners, so learners must learn Japanese on their own. Conversely, role-play dialogues are considered effective in the early stages of language learning and have been adopted as a teaching method in classes (Ministry of Education, Culture, Sports, Science and Technology of Japan, 2022). However, learning through role-play dialogues requires feedback from both the dialogue partner and the teacher, which is difficult for self-study. There are existing role-play language learning support services that allow self-study, however, they have drawbacks such as only being able to use preset situations (Takeda, M. & Chiba, A., 2019), or being able to set up situations freely however, at a high cost (Sano, et al, 2011), and few systems exist that allow teachers to set up situations at a low cost. To solve these problems, we have built a Japanese language learning support system that follows a general learning flow. In our system, teachers can set the details of how learners learn with the content they have prepared.

2. Position of this Study

The JF Standard for Japanese Language Education is a standard proposed by the Japan Foundation based on the international standard CEFR (Common European Framework of Reference for Languages). The Japan Foundation's long-term training programs for overseas Japanese-language teachers have adopted teaching methods based on the JF Standards for Japanese-Language Education since 2015.The JF Standards for Japanese-Language Education uses the term "Can do" to express the ability to perform tasks, "What can you do in Japanese and how well can you do it?" Shinohara et al. (Shinohara, A., & Tsukishima, F., 2015) report on JMOOC (Japan Massive Open Online Courses), an e-book type e-Learning material developed by the Japan Foundation Japanese-Language Institute.

Role-playing interactive learning is a well-known interactive learning method in which the learner simulates a situation by playing a role assigned to him or her according to a situation set by the

teacher (Iman, N., J. & Angraini, N, 2022; Zalavra, E. & Makri, K., 2022). It is a form of learning often employed in actual educational settings, and there are many learning warfare services based on role-play type interactive experiences and learning support systems. In this study, we construct a learning support system that enables role-play type interactive experiences.

In the field of learning, cooperative learning, in which learners work together by cooperating with learners of similar abilities to accomplish a task, is often employed (Alamdari, F., E. & Ghani, F., 2021). In the aforementioned role-playing dialogue experience, multiple learners play different roles so that all the learners can simulate the situation together. However, as mentioned above, since there are not enough Japanese language teachers available, self-study is necessary. There are learning support systems that enable self-taught cooperative learning by substituting agents for cooperative learners, who are the partners of cooperative learning. In this study, we construct a learning support system that enables self-taught cooperative learning.

Many of the various language learning support services and learning support systems that exist in the world allow learners to freely select and learn from well-developed content. However, from a teacher's viewpoint, there are few systems that allow teachers to maintain the contents of learning support systems or to specify learning methods in detail. In this study, we construct a learning support system that allows learners to systematically use the basic learning of "Introduction," "Basic Practice," and "Application Practice," which are specified in the standard method of learning in language learning proposed by The Japan Foundation (2007). "Application Practice" includes a dialogue experience with the role-playing strategy.

3. Construction of a Japanese Language Learning Support System that Enables Role-play Style Dialogue Experiences

In order to master a certain learning item, it is effective to proceed in the order of S1, S2, and S3 in Figure 1. We also define the learning methods (from (A) to (F) in Figure 1. Learning method (F) can be divided into the following three categories: (F1) Initial Setup for Role-playing Dialogue Experience (included in S3), (F2) Learners Experience Situational Dialogue with CLA (included in S3), and, (F3) Feedback (included in S3). Figure 1 illustrates the relationship between the teacher, learner, and system in the learning method proposed here. We built a prototype of this system in Python. We also prepared three sample lessons.

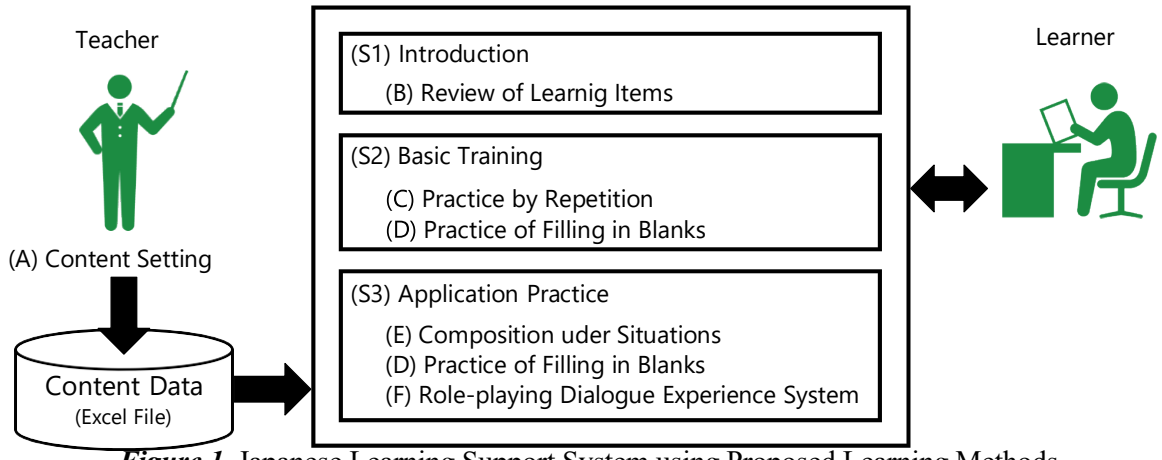


Figure 1. Japanese Learning Support System using Proposed Learning Methods.

4. Questionnaire Results for Japanese Language Teachers

To confirm the usefulness of the Japanese language learning support system developed in this study, we conducted a survey of six Japanese language teachers who teach or have taught Japanese at universities. The survey was conducted from December 20, 2021 to January 10, 2022. Considering the social situation, the questionnaire was administered remotely. We filmed the use of the system. Subjects

viewed a video recording of one of the authors using the system. After viewing the video, we asked the subjects to answer a questionnaire. The questionnaire items were the usefulness of the system in its entirety and each of the functions from (F1) to (F3) for beginners of Japanese. Subjects answered each item on a 5-point scale (1 to 5 points). For the overall system rating, "Overall, do you think this system is useful for the students in the classes you teach?" we received an average score of 2.83. This result suggests that the current system needs improvement. The following comments were received from the participants. “It is difficult to understand the situation because there is no visual presentation at all.” “I don't think it is suitable for in-class use, but it could be used for individual self-study.” “I think it could be used as self-study material for intermediate level students of B1 and above.” Next, we asked the subjects to answer the question "Do you think it would be useful for the students you teach?" for the role-played interaction portion only. The results were 3.67 for (F1), 3.17 for (F2), and 4.00 for (F3). The average score for the role-play type dialogue experience was above 3. Owing to social conditions, we were unable to have international students actually use the system this time. We would like to improve the system based on the opinions obtained from the teachers, have the improved system actually used by international students, and evaluate whether the system motivates them to learn or not.

5. Conclusion

In this study, we propose a Japanese language learning support system that allows teachers to freely practice situations and dialogue methods. In addition to basic learning such as repetitive exercises and fill-in-the-blank questions, the system enables applied learning through role-play type dialogue experiences. We actually built a prototype system. We asked Japanese language teachers to watch the system in action and provide us their opinions. The subjective evaluation of the usefulness of the system was not positive, however, we received feedback on how to improve the system. We are planning to improve the system based on the feedback we received, and to conduct an evaluation experiment with foreign students who were not able to participate in the experiment.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17K00483, 21K12172.

References

Alamdari, F., E. & Ghani, F. (2021). Enhancing foreign language motivation through the magic of cooperative learning: Dream or reality?, Foreign Language Annals, ACTFL, 55(1), 237-257.

Iman, N., J. & Angraini, N. (2022). Roleplaying in EFL classroom: EFL learner’ oral proficiency achievement and perception, English Journal, *16(1)*, 1-15.

Japanese Language Institute of the Japanese Foundation (2019). IRODORI Japanese Online Course, https://www.irodori-online.jpf.go.jp/ (referred on 21st, December 2021).

Japanese Language Division, Agency for Cultural Affairs, (2021). Overview of Japanese language education in Japan in 2020, https://www.bunka.go.jp/tokei_hakusho_shuppan/tokeichosa/nihongokyoiku_jittai/r02/pdf/93285801_01.pdf (referred on 21st, December 2021).

Ministry of Education, Culture, Sports, Science and Technology of Japan (2022). Section 3 How to teach, https://www.mext.go.jp/b_menu/shingi/chousa/shotou/024/report/attach/1370772.htm, referred on 6th, May 2022 (in Japanese).

Shinohara, A, & Tsukishima, F. (2015). The Online Japanese Course "NIHONGO Starter" : The Development and Usage of eBook Materials, International Japanese Language Education Program, The Japan Foundation, *13*, 71-85 (in Japanese).

The Japan Foundation (2007). The Japan Foundation Japanese Language Teaching Method Series, Volume 9: Teaching Beginner's Level, Hitsuji Syobo.

Zalavra, E. & Makri, K. (2022). Relocating online a technology-enhanced microteaching practice in teacher education: challenges and implications, The Electronic Journal of e-Learning, 20(3), 270-283.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Efficacy of an Online Course to Build Mindfulness in Adolescents

Renuka RAUTELA, Anurag DEEP, Mayank SHARMA & Nandini CHATTERJEE*
UNESCO Mahatma Gandhi Institute of Education for Peace and Sustainable Development, 35, Ferozshah Road, New Delhi – 110 001, India
*n.chatterjee@unesco.org

Abstract: Mindfulness intentionally focuses and sustains attention on immediate experience and adopts a curious, open, and accepting orientation towards that experience (Kabat-Zinn, 2005). Increasing research shows that students equipped with mindfulness can demonstrate higher levels of attention, awareness, obtain higher test scores, and report lower stress levels. This paper explores digital activities to build mindfulness in an online course on global citizenship administered to adolescents between 12-14 years. A total of 68 students from India took the course over twelve weeks. A pre-post paired t-test revealed a significant increase in the overall score of mindfulness ($p < 0.05$) as measured by the Five-Facet Mindfulness Questionnaire (FFMQ). Additionally, significant increases were seen on the subscales of "observing" ($p < 0.05$), "describing" ($p < 0.05$) and "non-judging" ($p < 0.05$) facets of mindfulness. Thematic analysis of the open-ended responses supported the findings from the pre-post.

Keywords: Mindfulness, Digital course, Global Citizenship, Social and Emotional Learning (SEL), Efficacy

1. Introduction

The adolescent years are periods of development where children face a myriad of physiological and psychological changes leading to an upsurge in their stress levels (OECD, 2017). According to the National Institute of Health, 1 in 8 children suffer from stress and anxiety disorders (Kessler R. C., 2005), and more than 25% of 13- to 18-year-olds will experience it by the time they turn 18. This data is alarming, especially when a growing body of literature informs us of the negative impacts of prolonged stress. Accumulating Social and Emotional Learning (SEL) research has shown that mindfulness training decreases stress, anxiety, and disruptive behaviors and even reduces depression (Hofmann, S. G., 2010). Mindfulness is intentionally focusing and sustaining attention on immediate experience, adopting a curious, open, and accepting orientation towards that experience (Kabat-Zinn, 2005). In addition, students with mindfulness training score higher grades (Franco, C., 2010), are more attentive (Weare, K., 2012) than their counterparts, and are less likely to engage in disruptive behaviour. These students have also demonstrated higher self-esteem and social skills that help them build deep and meaningful relationships (Durlak, J. A. et al., 2011).

These powerful effects of mindfulness have led to several studies exploring mindfulness-based intervention in the classroom. However, with worldwide school closures and online, exploring ways to teach mindfulness through digital modalities also becomes imperative. Increased research shows that using multiple digital pedagogies results in fun, playful, motivational, and rewarding learning experiences. We sought to harness this potential of technology and utilized the Libre pedagogical framework (Rautela et al., 2019) to cultivate different facets of mindfulness. Using this learning, activities on digital storytelling, dialogue, journaling, and games focused on Mindfulness were embedded in a course on global citizenship. The course is hosted on Framerspace (www.framerspace.com/course/global-citizenship), a free AI-based learning platform.

2. Efficacy of Digital Activities to Cultivate Mindfulness

To evaluate the efficacy of digital activities to cultivate mindfulness, a research study was conducted using a pre-post research design. It was hypothesized that adolescent students who undertook the course would demonstrate a higher understanding of self-awareness and mindfulness and its role in calming the brain.

2.1 Participants, Study Procedure, Data sources and Analysis Technique

Participants for the study were recruited from schools in India and were selected based on convenience and ease of course implementation. A total of 68 participants took part in the study. The mean age of the participants was 12.9 years (SD = 0.72 years and comprised of 40 females and 28 males). The research was conducted over five weeks in 4 phases – An induction workshop and a Pre-Assessment questionnaire, followed by participants undertaking the course and completing the post-Assessment questionnaire. The self-report Five-Facet Mindfulness Questionnaire (FFMQ) (Baer et al. 2006) was used to measure the course's efficacy. FFMQ measures mindfulness as a five-factor construct comprising describing, observing, non-judging, acting with awareness, and non-reactivity to inner experiences. Participants responded to these items on a 3-point Likert scale (Never, Sometimes, Always). Paired t-tests on students' responses to the FFMQ Questionnaire was performed using R version 4.0.2. Non-reactivity was not measured in this research study, as there was no activity in the course associated with it. Additionally, a deductive thematic analysis of two open-ended questions was conducted to gauge participant's understanding of mindfulness and their ability to identify mindfulness-based activities. Two independent coders carried out the coding process with one student response as the unit of analysis.

2.2 Results

3

2.2.1 Quantitative Analysis

After the intervention, an increase in the score on overall mindfulness was seen ($p < 0.05$). Additionally, increases were seen on the subscales of "describing" ($p < 0.05$), "observing" ($p < 0.05$), and "non-judging" ($p < 0.05$) (Figure 1). Besides this, no statistically significant difference was found in the construct of acting with awareness.

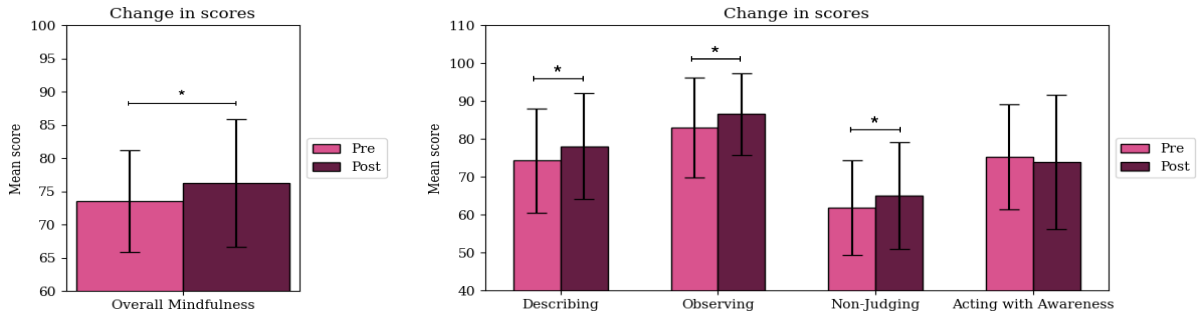


Figure 1. FFMQ Questionnaire Mean Score Variation Across Pre and Post Tests

2.2.2 Thematic Analysis

Data analysis of the response to questions on evaluating participants' understanding of mindfulness revealed that participants could identify instances when they were not mindful and the techniques, they could use to enhance calmness (Table 1).

Table 1: *Themes and Respective Sample Excerpts*

I am going to introduce you to a new way of life, 'Mindfulness'. This will help you to be attentive and aware of the moment and also help you to fully give yourself to whatever you do…..Tell me something about the times that you are not paying attention or are distracted…….		
S. No.	**Themes**	**An instance of response from participants artefacts**
1	**Place** (Where I am not mindful?)	I always get distracted during class. I can sit in one place and stare at a wall and daydream for an entire period straight. it happens at home too. my parents might be talking to me but I will just nod my head as if I heard everything but in reality, I was just looking at the floor. and I'm being 100 % honest. I go to auto-pilot mode A LOT.
2	**Instance** (When I am not mindful?)	
3	**Cognitive processes** (What do I do when I am not mindful?)	
How would you help Vivaani be calm? Would you have done something similar or different?		
S. No.	**Themes**	**An instance of response from participants artefacts**
1	Perspective-taking/Empathy	I would have told her to focus on what she is going to do and not to focus on the negative aspects....and to meditate for a while to calm down her mind.
2	Motivating/Positive intention	
3	Becoming calming/Focus	

3. Discussion and Conclusion

The research study reported in this paper explores the efficacy of a digital course in building mindfulness in adolescents. The quantitative analysis of the FFMQ scores revealed a significant increase in the overall score of mindfulness ($p < 0.05$) suggesting the course's efficacy in cultivating mindfulness. Additionally, significant increase was seen in three subscales of describing, observing, and non-judging facets of mindfulness. This significant difference can be attributed to the digital activities that were used in the course. The thematic analysis of open-ended responses supports the quantitative findings. The findings from this research study present promising opportunities for spurring the development of mindfulness competency among students through online interactive courses. An important limitation of this research is that the results presented in this study must be corroborated with results obtained from other behavioral/neuroscientific measures.

References

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Five Facet Mindfulness Questionnaire. *Assessment*, *13*(1), 27-45.

Durlak, J. A., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development*, *82*(1), 405-432.

Franco, C., Mañas, I., Cangas, A.J., and Gallego, J. *The Applications of Mindfulness with Students of Secondary School: Results on the Academic Performance, Self-concept and Anxiety*. University of Almería Department of Education, 2010, pp.83-97.[11])

Hofmann, S. G., Sawyer, A. T., Witt, A. A., & Oh, D. (2010). The effect of mindfulness-based therapy on anxiety and depression: A meta-analytic review. *Journal of consulting and clinical psychology*, *78*(2), 169.

Kabat-Zinn, J. (2005). *Coming to our senses: Healing ourselves and the world through mindfulness*. Hachette UK.

Kessler RC, Chiu WT, Demler O, Merikangas KR, Walters EE. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Arch Gen Psychiatry.* 2005 Jun;62(6):617-27.

OECD. (2017). *PISA 2015 Results (Volume III)*. Paris, France

Rautela, R., & Singh, N. C. (2019). Libre—Nourish the Brain So the Future Can Flourish. *Childhood Education*, *95*(5), 34-43.

Weare, K. *Evidence for the Impact of Mindfulness on Children and Young People*. Mindfulness in Schools Project, April 2012, pp. 1-12.[13].

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Detection of At-Risk Students in Programming Courses

Ikkei IGAWA[a*], Yuta TANIGUCHI[a] , Tsubasa MINEMATSU[a] , Fumiya OKUBO[a] & Atsushi SHIMADA[a]
[a] *Kyushu University, Japan*
*igawa@limu.ait.kyushu-u.ac.jp

Abstract: Since the demand for programmers is increasing, programming courses are being offered widely. In this context, students' motivation can be damaged by difficulties they encounter in their programming courses. Although teachers' support is necessary to prevent such an issue, it is impossible for teachers to directly monitor all students' programming activities at the same time and determine which students have troubles with programming. Therefore, several studies have been conducted to help teachers monitor students. However, these studies do not provide an understanding of the activities of students who do not run their code, which may lead researchers to miss students who are in trouble. In this paper, we propose an indicator for detecting students who need coding support by analyzing programming logs that are recorded even when the students do not run their code. This gives teachers deeper insight into the students' programming performance. Although further work remains, the validation of this indicator shows that it could detect those students who are in trouble.

Keywords: Programming education, educational data analysis, clustering, teacher support

1. Introduction

Recently, the demand for programmers has been rapidly increasing, and the lack of programmers has become serious. Therefore, developing programmers is an urgent need, based on which various changes are being implemented in the field of education. For example, programming education became mandatory in Japanese elementary schools in 2020.

Jenkins (2001) and Pilkington (2018) have reported that students' motivation to learn is important in order for them to succeed in programming courses. However, their motivation is often affected by the difficulties they encounter in coding, which may lead students to drop out (Martins et al., 2010; Feldgen et al., 2004). To prevent this, teachers must monitor students' programming activities and determine which students have troubles with programming (hereafter, at-risk students). However, the larger the number of students, the more difficult it is to monitor all of them at the same time.

To help the teachers monitor students, various studies have been conducted (e.g., Fonseca et al., 2018; Fonseca et al., 2021; Penmetsa et al., 2021). Many of them include the creation of a dashboard. Fu et al. (2017) developed a dashboard that enables teachers to identify the patterns of errors students frequently make by obtaining error messages from the compiler. Furthermore, Diana et al. (2017) analyzed programming logs to predict the students' final grade in a programming course. In those works, however, the researchers only used students' programming logs that were recorded when they ran their code, which means that they could not collect programming logs for analysis from students who wrote their code without execution. This leads to some problems in monitoring students. One such problem is the lack of logs to analyze. For students who seldom run their codes, researchers cannot collect a sufficient number of logs. This leads us to miss at-risk students due to the reduced accuracy of detection. If we can obtain programming logs while the students are coding regardless of execution, we can increase our sample size and deepen our understanding of students' programming activities.

In this paper, we design a better indicator to detect at-risk students by calculating the frequencies of execution errors and autosaves. Autosave is an event that automatically records the students' programming logs such as time and code even if they do not execute their code. This enables us to get a sufficient number of programming logs and identify students' programming activity while they are coding without execution. In addition, we validate how effective the frequencies of the execution errors and autosaves are at finding at-risk students.

2. Programming Log Collecting Tool

The WEVL is an online coding application currently being developed in our laboratory as well as being used in some programming courses in our university (Taniguchi et al., 2022). Figure 1 shows the user interface of WEVL. If a student writes a code in the text box on the left side and clicks the yellow execution button, the execution result of the code is displayed on the right side. While writing a code on WEVL, various pieces of information such as name, time, and code are automatically recorded to the WEVL database. This recording event occurs twice. The first time is when students run the code. In this case, the result of the execution is also saved. The second is when the student stops typing for three seconds. However, if the code is the same as the one that was saved previously, this event does not occur.



Figure 1. The User Interface of WEVL.

By automatically recording the programming logs, we can obtain a sufficient number of logs for analysis. For example, in the case of a certain 90-minute programming course in our university consisting of 87 students, the database includes 1,069 autosave logs and 245 execution logs.

If we line up the programming logs in chronological order, we can obtain one time series set for each student. In this paper, we call these series' elements "save-points" and use them to detect at-risk students.

3. Detection of At-Risk Students

3.1 Indicator for Detection

We hypothesized that there are some situations that lead a student to be in an at-risk situation; one example situation is repeatedly getting errors when they run their code. In this case, the students could be in an at-risk situation. Another situation is when students get stuck while writing their code. For example, when a student is unsuccessfully trying to solve a task in a programming course, s/he may frequently stop writing code. In other words, s/he would be repeating the trial-and-error process.

Next, we detected students in such situations based on their activity. Students who are repeatedly getting errors when running their code can be detected simply by checking the frequency of execution errors in their save-points. However, it is more difficult to detect students who are repeating trial-and-error processes. As mentioned in section 2, autosave happens when students stop typing a code for a short time. Therefore, we can say that autosave is a good candidate for the detection of trial and error.

We then designed an indicator to detect at-risk students based on the idea that the frequencies of certain kinds of save-points have fruitful information.

3.2 Designing the Indicator

We designed an indicator for detecting at-risk students by using students' save-points. As mentioned in section 2, there are three kinds of save-points when students are coding on WEVL: run and success, run and fail, and autosave. Figure 2 shows example save-points of a certain student.

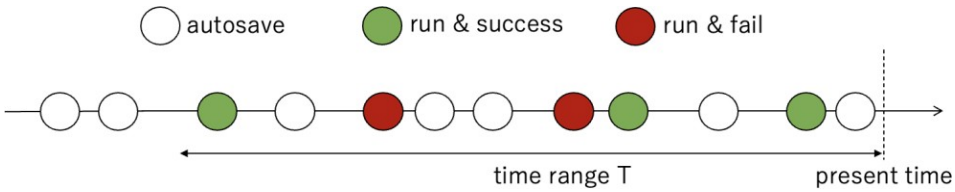


Figure 2. An Example of a Student's Save-points.

In Figure 2, the white points represent autosaves, the green points represent run and success, and the red points represent run and fail. As noted in section 3.1, at-risk students can be found by checking the frequencies of failures and autosaves. When calculating these frequencies, we only take account of recent save-points. As shown in Figure 2, we set the time range T, which covers the past for a certain period to the present time. We calculated the frequencies of the number of save-points in this time range. When we set time range T to T minutes at the present time t, we defined the ratio of errors from time t-T to time t as "failure ratio" and the ratio of autosaves as "autosave ratio." In the case of Figure 2, the number of the save-points in the time range T is 10. There are five autosaves and two failures in the time range T, so the autosave ratio is 50% and the failure ratio is 20%. In this research, we use these ratios as an indicator for detecting at-risk students and validate whether they can detect such students via a classification experiment.

3.3 Identifying Coding Situations

In this section, we identify students' coding situations based on the ratio of failures and autosaves. Table 1 shows four possible groups classified by the size of the students' failure and autosave ratio.

Table 1. *Four Groups Classified by Failure and Autosave Ratio*

		autosave ratio	
		high	**low**
failure ratio	**high**	Group A	Group B
	low	Group C	Group D

Group A consists of students who have a high failure and autosave ratio. As mentioned before, a high autosave ratio represents a high probability of repeating trial and error. Therefore, students in this group can be identified as those who are trying to run their code but not finding success. Thus, they can be at-risk students.

Group B is a group of students who have a high failure ratio and a low autosave ratio. A low autosave ratio represents that the student can smoothly write a code. Given that such students have a high failure ratio, this group indicates students who are making several errors occur but are writing code to successfully solve them.

In group C, students have a high autosave ratio and a low failure ratio. In other words, they are repeating trial-and-error without making frequent errors. In general, students with a low failure ratio are less likely to be at-risk. However, we hypothesized that high frequency of trial-and-error leads a student to be in an at-risk situation in section 3.1. Therefore, we cannot easily say that they are low-risk students.

Finally, group D includes students with a low failure and autosave ratio. Naturally, this ratio represents a high frequency of success. Therefore, students in this group can be identified as those who run their code several times, typically successfully, which is a low-risk situation.

To sum up, group A includes at-risk students while group B and D does not. In group C, it seems difficult to say whether the students are at risk. For this reason, we conducted a classification experiment and validated the indicator.

4. Validation of Indicator

4.1 Dataset

In this research, we conducted an experiment using the programming logs collected from the WEVL database. These logs were recorded in the database during a Python programming exercise course offered in our university in 2020, and we used the logs for three sessions of the course.

In Figure 2, we calculated the failure and autosave ratio from save-points in the time range T. If this time range is long enough, a sufficient number of save-points can be collected. If it is too long, however, both the failure and autosave ratios are affected by past data, which may lead to missing students that are currently having troubles with programming. In this experiment, we set the time range T to 30 minutes. Whether this length of the range is appropriate may require discussion. The course was offered from 8:40 am to 10:10 am, and the latter part of the class was used for exercises. Thus, we collected students' programming logs that were recorded from 9:30 am to10:00 am, obtaining 165 pieces of time series data of save-point logs. These save-point logs comprise 6,647 programming logs (1,255 execution logs and 5,392 autosave logs).

The purpose of this research is to find at-risk students. However, we cannot know whether a given student had trouble in a class period based solely on the logs in the database. Therefore, we used the assignment's scores to evaluate the classified students. This assignment was imposed every week in the programming course, and a full score is 5. We conducted an experiment based on the idea that this score is related to the performance in the course.

4.2 Setting the Thresholds

In section 3.3, we created a table divided by whether the ratio was high or low. To do so, a threshold is required for both the failure and autosave ratios.

First, we set the thresholds of failure and autosave ratio at 50%. We then calculated the average score of the assignments for each group. Finally, we maximized the differences between groups A and C, which consist of students with a high autosave ratio, and groups B and D, which consist of students with a low autosave ratio.

Figure 3 shows how the differences of the average score changes according to the thresholds. The vertical axis represents the threshold of the autosave ratio, and the horizontal axis shows the threshold of the failure ratio; the gradation represents the magnitude of the difference. The greater the difference, the brighter the color of the point. According to this figure, the ratios that maximize the differences of the scores are 20% for the autosave ratio and 70% for the failure ratio. Thus, we set these as the thresholds.

4.3 Results of the Experiment

After setting the thresholds, we classified the students into four groups. Table 2 shows the statistics of each group's assignment score; 5 represents the highest score. In this section, we verify how the failure and autosave ratio are effective to detect at-risk students by examining the relationship between the identified situation and the weekly assignment score.

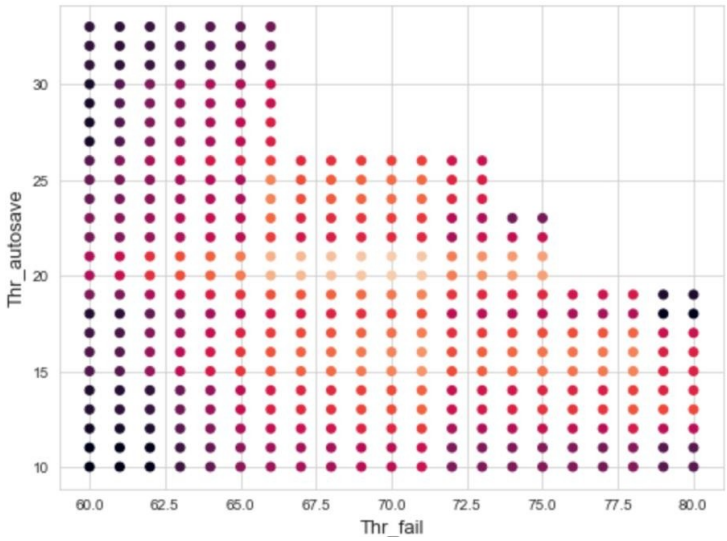


Figure 3. The Differences of the Average Assignment Score.

Table 2. *The Statistics of Each Group's Assignment Score*

	Group A	**Group B**	**Group C**	**Group D**	**all**
number of data	11	25	114	15	165
average	3.38	3.88	3.92	4.38	3.92
std.	0.83	0.92	1.13	1.10	1.09

Students in group A have high failure and autosave ratios, which suggests that they try to run their code successfully but it does not work well. Table 2 shows that this group's average score is the lowest of the four groups' scores. Therefore, we can say that they are likely to be at-risk students.

In group B, students have a high failure ratio and a low autosave ratio. They are identified as getting several coding errors, but being able to solve them smoothly. Although they have as high a failure ratio as group A, Table 2 shows that they received scores that were not so low compared to the average across the groups.

It seemed difficult to judge whether the students in group C are successful in coding. Table 2 shows that students in this group scored almost the same as the average of all the students, and about 70% of the data points were classified into this group.

Finally, students in group D exhibited low failure and autosave ratios, suggesting that they are coding successfully without getting errors. This group's average score is the highest of the four. Therefore, we can say that they are not likely to be at risk.

4.4 Discussion

Students frequently getting coding errors seem to be at risk in general. However, our results shows that a high error frequency does not always lead to a poor grade. If we classify the students only by failure ratio, groups A and B are classified together. That group's average assignment score is 3.73 out of 5, which does not differ much from the overall average score. Therefore, we can say that failure ratio is not sufficient to classify students as having a poor grade, and we can extract at-risk students from those who have a high failure ratio by using the autosave ratio. Thus, the autosave ratio is an important value for detecting at-risk students. The same applies to groups C and D, as those groups' joint average assignment score is 3.97. From the above, we can also say that the autosave ratio is useful for detecting low-risk students.

The results for Group C show that they received a similar assignment score as the overall average, and many more students were classified into Group C than any other group; both successful and unsuccessful students were included in this group. Thus, more details should be used to classify this group.

5. Conclusion

In this paper, we proposed an indicator to detect at-risk students in programming courses to help teachers comprehend students' programming activities and detect those who are at risk. By making good use of WEVL, which can record students' codes even if they are not executed, we can analyze many more programming logs and capture students' activities while they are coding. We then designed an indicator calculating the frequencies of execution errors and autosaves. We classified students into four groups and validated how important the ratios are to detect at-risk students. The results showed that in addition to failure ratio, autosave ratio, or the amount of trial and error in coding, is valuable for detection. Group A, which had high failure and autosave ratios, are at-risk students. However, about 70% of all students were classified into group C, and the score distribution is similar to the overall average. Thus, this group should be classified in more detail. In this research, we used only two parameters, autosave ratio and failure ratio, to classify students. Therefore, we are conducting further experiments using other parameters such as the length of the code. We are conducting further research to improve detection accuracy and hope that it will be of great help for teachers in the field of programming education.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP21K17863 and JP22H00551, Japan.

References

Fu, X., Shimada, A., Ogata, H., Taniguchi, Y., & Suehiro, D. (2017). Real-time learning analytics for C Programming Language courses. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. doi:10.1145/3027385.3027407

Diana, N., Eagle, M., Stamper, J., Grover, S., Bienkowski, M., & Basu, S. (2017). An instructor dashboard for real-time analytics in interactive programming assignments. *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. https://doi.org/10.1145/3027385.3027441

Jenkins, T. (2001). The motivation of students of programming. *ACM SIGCSE Bulletin*, *33*(3), 53–56. https://doi.org/10.1145/507758.377472

Taniguchi, Y., Minematsu, T., Okubo, F., & Shimada, A. (2022). Visualizing Source-Code Evolution for Understanding Class-Wide Programming Processes. *Sustainability 2022*, *14*, 8084. https://doi.org/10.3390/su14138084

Fonseca, N. G., Macedo, L., Marcelino, M. J. & Mendes, A. J., (2018). Augmenting the teacher's perspective on programming student's performance via permanent monitoring, *IEEE Frontiers in Education Conference (FIE)*, 2018, pp. 1-9, doi: 10.1109/FIE.2018.8658924.

Penmetsa, P., Shi, Y., & Price, T. (2021). Investigate Effectiveness of Code Features in Knowledge Tracing Task on Novice Programming Course. *Work-In-Progress Track, CSEDM Workshop @ EDM'21*

Fonseca, N. G., Macedo, L. & Mendes, A. J., (2021). The importance of using the CodeInsights monitoring tool to support teaching programming in the context of a pandemic, *IEEE Frontiers in Education Conference (FIE)*, pp. 1-8, doi: 10.1109/FIE49875.2021.9637292.

Pilkington, C. (2018). A Playful Approach to Fostering Motivation in a Distance Education Computer Programming Course: Behaviour Change and Student Perceptions. *The International Review of Research in Open and Distributed Learning, 19*(3), –. doi:10.19173/irrodl.v19i3.3664

Martins, S. W., Mendes, A. J. & Figueiredo, A. D., (2010). A strategy to improve student's motivation levels in programming courses, *IEEE Frontiers in Education Conference (FIE)*, pp. F4F-1-F4F-7, doi: 10.1109/FIE.2010.5673366.

Feldgen, M. & Clua, O., (2004). Games as a motivation for freshman students learn programming, *34th Annual Frontiers in Education,* pp. S1H/11-S1H/16 Vol. 3, doi: 10.1109/FIE.2004.1408712.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Learning Support System Visualizing Relationships Among Classes and Objects Based on Teacher's Intent of Instruction

Koichi YAMASHITA[a*], Yusuke SUZUKI[b], Satoru KOGURE[b], Yasuhiro NOGUCHI[b], Raiya YAMAMOTO[c], Tatsuhiro KONISHI[b] & Yukihiro ITOH[d]
[a]*Faculty of Business Administration, Tokoha University, Japan*
[b]*Faculty of Informatics, Shizuoka University, Japan*
[c]*Faculty of Engineering, Sanyo-Onoda City University, Japan*
[d]*Shizuoka University, Japan*
*yamasita@hm.tokoha-u.ac.jp

Abstract: While several learning support systems that visualize program behaviors for various object-oriented languages have been developed to date, many of them are not sufficient for learners to understand the concepts specific to object-oriented languages. We have developed a program visualization system that supports learners' understanding by visualizing relationships among objects and classes. We evaluated the effectiveness of our system by introducing it into an actual class and measuring the degree of improvement in learners' understanding based on tests. The evaluation results suggest that our system would have a certain degree of effectiveness for learning programming using object-oriented languages.

Keywords: Programming education, program visualization system, object-oriented conceptual model, educational authoring tool

1. Introduction

To date, several program visualization (PV) systems have been developed to support learners' understanding of programs (Sorva, Karavirta, & Malmi, 2013). We have developed a PV system for C programs called Teacher's Explaining Design Visualization Tool (TEDViT), and have conducted several classroom practices using this system (Kogure et al., 2014). On the other hand, Java, rather than C, has been increasingly adopted as a language used in introductory programming education for novice students in recent years. Therefore, this study attempts to extend TEDViT to support Java and to develop a PV system that can be adapted to a variety of programming learning environments for novice learners.

As an approach to make TEDViT capable of object-oriented languages, we have developed a visualization model called Object-Oriented Conceptual Model (OOCM) and have evaluated effectiveness of PVs based on the model for understanding class and object concepts based on questionnaire surveys (Kogure et al., 2019). However, the PV system based on OOCM had implementation problems and required a non-trivial amount of manual coding works to generate PVs. The purpose of this study is to develop a system visualizing PVs based on OOCM with no manual interventions and to measure their effectiveness by introducing the system into an actual class.

2. PV System Visualizing OOCM Based on Teacher's Intent of Instruction

Object-oriented languages such as Java are increasingly being used in programming education for novice learners. Based on this background, we adapted Object-Oriented Conceptual Model (OOCM) (Kogure et al., 2019) as our approach to make TEDViT support Java. OOCM is a visualization model based on Unified Modeling Language (UML) and can visualize the relationships among objects and classes, including the concepts of inheritance and polymorphism. In this paper, we refer to the PV system developed by Kogure et al. as Kogure system.

Kogure system is a PV system based on TEDViT which visualizes program behavior based on the execution history. Execution history is a log file generated by executing the target program in which the system embedded statements for recording execution statuses of the program. However, due to implementation problems, Kogure system requires manual modification of execution history to visualize OOCM-based PV of Java programs. Therefore, in this study, we developed a new PV system by radically modifying Kogure system to eliminate the need for manual intervention. In our system, the execution history is generated from Java Debug Interface (JDI) framework, so there is no need to add statements to observe the execution status.

Figure 1 shows a screenshot of our system. Our system generates a learning environment consisting of three main visualization fields. Field (A) visualizes the target program code and highlights the statement being executed. Learners can perform stepwise execution of the target program by clicking “Next” or “Prev” button. The source code of each class can also be visualized by changing the target program file with pull-down menu. The data structure to be processed by the method being executed is visualized in field (B). Field (C) visualizes the classes and objects that are interrelated in the target program based on OOCM. In fields (B) and (C), teachers can define the visualization by providing drawing rules to specify when, what, and how variables, classes, and objects are visualized.

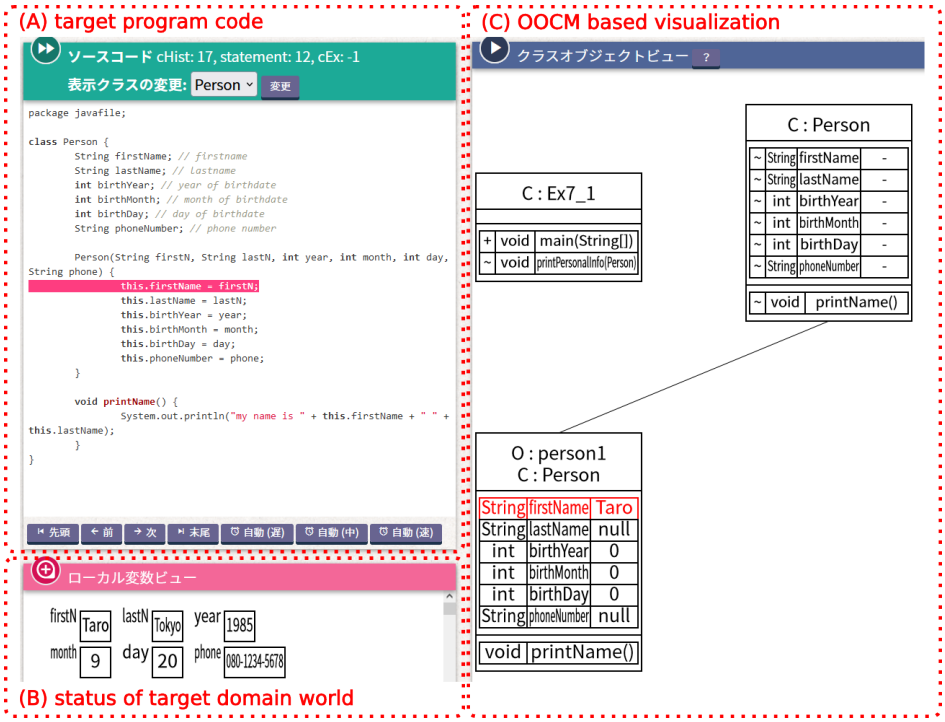


Figure 1. Learning environment generated by our system.

3. Evaluation

In order to evaluate the effectiveness of our system, we conducted a classroom practice using our system, incorporating it into an actual introductory programming course. The hypothesis to be verified is that, compared to the usual classroom exercise consisting of code reading and program execution of sample program, observing the behavior of sample program using our system would cultivate better understanding of the concepts specific to object-oriented programming. The actual class in which our system was introduced was an on-demand class offered to first-year university students majoring in computer science. 68 students with less than one year of programming experience participated in this practical class.

The practiced class includes two exercises: Exercise A, which was usual-style exercise as in regular classes, and Exercise B, which used our system. Both exercises were based on the same sample program, but Exercise A required only code reading and program execution, while Exercise B required only learning in the learning environment of our system. We divided the participants into two groups,

Group 1 and Group 2, and reversed the order in which Exercise A and Exercise B were performed. We tested the participants' understanding three times: before the class (pre-test), after the first exercise (middle-test), and after the second exercise (post-test). Because the classes were on-demand, we did not perform grade equalization based on the pre-test score, and grouping into two groups was done randomly. Each test consisted of 20 questions: 3 questions on the properties of concepts specific to object-oriented languages, 6 questions on the behavior of class and object variables, 5 questions on the behavior of programs with inheritance, and 6 questions on the behavior of programs with overriding. The score was calculated with 100 as the maximum score.

Table 1 shows the results of the tests for each of the two groups based on the average of the scores. The underlined scores in Table 1 show the scores immediately after Exercise B where participants use our system. We can see that both groups improved their test scores by learning with our system, suggesting that our system has a certain effect on learning concepts specific to object-oriented languages. On the other hand, especially in Group 1, the improvement in test scores can be seen even when learning without using our system. Hence, the results do not positively support the hypothesis that our system is more effective than the usual-style exercises. However, there are some positive indications for the use of our system, such as a higher improvement in scores in Group 2 compared to the usual-style exercises and a significant improvement in the lowest test scores after the exercises using our system. These results suggest that our system is effective for learning programming in object-oriented languages such as Java.

Table 1. *Average Scores of Pre-, Middle- and Post-test*

Group	N.	1st exercise	Pre-test	Middle-test	Post-test
Group 1	29	No system	75.3	80.0	83.2
Group 2	39	With system	80.7	83.6	84.1

4. Conclusion

In this paper, we describe a PV system which visualizes program behaviors based on teachers' intents of instructions, employing OOCM as an approach to visualize concepts specific to object-oriented languages. We evaluated the effectiveness of our PV system by introducing it in an actual class. The evaluation results did not positively support our hypothesis that observing the behavior of the sample programs using our system would cultivate better understandings of object-oriented concepts, compared to the usual classroom exercises consisting of code readings and program executions. However, some positive indications were obtained from the distribution of scores, suggesting that our system would have a certain degree of effectiveness on learning programming with object-oriented languages such as Java.

Acknowledgements

This study was supported by JSPS KAKENHI Grant Numbers JP18K11567, JP19K12259, and JP22K12290.

References

Kogure, S., Fujioka, R., Noguchi, Y., Yamashita, K., Konishi, T., & Itoh, Y. (2014). Code reading environment according to visualizing both variable's memory image and target world's status. *Proceeding of the 22nd International Conference on Computers in Education (ICCE2014)*, 343-348.

Kogure, S., Ogasawara, K., Yamashita, K., Noguchi, Y., Konishi, T., & Itoh, Y. (2019). Application of Programming Learning Support System to Object-Oriented Language. *Proceedings of the 26th International Conference on Computers in Education*, 348-350.

Sorva, J., Karavirta, V., & Malmi, L. (2013). A Review of Generic Program Visualization Systems for Introductory Programming Education. *ACM Transactions on Computing Education (TOCE), 13*(4), 15.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Combining Data and Human Intelligence through Predictive Visual Analytics to Improve Educational Assessments

Yancy Vance PAREDES[a*] & I-Han HSIAO[b]
[a]*Arizona State University, USA*
[b]*Santa Clara University, USA*
*yvmparedes@asu.edu

Abstract: Assessments are widely used in higher education. It is often the case that most teachers learn the tacit knowledge of creating tests throughout their careers. Despite the proliferation of educational technologies, existing datasets are often still untapped for their potential use to improve test quality. This paper proposes a novel approach to leveraging students' test data to aid teachers in the test construction process. Based on the various student profiles and the domain model learned by the system, teachers can administer their newly created tests through simulation, effectively helping them identify items for revision and improvement without additional effort. The affordances of interactive data visualization are utilized to make such predictive models accessible for the teachers. In addition to the simulation, the system can determine topic coverage, ensuring such test matches a predetermined test blueprint. This system integrates machine intelligence with human intelligence that benefits the teacher in making informed decisions. Students indirectly benefit as they are assessed with better quality tests that could capture their mastery of the domain.

Keywords: Assessments, learning analytics, knowledge tracing, human-data interaction

1. Introduction

Assessments or tests, both formative and summative, are widely used by teachers in higher education to determine their students' mastery. The outcomes of these assessments often form the basis of several important decisions affecting students. Teachers, therefore, have to ensure that these are fair, reliable, valid, and high quality. Unfortunately, there is an abundance of teachers who lack formal training on test construction. Teachers typically learn this skill from experience as they repeatedly create, administer, grade, and revise tests throughout their profession. Ideally, they become better as time progresses, but there is no guarantee. It is typically uncertain how students–even the excellent ones–would perform on a poorly devised test. Just like students, if teachers are unguided or lack feedback, they learn the process inefficiently. However, unlike students, such ineffective performance would affect those beyond themselves.

Existing techniques in psychometrics concerning the quality of a test (e.g., item response theory) often leverage student answers or responses. Several educational systems exist that can capture these data at a finer level. For example, WebPGA is a home-grown web system that streamlines the process of digitizing, grading, and distributing paper-based tests (Paredes & Hsiao, 2021). In addition to performance, it also captures behaviors relating to how students review their mistakes. These types of student data are rich and often untapped. Beyond the overall score, detailed and personalized feedback can be provided to students and teachers by building diagnostic and predictive models.

This paper proposes a novel approach to extending WebPGA by providing a tool to support teachers in test construction. This work takes inspiration from the intelligent tutoring system (ITS) literature, where students' mastery of topics is estimated based on their answers to practice opportunities (Baker, 2016). However, unlike ITS, where there is a learning component between practices, there is not generally a learning opportunity, per se, in our system as students mainly use it to access their graded tests. However, such summative tests could be framed as an opportunity to diagnose student misconceptions to tailor the subsequent tests to their capabilities.

2. Extended Research Platform

VanLehn and colleagues (1994) have outlined possible uses and benefits of simulated students in teacher education. However, most of the approaches mainly focused on instruction rather than testing. With the growing ease of collecting student test data, it is worth exploring the use of simulated students and test data in the professional development of teachers. However, this raises the question regarding the scope of the system's control. How authoritative will the system be? Will it make the decisions for the teachers? Baker (2016) articulated a paradigm shift toward designing educational systems intelligently rather than designing intelligent systems. He further argues that, ultimately, humans should make such critical decisions. Systems could help produce distilled reports that humans can use. Pelánek (2021) outlined several approaches to visualizing learner data and further notes how some of these data are often underutilized. Most of these typically pertain to student test data, such as an item's difficulty and amount of time to answer. Therefore, these prior works motivated the extension and further development of our existing system.

Existing test data is used to create the domain model (e.g., hierarchy and pre-requisite of topics) and the student profiles (e.g., multiple student models). Each student profile will have its associated proficiency parameters for each domain topic (i.e., knowledge component). The predictive model would then use these to determine the probability of a student answering the item correctly. Each item is typically composed of multiple skills. There are several ways in which student models can be developed. Item response theory (IRT) is used in psychometrics to determine students' abilities (e.g., the Rasch model). One issue with the canonical implementation of IRT is that it is unidimensional. This means students are only classified based on their overall ability, which may not be necessarily beneficial in providing a comprehensive and specific view of students' mastery. Another approach is through cognitive diagnostic models (e.g., DINO) (DiBello, Stout, & Roussos, 1995). These models are capable of associating multiple skills to a single item. However, one such limitation is that the latent features (i.e., skills) are dichotomous. This means that a given skill could either be mastered or not by the student with no varying levels of certainty. Another approach is performance factor analysis (Pavlik, Cen, & Koedinger, 2009). It is a logistic regression model that modifies the learning factors analysis. It attempts to address the issues mentioned above. In these models, student performance serves as the dependent variable. These are some of the many approaches to knowledge tracing. Liu and colleagues (2021) provided an overview of several techniques for knowledge tracing and proposed a taxonomy from a technical perspective, namely probabilistic, logistic, and deep learning. Moreover, Gervet and colleagues (2020) compared the predictive performance of these models using various educational datasets and determined the nature of the dataset on which they performed better.

Once the system learns the models, these will be stored for later instantiation in the simulation. The simulation can be configured so that tests can be administered to all the various student profiles previously seen by the system (i.e., for multiple offering courses) or those limited to the current classroom distribution only. Teachers then upload a newly created test which the system parses and automatically assigns the relevant topics. Additionally, it builds the needed inputs for the predictive models (e.g., Q-matrix). Afterward, it predicts the likelihood of simulated students answering the questions correctly (Figure 1). To further aid the process, items would be measured for similarity to make suggestions for teachers to consider. For example, a given topic is assessed multiple times, resulting in redundancy, inefficiency, or insufficient coverage.

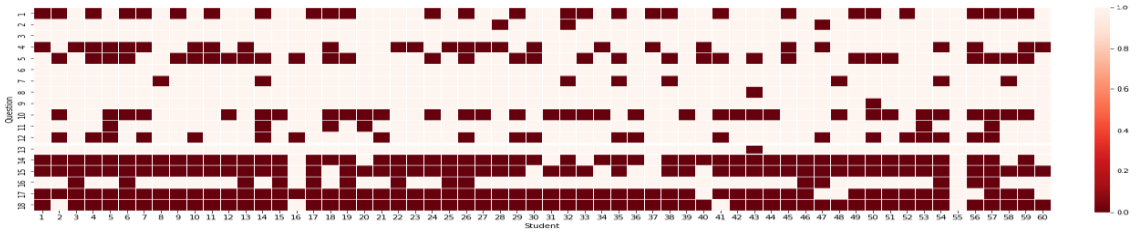


Figure 1. A prototype interactive visualization, rendered using the predictive models, illustrating the likelihood that a simulated student (x-axis) correctly answers a question (y-axis).

A typical concern of developing predictive models is their complex nature and interpretability. This could be addressed by leveraging the affordances of data visualization so that teachers are

presented with easily understood information. Following the principles of Shneiderman's (1996) Information Seeking Mantra or the "overview first, zoom and filter, then details-on-demand" (p. 337), teachers are not overwhelmed with information. Nevertheless, they are provided with a toolkit to encourage exploratory analyses. Interesting questions could be answered with the aid of the system, such as "which items are predicted to be difficult", "what topics do most students find difficult", or "what questions seem to be interdependent". The described workflow is similar to performing a modern test analysis as in psychometrics (e.g., IRT). The only difference is the timing when it is done. In this case, before a test is administered to the actual students giving the teachers a chance to improve. This knowledge could be used to objectively understand whether the students are ready to be tested or not. The created questions could be tailored to the level of performance of students. Ultimately, the teacher decides to stick with the questions or do any revisions. However, the system captures these actions for future analysis, essentially to know how this knowledge impacts teachers' decision-making.

3. Conclusion and Future Work

This work proposes a novel approach to using existing students' test data to support the process of test construction. By extending an existing educational system, insights can be provided to both students and teachers. Modeling approaches from psychometrics and artificial intelligence, such as performance factors analysis, are utilized to create various student profiles and learn the domain model. With the aid of interactive visualizations, teachers are provided with a simulation tool that allows them to forecast the outcome of a newly created test based on students' current level of mastery. Such knowledge empowers them to make informed decisions to improve the test's quality, leading to better and more effective assessments that precisely measure students' knowledge. Arguably, the role of the machine is to aid the amplification of human intelligence (Baker, 2016) with the aid of visualizations (Bassen, Howley, Fast, Mitchell, & Thille, 2018; Pelánek, 2021). In the long term, these revision behaviors of teachers could uncover best practices that can be shared with teachers early in their careers. This helps fully understand how to close the loop toward a shift to automated educational assessments.

To truly determine the system's effectiveness, it will be evaluated in two ways. First, the accuracy of the predictive models needs to be determined and compared. This could be conducted through the use of existing datasets along with the collection of new datasets. Second, the overall perceived usefulness of the system to the teachers should be determined. This will be done through subjective evaluation and surveys to solicit the teachers' feedback.

References

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education, 26*, 600–614.

Bassen, J., Howley, I., Fast, E., Mitchell, J., & Thille, C. (2018). OARS: Exploring instructor analytics for online learning. *Proceedings of the 5th Annual ACM Conference on Learning at Scale*.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cognitively diagnostic assessment* (pp. 361–389).

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep earning the best approach to knowledge tracing? *Journal of Educational Data Mining, 12*(3), 31–54.

Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). A survey of knowledge tracing. https://doi.org/10.48550/arXiv.2105.15106

Paredes, Y. V., & Hsiao, I.-H. (2021). WebPGA: An educational technology that supports learning by reviewing paper-based programming assessments. *Information, 12*(11), Article 450.

Pavlik, P. I., Jr, Cen, H., & Koedinger, K. R. (2009). Performance factors analysis–a new alternative to knowledge tracing. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction, 27*, 313–350.

Pelánek, R. (2021). Analyzing and visualizing learning data: A system designer's perspective. *Journal of Learning Analytics, 8*(2), 93–104.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the 1996 IEEE Symposium on Visual* Languages (pp. 336–343).

VanLehn, K., Ohlsson, S., & Nason, R. (1994). Applications of simulated students: An exploration. *Journal of Artificial Intelligence in Education, 5*(2), 135–175.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

The Influence of a Virtual Physics Experiment Learning Environment on Grade 9 Students' Motivation towards Physics Learning

Yushi HUANG[a*], Alex Wing Cheung TSE[a*]
[a]*Faculty of Education, The University of Hong Kong, Hong Kong*
*u3586759@connect.hku.hk, *awctse@hku.hk

Abstract: Experiments are essential in physics to help students comprehend abstract concepts. Although studies had been conducted to explore whether virtual physics experiments could affect students' learning motivation, they showed inconclusive results. Therefore, this research aimed to investigate the influence of a virtual physics experiment learning environment on the physics learning motivation of Grade 9 students, combining quantitative (quasi-experimental design) and qualitative (student interviews) research methods. Participants of this research were from two different classes, divided into an experimental group (n=37) and a control group (n=37). The intervention lasted for three weeks, with one 45-min physics experiment class per week. Learning motivation was measured by the Physics Learning Motivation Test, which included dimensions like interest-enjoyment, tension-pressure, perceived choice, perceived competence, and perceived value. Based on the data analysis, we found that the virtual experiment learning environment could significantly increase students' learning motivation, especially for the perceived value dimension. Moreover, students who perceived a higher level of competence in the virtual environment were more likely to appreciate its value and utilize virtual experiments again. We expect that the implications of this study and intervention design can be a reference for teachers in incorporating virtual experiments in future physics education and provide a possible solution for conducting physics lessons during the COVID-19 pandemic. More in-depth teacher interviews are recommended to investigate the issues from different pedagogical perspectives.

Keywords: Virtual physics experiment, motivation, learning environment

1. Introduction

Physics could be abstract and complex, which could reduce learners' interest (Awan et al., 2011). Experiments are essential in physics because they are considered to be a useful tool to help students grasp abstract concepts and acquire hands-on skills (Hamed & Aljanazrah, 2020). However, the outdatedness and inadequacies of experimental equipment negatively impacted teaching and learning in traditional physics instruction (Huo, 2015), which dampened students' enthusiasm to learn physics because they didn't have enough opportunities to experience and study physics experiments. Research pointed out that the inadequacies of traditional physics experiments could be compensated by virtual physics experiments (Hamed & Aljanazrah, 2020). Virtual physics experiments could simulate real experiments by visualising some difficult-to-observe experiment phenomena and providing a flexible learning environment (Li, 2015), thus boosting student achievement (Al-Amri et al., 2020), learning interest and engagement (Ogbuanya & Onele, 2018). Although studies had been conducted to explore whether the virtual physics experiment learning environment could influence students' learning motivation, they showed inconclusive results, with an increase (Gunawan et al., 2018), a decrease (Aliane et al., 2010) and no significant effect (Aslan & Duruhan, 2021). Therefore, this study aimed to explore whether the virtual physics experiment learning environment could shape the motivation of Grade 9 students in a middle school in Mainland China. This study investigated two research questions as follows:

1) Does the virtual physics experiment learning environment affect Grade 9 students' motivation towards physics learning?
2) If so, how does the virtual physics experiment learning environment affect Grade 9 students' motivation towards physics learning?

2. Literature Review

2.1 Virtual Experiments to Create Physical Learning Environments

The virtual physics experiment is the product of virtual technology applied to physics education, allowing students to complete physics experiments online without considering time and space (Carnevale, 2003). The effectiveness of using virtual experiments to create learning environments has received considerable attention in research. For instance, Hamed and Aljanazrah (2020) found that flexible virtual experiment learning environments helped undergraduate students better prepared for real experiments. Although many studies had examined the impact of virtual experiments in teaching and learning, most of them focused on higher education and little research investigated the effectiveness in secondary education. Thus, this research aimed to investigate whether Grade 9 students' motivation would be influenced in the virtual physics learning environment and to see if the virtual physics learning environment could be exploited to promote secondary school students' physics learning.

2.2 Virtual Physics Experiment and Learning Motivation

Of the research that examined the impact of virtual physics experiments on students' learning motivation, many studies demonstrated that they could increase students' motivation, while other studies showed different results. For instance, Gunawan et al. (2018) found that students who utilised virtual experiments had higher levels of conceptual understanding because the enjoyment of using virtual experiments increased their motivation. However, Aliane et al. (2010) described the limitations of virtual labs, such as negatively affecting student motivation because of the lack of hands-on activities. Aslan and Duruhan (2021) found no significant difference in motivation between virtual and real learning environments. The above review of studies revealed different findings on the impact of virtual physics experiments on students' motivation. To better stimulate students' learning motivation in physics, it is necessary to conduct this research to explore how the virtual physics experiment learning environment may influence learning motivation.

3. Methodology

3.1 Research Design

This research was conducted in physics experiment classes in a middle school in southern China. Using mixed methods with quantitative (embedded quasi-experimental design) and qualitative (student interviews) approaches, this research investigated the influence of the virtual physics experiment learning environment on Grade 9 students' motivation toward physics learning. Figure 1 shows the overview of the research design. The two groups came from two classes with a similar academic background and were divided into an experimental group (n=37) and a control group (n=37). They were taught by the same teacher, ensuring that their learning processes did not differ significantly. Figure 1 provides an overview of the research design.

Questionnaires and student interviews were applied to collect data. Students' learning motivation was measured by the Physics Learning Motivation Test (PLMT), adapted from the Intrinsic Motivation Inventory (McAuley et al., 1989; Yin et al., 2020). The PLMT includes five areas: interest-enjoyment (IE), tension-pressure (TP), perceived choice (PCh), perceived competence (PCo), and perceived value (PV). Both groups finished the same pre-test and post-test questionnaires. Additionally, six students of the experimental group were randomly selected for interviews.

The intervention lasted for three weeks, with one 45-min physics experiment class per week. Students in the experimental group completed the intervention in the virtual environment, while the control group in the real laboratory. The virtual experimental platform created the virtual experimental environment provided in this study. An experienced physics teacher who joined this study designed the experimental content. The entire intervention consisted of three physics experiment sessions on electrical experiments.

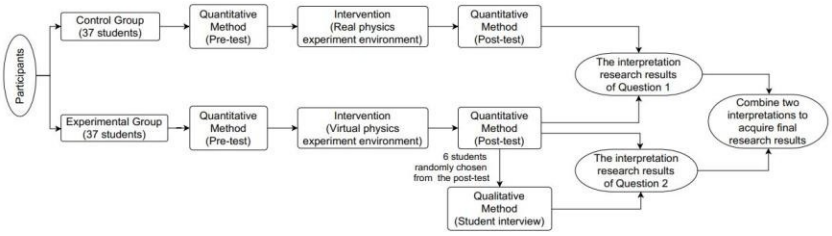


Figure 1. Overview of the Research Design

3.2 Virtual Physics Experiment Platform: Guangdong Virtual Learning Platform

The Guangdong Virtual Learning Platform (*Guangdong Virtual Learning Platform*, 2018) is an innovative digital learning platform widely used by secondary schools in Guangdong Province, a province of southern China. It provides high-quality virtual physics experiments for students and teachers to learn and simulate online, avoiding wasting time restarting experiments due to errors in the real lab. Users can also conduct virtual experiments at any time, which is not feasible in a real lab. Figure 2 shows the translated screenshot of the virtual physics experiment platform.

In this study, the virtual environment could help students learn knowledge and skills by allowing them to complete online experiments related to circuits through interaction with the virtual physics experiment platform. For instance, one of the experiments for Ohm's Law was to investigate the relationship between current and voltage. Students were required to select the appropriate range of ammeter, voltmeter, and other experimental instruments, measure the current through the resistor several times by changing the voltage, record and analyse the current law in the virtual platform, and draw conclusions.

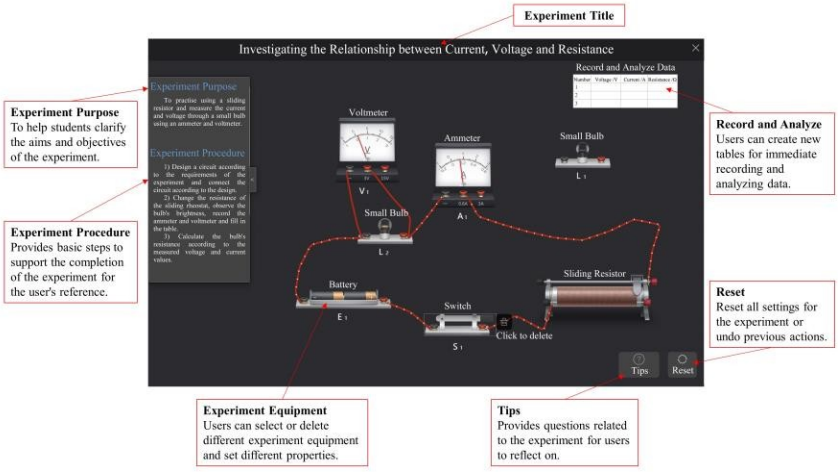


Figure 2. The Translated Screenshot of the Virtual Physics Experiment Platform

4. Results and Findings

The internal reliability of the PLMT was investigated based on Cronbach's alpha. The overall α for the questionnaire was .908, indicating that the questionnaire's data was high quality and reliable enough to evaluate the learning motivation. An independent t-test of pre-test motivation revealed no significant difference between the two groups ($p=.292>0.05$), indicating that these two groups had the same level of motivation before the intervention.

4.1 Impact of the Virtual Experiment Environment on Students' Physics Learning Motivation

The ANOVA was conducted to examine whether there were significant differences in learning motivation in the two learning environments after the intervention. The learning environments were set as independent variables, and the five dimensions of the PLMT were set as the dependent variables. Table 1 shows the result.

Statistical analysis showed that there was no significant difference in IE ($F=1.653$, $p=.203>.05$), TP ($F=.060$, $p=.807>.05$), PCh ($F=3.023$, $p=.086>.05$) and PCo ($F=1.816$, $p=.182>.05$), indicating that the virtual experimental environment could produce the same learning effect (e.g., learning interest, learning pressure) in these four dimensions as the real experimental environment. Moreover, the mean of IE ($M_E=4.927>M_C=4.517$), PCh ($M_E=4.439>M_C=3.946$) and PCo ($M_E=4.378>M_C=3.930$) in the experimental group were higher than in the control group, suggesting that after the intervention, students reported higher motivation in these three dimensions in the physics learning. This meant that students enjoyed the virtual environment more, were more likely to choose the virtual experiment environment if possible and believed that the virtual experiment was more capable of helping them improve their competence in learning physics. The mean of TP ($M_C=2.773>M_E=2.692$) in the control group was higher than in the experimental group, which indicated that students were slightly less stressed and nervous in the virtual environment. However, there was a significant difference in PV ($F=6.543$, $p=.013<.05$), suggesting that the intervention significantly impacted students' perceived value. In addition, based on the mean of PV post-test scores ($M_C=4.345<M_E=5.196$), students perceived a higher value in using the virtual physics experiment.

Table 1. *ANOVA of the Pre-test and Post-Test Learning Motivation*

Test	Dependent Variables	Groups	M	SD	F-value	P-value
Post-test	Interest-enjoyment (IE)	Control	4.517	1.509	1.653	.203
		Experimental	4.927	1.213		
	Tension-pressure (TP)	Control	2.773	1.309	.060	.807
		Experimental	2.692	1.525		
	Perceived Choice (PCh)	Control	3.946	1.403	3.023	.086
		Experimental	4.439	1.004		
	Perceived Competence (PCo)	Control	3.930	1.452	1.816	.182
		Experimental	4.378	1.411		
	Perceived Value (PV)	Control	4.345	1.500	6.543	.013
		Experimental	5.196	1.359		

The first research question can be answered based on the previous findings: the virtual physics experiment environment can increase students' physics learning motivation. Furthermore, the ANOVA results showed that the perceived value dimension presented the most significant changes, suggesting that the students in the experimental group perceived the value of using the virtual physics experiment, which boosted their motivation.

4.2 Students' Perceptions of Virtual Physics Learning Environment

Student interviews were conducted to investigate how the virtual physics experiment learning environment affects Grade 9 Students' motivation toward physics. Six participants from the experimental group were randomly invited. To protect privacy, we coded participants from ST1 to ST6. For example, ST1 is the first participant to participate in the interview.

Most students (n=5) thought they were more motivated to learn physics because they perceived improved competence in the virtual learning environment. ST1 and ST3 believed that the virtual platform had improved their problem-solving ability. ST2 improved experimental ability by repeatedly performing experimental procedures in the virtual platform. ST4 and ST5 believed that their concept-understanding ability had improved because the virtual platform helped them understand the knowledge better and made physics less complicated and abstract. Those who felt their competence had improved in the virtual environment also recognised its value. The students perceived

the increased competence as the most significant value that the virtual environment brought them (ST1, ST2, ST4, and ST6), which motivated them to be choose the virtual environment again. For example, ST2 said, "*after using the virtual experiment, I felt my experimental ability and knowledge improved. This made me more courageous to challenge difficult physics problems because I can use virtual experiments to simulate the scenes and help me understand the problems.*" This indicated that once the students perceived good experimental competence in the virtual environment, they would choose the virtual environment again to challenge difficult problems.

In general, the flexible virtual experimental environment could provide students with a good learning experience, help them improve their abilities and solve physics challenges, and therefore have a high potential to increase their motivation. Especially those students with high perceptual competence could feel a higher value in using virtual experiments to learn physics and thus be more motivated.

5. Discussion: Virtual Physics Experiment Environment Could Increase Motivation

Although the results demonstrated that students' motivation rose in both learning environments, students in the experimental group showed significant changes after the intervention, suggesting that students perceived higher motivation in the virtual environment. It was consistent with the findings of Maksoud (2018) and de Vries and May (2019). As a novel teaching tool, virtual experiments have considerable potential to motivate students to learn physics. Based on the student interviews, this is mainly attributed to the flexibility of the virtual environment. It respected students' autonomy and provided opportunities for repetitive practice, significantly increasing participants' motivation in the experimental group. When students were given choices to do what they wanted, it positively affected their intrinsic motivation (Deci & Ryan, 1987). In addition, student interviews also confirmed that those who perceived increased experimental ability in the virtual environment were more self-motivated, which explained why the experimental group witnessed a higher level of interest-enjoyment, perceived choice, and perceived competence than the traditional group. Such qualities are of immense significance to students' academic development. Therefore, the numerous potentials of virtual physics experiments as a more effective and efficient teaching aid to motivate students remains to be tapped to a greater extent.

As shown in this research, the virtual experimental environment significantly increased perceived value. Further research found that students with high perceived competence had higher perceived value, and thus were more motivated to learn physics. One possible interpretation was that students perceived increased competency in the flexible virtual learning environment, thus recognising the value of virtual learning and gaining confidence to challenge difficulties (Rodriguez et al., 2021), so they may choose to use it again as a learning tool. This is because the flexible virtual learning environment allowed students to repeat the experiment at any time, which was restricted in the real lab. Therefore, if learning designers want to increase students' motivation to learn physics, they are recommended to design instructional strategies that would enhance students' perceived competence to increase their recognition of the value of the virtual physics experiment environment.

6. Conclusion

This study used a mixed methods design with a quantitative (quasi-experimental design) and qualitative approach (student interviews) to explore whether virtual physics experiments could affect Grade 9 students' learning motivation. The results showed that the virtual physics experiment learning environment could significantly increase participants' motivation, especially in the perceived value dimension. The student interviews also aligned with the results of the quantitative analysis that students with a high perceived competence were more likely to increase their motivation through the virtual physical experiment environment. This research has contributed to filling the gap left by current findings on using virtual physics experiments to motivate students. Furthermore, the intervention in this study created a virtual physics experiment-based instructional design, which can be a reference for learning designers and teachers in incorporating virtual experiments in future physics education and provide a possible solution for conducting physics lessons during the COVID-19 pandemic. The limitation of this study is that the intervention lasted only three weeks due to the course content. If the duration could be extended, it might deepen the impact and thus increase the

potential for motivational change. In addition, teacher perspectives could be considered to investigate the issues from different pedagogical perspectives.

Acknowledgements

We would like to thank the teachers and students who participated in the research process and provided data.

References

Al-Amri, A., Osman, M. E., & Musawi, A. (2020). The Effectiveness of a 3D-Virtual Reality Learning Environment (3D-VRLE) on the Omani Eighth Grade Students' Achievement and Motivation towards Physics Learning. *International Journal of Emerging Technologies in Learning*, *15*(05), 4-16.

Aliane, N., Pastor, R., & Mariscal, G. (2010). Limitations of Remote Laboratories in Control Engineering Education. *International Journal of Online Engineering*, *6*(1), 31-33.

Aslan, S. A., & Duruhan, K. (2021). The effect of virtual learning environments designed according to problem-based learning approach to students' success, problem-solving skills, and motivations. *Education and Information Technologies*, *26*(2), 2253-2283.

Awan, R.-U.-N., Sarwar, M., Naz., & A., N., G. (2011). Attitudes Toward Science Among School Students Of Different Nations: A Review Study. *Journal of College Teaching & Learning (TLC)*, *8*(2).

Carnevale, D. (2003). The Virtual Lab Experiment: Some Colleges Use Computer Simulations to Expand Science Offerings Online. *The Chronicle of Higher Education*, *49*(21), A.30.

de Vries, L. E., & May, M. (2019). Virtual laboratory simulation in the education of laboratory technicians-motivation and study intensity. *Biochemistry and Molecular Biology Education*, *47*(3), 257-262.

Deci, E. L., & Ryan, R. M. (1987). The Support of Autonomy and the Control of Behavior. *Journal of Personality and Social Psychology*, *53*(6), 1024-1037.

Guangdong Virtual Learning Platform. (2018). Guangdong Virtual Learning Platform Website. Retrieved July 31, 2022 from https://www.gdtextbook.com/

Gunawan, G., Nisrina, N., Y Suranti, N. M., Herayanti, L., & Rahmatiah, R. (2018). Virtual Laboratory to Improve Students' Conceptual Understanding in Physics Learning. *Journal of Physics. Conference Series*, *1108*(1), 12049.

Hamed, G., & Aljanazrah, A. (2020). The Effectiveness of Using Virtual Experiments on Students' Learning in the General Physics Lab. *Journal of Information Technology Education: Research*, *19*, 977-996.

Huo, D. Y. (2015). The Design of Teaching Experiment System Based on Virtual Instrument Technology. *Proceedings of the 2015 International Conference on Management, Education, Information and Control*, *125*, 53-57.

Li, Y. F. (2015). The Application of the Virtual Experiment in Physics Teaching. *Proceedings of the 2015 International Conference on Education Technology, Management and Humanities Science (Etmhs 2015)*, *27*, 1225-1228.

Maksoud, N. F. A. (2018). When Virtual Becomes Better than Real: Investigating the Impact of a Networking Simulation on Learning and Motivation. *International Journal of Education and Practice*, *6*(4), 253-270.

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*(1), 48-58.

Ogbuanya, T. C., & Onele, N. O. (2018). Investigating the Effectiveness of Desktop Virtual Reality for Teaching and Learning of Electrical/Electronics Technology in Universities. *Computers in the Schools*, *35*(3), 226-248.

Rodriguez, S., Estevez, I., Pineiro, I., Valle, A., Vieites, T., & Regueiro, B. (2021). Perceived Competence and Intrinsic Motivation in Mathematics: Exploring Latent Profiles. *Sustainability*, *13*(16), 8707.

Yin, J., Goh, T.-T., & Yang, B. (2020). Conversation Technology With Micro-Learning: The Impact of Chatbot-Based Learning on Students' Learning Motivation and Performance. *Journal of Educational Computing Research*, *59*(1), 154-177.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Extracting Students' Self-Regulation Strategies in an Online Extensive Reading Environment using the Experience API (xAPI)

Chia-Yu HSU [a*], Izumi HORIKOSHI[b], Huiyong LI[b], Rwitajit MAJUMDAR[b] & Hiroaki OGATA[b]
[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*hsu.chiayu.25t@st.kyoto-u.ac.jp

Abstract: Extensive Reading (ER) activity is useful in language learning where learners pick any reading materials in target language by themselves and continue reading. In this study, we aim to understand students' self-regulation behaviors and strategies in ER so that facilitators can give learning strategy-based instruction. For this purpose, we first explored a potential structure of a learners' model that can highlight their learning strategies and extracted from multi-system interaction logs recorded in Experience API (xAPI) format. To demonstrate, we collected data from 120 students across 3 months when they did ER activities. We analyzed and extracted the self-regulation strategies of the learners and found 2 groups by applying K-means cluster analysis. The results inform dashboard design and instructional support based on the visualized attributes of the cluster members. This study contributes towards using interoperability standards to record learner's online reading behaviors and demonstrate how teaching and learning activities can be supported by xAPI when such experiences are distributed across various learning tools.

Keywords: xAPI, Learning Analytics, Extensive Reading, Learning Strategies

1. Introduction

Considering the effect of improving students' reading efficiency, Extensive Reading (ER) is regarded as an important component in reading skills (Mustapa, Rahman, Ghani, Saad, & Mohamed, 2018; Takii, Flanagan, & Ogata, 2021). In recent years, the advance of ICT technology has enabled a variety of support for the learning from higher education to K12 (Lin et al., 2016). In terms of the selection of ER materials, some personalized e-learning recommender systems can adapt to learners' different interests and levels, such as the system proposed by Takii et al. (2021). An online tool, M-Reader, proposed by Robb and Kano (2013) allowed students to take a short quiz after finishing a book and reduced teachers' workload on keeping track of their students' reading. To support self-regulated learning, the GOAL system, proposed by Li, Majumdar, Chen and Ogata (2021), scaffolds students to schedule their learning activities and enables their data informed decision making.

The researchers suggested that to get students to read more, it may be necessary to create monitoring systems that are built to facilitate the implementation of ER by teachers (Chang & Renandya, 2017; Renandya & Jacobs, 2002; Mustapa et al., 2018). Frequent questions are often raised about how to integrate student data from heterogeneous learning contexts with different software components, and what should be tracked to obtain useful information and avoid noisy data (Carrillo, Lavoué, & Prié, 2016). Tin Can API, also known as Experience API or xAPI, has a simple vocabulary able to capture and share the activity flow of almost any tool, which makes it a potential mechanism to manage and visualize the data of learning experience, leading to more adaptive and personalized learning services (Lin et al., 2016).

In terms of teachers' monitoring for ER activity, it is crucial to understand students' self-regulation behaviors and strategies in ER so that teachers can give learning strategy-based instruction. Hence, we find it necessary for a potential structure that can help describe learners' behaviors and extract their self-regulation strategies from logs recorded in xAPI considering that ER activity can be supported by different technologies. 2 research questions are: (1) How can students'

learning behaviors and strategies be extracted from logs recorded in xAPI? (2) What are the students' self-regulation behaviors and strategies in ER?

2. Foundation

2.1 Experience API (xAPI) to Extract Learning Strategies

Vast quantities of data on student's online activity are recommended as observable indicators, supporting valid inferences about a learner's learning if taken advantage of. Different Learning Analytics (LA) methods can be used to help make meaningful sense from the data of massive volume and high rate of velocity (Viberg, Khalil, & Baars, 2020; Raga Jr, Raga, & Cariño, 2018). Saint, Gašević and Pardo (2018) used process mining techniques to identify strategic and tactical learner behaviors and found that certain temporal activity traits relate to performance in the summative assessments attached to the course, mediated by strategy type. Carlson, Genin, Rau and Scheines (2013) used an expectation-maximization approach to extract the error-making and hint-seeking behaviors of each student to characterize their learning strategy. However, many of these methods are limited in their ability to describe and interpret differences between extracted latent representations at varying levels of granularity, such as underlying data of student actions and behavior.

xAPI is a specification for learning technology that allows recording a wide range of experiences that a user can have in different technologies or tools (Lin et al., 2016). An event is captured as a Statement. The main elements in a Statement are the Actor, the Verb (action) and the Object, which forms the simplest event case: "[actor] [verb] [object]" (Manso-Vázquez, Caeiro-Rodríguez, & Llamas-Nistal, 2018; Carrillo et al., 2016). The limited but expandable vocabulary makes it easy to record new actions, and hence makes the creation of specific domain profiles possible (Manso-Vázquez et al., 2018; Serrano-Laguna et al., 2017). The concept of application profile was defined as "a way to declare which elements from which namespaces are used in a particular application or project". xAPI-SRL, for example, provides a standard way to record all the actions and activities typical in self-regulated learning. Different use cases covered in an xAPI profile are usually formed by several recipes based on the development of Communities of Practice. In xAPI-SRL, the large number and variety of strategies results in different recipes into which the identified actions are divided. (Manso-Vázquez et al., 2018). That is, an xAPI profile can give structure to a whole learning activity, describing learning strategies as well as the study tactics that compose them.

2.2 Extensive Reading in the Learning Evidence Analytics Framework (LEAF)

In this study, we investigated students' behaviors of extensive reading in a Japanese junior high school, in which the learning activities are supported with various tools in the Learning Evidence Analytics Framework (LEAF). Figure 1 shows the learning tools composed of the framework along with the affordances of each tool. Examples of learner actions and the trace data generated are also given.

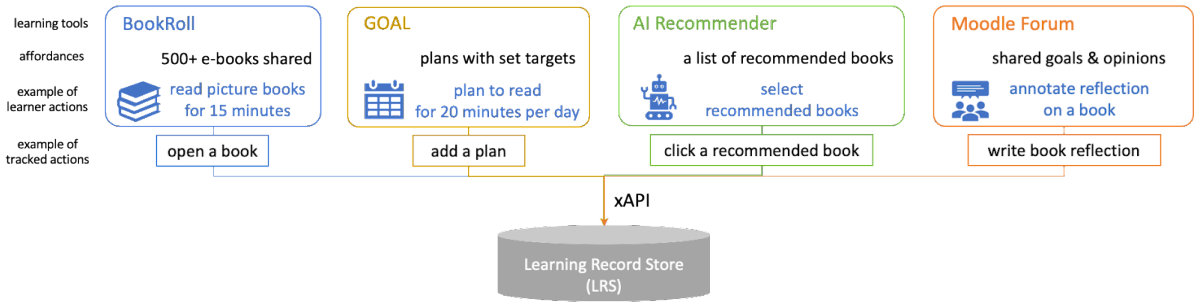


Figure 1. Extensive Reading in LEAF.

In the ebook reader, BookRoll (Ogata et al., 2015), students could read more than 500 English picture books with different levels. The self-directed learning scaffolding system, Goal-oriented Active Learning System (GOAL), supports students to make reading plans to achieve targets set by themselves. To help students find a suitable book, the material recommending system, AI Recommender, generates recommendations based on students' previous learning activities. With the Learning Management System (LMS), Moodle, teachers launched a forum in which the students could share their opinions with each other every month. The following describes an example action for which the xAPI object is generated in those tools. "open a book" is generated when students read picture books in BookRoll. Plans made in the GOAL system are recorded as "add a plan". When a book is selected in AI Recommender, students are recorded to "click a recommended book". The reflection shared in the Moodle Forum generates the action: "write book reflection". Trace data of students' actions are logged in the Learning Record Store (LRS).

3. Extracting Self-regulation Extensive Reading strategies

3.1 Aggregating the Actions

To extract the learning strategies implied by the actions, we aggregated the actions into a common xAPI statement structure with the xAPI-SRL profile (Manso-Vázquez et al., 2018), introduced in Section 2. Among 11 proposed recipes, we identified 6 recipes according to the functionality of the systems where the actions were taken. Table 1 summarizes the actions in the systems in LEAF as well as the verbs and activity types of objects which were used on the recommendation of different recipes. Different recipes can be used to describe different system functionality. For instance, the 3 recipes used for GOAL reflect the system's functionality of supporting learners' task and goal management, including planning their goals, and controlling the time and flow to complete tasks and achieve goals. In each of the recipes, different statements are used to describe actions with recommended vocabulary of verbs and activity types of objects for the actions. In the GOAL system, its functionality of task and goal management supports learners to add a plan. The recipe reflecting this function controls that the verb, created, is used to indicate the activity type of object, task. Hence, in the common xAPI statement structure, we defined the action of adding a plan as follows: the verb, created, should be used to indicate a plan added by a learner, and the plan (object) should be indicated as the activity type, task.

Table 1. *The Recipes Used to Aggregate the Actions in LEAF*

Recipe	Task and goal management			Planning	Time and flow control		View control	Annotation	Curation
SRER[1] Actions	add a plan	edit a plan	delete a plan	select WPM[2] as target	estimate reading time as target	suspend a plan	open a book	write book reflection	click a recommended book
Verbs	created	edited	deleted	selected	estimated-duration	suspended	read	annotated	selected
Activity Types	task	task	task	goal	goal	task	book	note	book
System	GOAL						BookRoll	Moodle Forum	AI Recommender

[1]SRER: self-regulated Extensive Reading
[2]WPM: words per minute

3.2 Extracting the Strategies

When developing the profile, Manso-Vázquez et al. (2018) divided complex composite strategies into simple strategies and identified the actions and steps learners should follow to use the strategy. They defined the information and data resulting from the performance of the strategy in the software, including the actors, action, object of the action, results, context, time, and so on. On the basis of the verbs and activity types of objects, we corresponded the actions in LEAF to 7 simple strategies as summarized in Table 2. For example, we defined the action of adding a plan as: a task is created. This action involves the simple strategy: Task definition/identification. Based on the categorization of

Manso-Vázquez et al. (2018), we also corresponded the strategy to the composite strategy it belongs to: Metacognitive strategies.

Table 2. *The Categorization of the Strategies Corresponding to the Actions in LEAF*

composite strategy	simple strategy	actions in LEAF
Metacognitive strategies	Task definition/identification	add/edit/delete a plan
	Time estimation	estimate reading time as target
	Time use evaluation	suspend a plan
Cognitive strategies	Reading speed	select WPM* as target
	Content/information review	open a book
	Annotation (note taking, highlighting)	write book reflection
	Selection of information sources	click a recommended book

*WPM: words per minute

4. Implementation of the Extraction from Multi-system Trace Data

4.1 Methods

The data were collected from a junior high school in Japan. 120 seventh-graders participated in ER activity from April to June in 2021. We aggregated the action logs into the xAPI statement structure described in Table 1. We then extracted the learning strategies from the actions using the categorization in Table 2. The strategies are represented by the total numbers of the corresponding actions taken. For example, we counted how many times a student added/edited/deleted a plan respectively. Since these actions involve the simple strategy: Task definition/identification, we considered the occurrence where the students had this strategy by summing up the times s/he added/edited/deleted a plan during the period. Based on the numbers of the simple strategies extracted, we used K-means Cluster Analysis to cluster the students. We identified 2 as the optimal number of clusters with the Average Silhouette Method.

4.2 Results

We could label the two clusters which emerged in K-means Cluster Analysis as: Metacognitive & Cognitive Learners (n=11) and Metacognitive Learners (n=109). Table 3 summarizes the average counts of each simple strategy in the 2 clusters.

Table 3. *The Statistics and Differences of Strategies in the 2 Clusters*

	Metacognitive & Cognitive Learners		Metacognitive Learners	
	M	SD	M	SD
Task definition/identification	3.73	0.90	4.60	4.04
Reading speed	1.64	0.67	1.72	0.97
Time estimation	1.64	0.50	1.61	1.05
Time use evaluation	0.00	0.00	0.26	1.88
Content/information review ***	39.10	12.89	5.90	7.63
Annotation (note taking, highlighting) ***	13.91	8.63	0.61	1.45
Selection of information sources	3.55	8.41	0.07	0.30

*** $p < .001$

The clusters indicate the composite strategies derived from the simple strategies extracted from the actions according to categorization in Table 2. In the 3-month ER activity, Metacognitive Learners tended to focus on organizing their reading plans but didn't spend much time on reading books and sharing their reflection (where the average counts of Metacognitive strategies are higher than those of Cognitive strategies). On the other hand, Metacognitive & Cognitive Learners made reading plans as often as Metacognitive Learners did (where their average counts of Metacognitive strategies are similar to those of Metacognitive Learners), and they also engaged in reading books either themselves or on the recommendation of AI Recommender and sharing their opinion in the Moodle forum (where the average counts of Cognitive strategies outstand the Metacognitive cluster).

5. Discussion and Conclusion

To facilitate teachers' monitoring for ER activity, we explored a potential structure of a learners' model that can highlight their self-regulation strategies and extracted from multi-system interaction logs recorded in xAPI format. This shows an inventive use of interoperability standards to record learner's ER behaviors. 2 clusters of the students were identified and labeled as: Metacognitive & Cognitive Learners and Metacognitive Learners. We find the results can inform dashboard design and instructional support based on the visualized attributes of the cluster members, including the behavior of a single student and the behaviors of the clusters implying different learning strategies. This helps teachers monitor their behaviors, be aware of the implied strategies, and thus give learning strategy-based instruction.

We propose to design the visualization as 2 parts. On one hand, teachers can have a glance at students' actions by individual, which implies their behaviors in ER activity. On the other hand, teachers can also look at students' learning strategies by cluster, which implies their self-regulation strategies in ER activity. In terms of the former design, teachers can understand students' learning status, such as how many books were read, how many plans were made, how many reflections were shared, etc. As for the latter design, teachers can monitor students' strategies in a similar way. The visualizations may enable teachers to be aware that Metacognitive Learners only focused on organizing their reading plans but didn't read many books and share their reflection. Then, teachers can encourage these students to read more books so that cognitive strategies can be prompted.

The proposed design demonstrates how teaching and learning activities can be supported by xAPI when such experiences are distributed across various learning tools. The visualizations are expected to solve the difficulty of monitoring encountered in ER implementation. However, further evaluation is needed in future research. In addition, xAPI allows integrating student data from heterogeneous learning contexts with different software components. Hence, future research can also explore the possibility of extension to other learning activities.

Acknowledgements

This work is funded by JSPS KAKENHI 20K20131, 22H03902, NEDO JPNP18013, JPNP20006.

References

Carlson, R., Genin, K., Rau, M. A., & Scheines, R. (2013). Student Profiling from Tutoring System Log Data: When do Multiple Graphical Representations Matter?. In *Proceedings of the 6thInternational Conference on Educational Data Mining*.

Carrillo, R., Lavoué, É., & Prié, Y. (2016, April). Towards qualitative insights for visualizing student engagement in web-based learning environments. In *Proceedings of the 25th international conference companion on world wide web* (pp. 893-898).

Chang, A. C., & Renandya, W. A. (2017). Current practice of extensive reading in Asia: Teachers' perceptions. *The Reading Matrix: An International Online Journal, 17*(1), 40-58.

Li, H., Majumdar, R., Chen, M. R. A., & Ogata, H. (2021). Goal-oriented active learning (GOAL) system to promote reading engagement, self-directed learning behavior, and motivation in extensive reading. *Computers & Education, 171*, 104239.

Lin, C. H., Hu, S. S., Lai, H. Y., Chiang, C. F., Tseng, H. C., & Cheng, Y. C. (2016, October). Visca: A dashboard system to visualize learning activities from e-learning platforms. In *International symposium on emerging technologies for education* (pp. 422-427). Springer, Cham.

Manso-Vázquez, M., Caeiro-Rodríguez, M., & Llamas-Nistal, M. (2018). An xAPI application profile to monitor self-regulated learning strategies. *IEEE Access, 6*, 42467-42481.

Mustapa, A. M., Rahman, Z. A., Ghani, M. Z. A., Saad, M. F. M., & Mohamed, F. A. (2018). Qiraahbot's Prototype Development for An Extensive Reading Activity. *International Journal of Civil Engineering And Technology (IJCIET)*, *9*(9), 1494-1503.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015, January). E-Book-based learning analytics in university education. In *International conference on computer in education (ICCE 2015)* (pp. 401-406).

Raga Jr, R. C., Raga, J. D., & Cariño, I. V. (2018). Visualizing Student Activity in Blended Learning Classroom by Mining Course Log Data. *International Journal of Learning and Teaching, 4*(1), 1-6.

Renandya, W. A., & Jacobs, G. M. (2002). Extensive reading: Why aren't we all doing it. *Methodology in language teaching: An anthology of current practice*, 295-302.

Robb, T., & Kano, M. (2013). Effective extensive reading outside the classroom: A large-scale experiment. *Reading in a Foreign Language, 25*, 234-247.

Saint, J., Gašević, D., & Pardo, A. (2018, September). Detecting learning strategies through process mining. In *European Conference on Technology Enhanced Learning* (pp. 385-398). Springer, Cham.

Serrano-Laguna, Á., Martínez-Ortiz, I., Haag, J., Regan, D., Johnson, A., & Fernández-Manjón, B. (2017). Applying standards to systematize learning analytics in serious games. *Computer Standards & Interfaces, 50*, 116-123.

Takii, K., Flanagan, B., & Ogata, H. (2021). An English Picture-book Recommender System for Extensive Reading Using Vocabulary Knowledge Map. In *Companion Proceedings of the 11th International Conference on Learning Analytics and Knowledge*.

Viberg, O., Khalil, M., & Baars, M. (2020, March). Self-regulated learning and learning analytics in online learning environments: A review of empirical research. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 524-533).

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Process Models Enhancement with Trace Clustering

Wiem HACHICHA[a,b*], Ronan CHAMPAGNAT[b], Leila GHORBEL[a] & Corinne Amel ZAYANI[a]
[a] *MIRACL Laboratory, Sfax University, Tunisia*
[b] *L3i Laboratory, La Rochelle University, France*
*wiem.hachicha1@univ-lr.fr

Abstract: Learning Management Systems collect data (such as event logs) about learners and trainers. There are techniques for analysing this data such as Educational Process Mining. In our previous work, we proposed an approach that extracts knowledge about learning paths based on process mining algorithms by generating process models. The latter are used for learning resource recommendation by taking into account learning features (learning style, interests, learning results, etc.). This approach is available for a limited size of event logs. In fact, process models generated from event logs of large classes are not expressive. Trace clustering is one of the successful methods that lead to overcome this limitation. For this reason, we aim to improve the previous approach by using trace clustering in order to characterise learners before the discovery of the corresponding process models. We applied the proposed approach on a Moodle dataset of 100 undergraduate students. Results show that trace clustering improve the general quality of discovered process models.

Keywords: Moodle, event logs, process mining, process discovery, trace clustering, recommendation

1. Introduction

Process mining aims to discover, monitor, and improve real processes by extracting knowledge from event logs readily available in current information systems (Van Der Aalst et al., 2012). One of the uses of process mining is process discovery. A discovery algorithm takes an event log and produces a model without using any prior information (Van Der Aalst et al., 2012). If process mining techniques are applied to educational data, they are often referred to as Educational Process Mining.
In order to evaluate the results of process discovery algorithms, three main quality dimensions (metrics) are invested (Buijs et al., 2012): 1) Fitness (F) which determines how well the model allows the behaviour present in the event log, 2) Precision (P) that corresponds to the rate of activities in the event logs compared to the total of activities observed in the process model and 3) Generalization (G) measures the ability of the model to generalize the behaviour present in the event log.

In our previous work (Hachicha et al., 2021), we proposed an architecture leading to discover process models and recommend to the learner not only learning resources but also process models, each of which is relative to a specific learning resource. We applied process mining techniques from event logs, learning results, and learner's profiles in order to make useful recommendations to learners that would enable them to build up their own learning scenario. This approach was validated on a Moodle-based course of 100 undergraduate students. Results are encouraging but may be difficult to scale to large classes such as MOOCs. Indeed, process models generated from event logs of large classes are not expressive.

In order to overcome this limitation various works in literature used trace clustering (Song et al., 2009), (Xu & Liu, 2019) and (Faizan et al., 2021). The trace represents a sequence of events of the same process instance (in our case the learners) (Van der Aalst, 2016). An event contains information such as the identifier of each learner, the activity name, the timestamp, etc. Indeed, trace clustering identifies homogeneous sets of traces in a heterogeneous event log and allows the discovery of many, simpler process models. For example, authors in (Faizan et al., 2021) have used the concept of trace profiles in order to cluster traces. They have employed the activity profile where each trace is transformed into a vector of features that represents the frequency of activities.

We aim in this work to improve the previous architecture by applying the trace clustering. The research questions tackled in this study are: is it possible to use trace clustering for e-learning traces? and does it give an advantage compared to applying process mining directly?

2. Proposed Approach

In our previous work, we proposed an architecture that use process mining for learning resource recommendation (Hachicha et al., 2021). This architecture is composed of four layers which are source, client, recommendation, and process mining. The source layer involves distributed databases of event logs and process models. The event logs are files that contain a large amount of raw data about the interaction of learners with Learning Management Systems (LMS). The client layer allows the interaction between the learner and e-learning systems (LMS). The recommendation layer recommends not only learning resources but also process models generated from the process mining layer. The latter allows the discovery of process models based on event logs. Each process model demonstrates the most common usage behaviour by learners in an LMS.

In this work, we propose to extend the process mining layer by the step of trace clustering to enhance the quality of process models. This step aims at finding such homogenous groups of process instances. First, we filter the event log by instances to ensure that only learners' activity logs were kept in the event log. Second, we anonymize the learners. We converted the learners' names into IDs to maintain their anonymity and ensure the principles of ethics. Finally, we used the trace clustering method to split the event log into several homogeneous sub-event logs, which are used to generate the corresponding process models. We created the activity profile which is used to split the event log. The activity profile is an NxM matrix; N represents the number of process instances and M represents the number of activities. Each row in this matrix corresponds to a vector trace which is composed of activity frequencies (the number of occurrences of each activity in the trace).

3. Experimental results

The case study uses an event log extracted from the Moodle platform of the University of La Rochelle (France). The event log includes 42,438 traces of 100 students that learned a course over one semester.

We implemented the different steps of the process mining layer using the PyCharm tool. In the trace clustering method, we created the activity profile. Based on this activity profile, k-means algorithm is employed to split the event log. Actually, we conclude that the optimal number of clusters for k-means clustering in our case study is 5 using the elbow method.

At last, we evaluated process mining algorithm resting on the result of trace clustering. In fact, we used the heuristic miner process discovery algorithm to obtain the process models representing the event log's behaviour. Subsequently, the Fitness (F), Precision (P) and Generalization (G) of discovered models were measured (cf. **Error! Reference source not found.**). These metrics are foregrounded in section **Error! Reference source not found.**.

To compare our new results with the results obtained during our previous work (Hachicha et al., 2021), we have calculated the average of each metric (cf. **Error! Reference source not found.**).

Table 1. *Evaluation metrics obtained on process models mined from the used data set*

	F	P	G
Cluster0	0.9922	0.1951	0.7718
Cluster1	0.9847	0.2218	0.8235
Cluster2	0.9652	0.1624	0.6668
Cluster3	0.9838	0.2884	0.7663
Cluster4	0.9923	0.1486	0.7938

Table 2. *Comparison of results*

	F	P	G
Previous approach	0.9903	0.1597	0.7609
New approach	0.9836	0.2023	0.7644

As plotted in **Error! Reference source not found.**, the heuristic miner used in the new approach proves better performance in the precision value (0.2023) and generalization value (0.7644) than the heuristic miner used in the previous approach.

4. Conclusion

In this work, we improve our proposed architecture by extending the process mining layer by trace clustering step. This step is based on the activity profile. We experimented the proposed architecture on a dataset extracted from the Moodle LMS. We derive 5 clusters corresponding to 5 variant of learning scenarios. Therefore, we compare the quality metrics of extracted process models with our previous approach. The study shows that the results have been enhanced even on a, relatively, small dataset. The results validate the possibility of using trace clustering on data coming from e-Learning traces.

As future work, we would enhance the activity profile. We plan also to conduct experiments on larger e-learning data such as MOOC datasets.

Acknowledgements

This work was financially supported by the PHC Utique program of the French Ministry of Foreign Affairs and Ministry of Higher Education and Research and the Tunisian Ministry of Higher Education and Scientific Research in the CMCU project number 22G1403.

References

Buijs, J. C. A. M., Van Dongen, B. F., & Van Der Aalst, W. M. P. (2012). On the role of fitness, precision, generalization and simplicity in process discovery. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *7565 LNCS*(PART 1), 305–322. https://doi.org/10.1007/978-3-642-33606-5_19

Faizan, M., Zuhairi, M. F., & Ismail, S. (2021). Process discovery enhancement with trace clustering and profiling. *Annals of Emerging Technologies in Computing*, *5*(4), 1–13. https://doi.org/10.33166/AETIC.2021.04.001

Hachicha, W., Ghorbel, L., Champagnat, R., Zayani, C. A., & Amous, I. (2021). Using process mining for learning resource recommendation: A Moodle case study. *Procedia Computer Science*, *192*, 853–862. https://doi.org/10.1016/j.procs.2021.08.088

Song, M., Günther, C. W., & Van Der Aalst, W. M. P. (2009). Trace clustering in process mining. *Lecture Notes in Business Information Processing*, *17 LNBIP*, 109–120. https://doi.org/10.1007/978-3-642-00328-8_11

Van der Aalst, W. (2016). Process mining: Data science in action. *Process Mining: Data Science in Action*, 1–467. https://doi.org/10.1007/978-3-662-49851-4

Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., Bose, J. C., Van Den Brand, P., Brandtjen, R., Buijs, J., Burattin, A., Carmona, J., Castellanos, M., Claes, J., Cook, J., Costantini, N., Curbera, F., Damiani, E., De Leoni, M., … Wynn, M. (2012). Process mining manifesto. *Lecture Notes in Business Information Processing*, *99 LNBIP*(PART 1), 169–194. https://doi.org/10.1007/978-3-642-28108-2_19

Xu, J., & Liu, J. (2019). A Profile Clustering Based Event Logs Repairing Approach for Process Mining. *IEEE Access*, *7*, 17872–17881. https://doi.org/10.1109/ACCESS.2019.2894905

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Applicability and Reproducibility of Peer Evaluation Behavior Analysis Across Systems and Activity Contexts

Izumi HORIKOSHI[a*], Changhao LIANG[b], Rwitajit MAJUMDAR[a] & Hiroaki OGATA[a]

[a]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
[b]*Graduate School of Informatics, Kyoto University, Japan*
* horikoshi.izumi.7f@kyoto-u.ac.jp

Abstract: Learning Analytics research provides findings and analytical methods. However, those have not been frequently shared and reused in other studies. The objective of this study is to clarify the possibilities and challenges in applying a behavioral analysis method to other system contexts. This study uses the Evaluation Behavior Analysis (EBA) as an example of an analytical method, and compared the applicability of the method and reproducibility of findings in two different studies: the original dataset (Study A), and which applied EBA (Study B). Not all methods were able to be applied as the data and activity were different. However, the reproducibility was far higher than we expected. This research contributes to expanding the reuse of analysis methods for small-scale systems such as EBA, which have not usually been reused, and further development in Learning Analytics.

Keywords: Peer Evaluation, Behavior, Learning Analytics, Applicability, Reproducibility

1. Introduction

Learning Analytics research has been conducted for a range of different purposes, and has focused on various learning activities (Lang, Siemens, Wise, & Gasevic, 2022). In fields where knowledge sharing and accumulation have begun, it is necessary to consider the reproducibility and replication of research findings. In education science research, the US National Science Foundation (NSF) and the Institute of Education Sciences (IES) jointly released the Companion Guidelines on Replication and Reproducibility in Education Research (NSF & IES, 2018). Based on these guidelines, McGill et al. (2019) organized problems and their solutions from the perspectives of reproducibility and replication for the maturity of the field of computer education. Learning Analytics research provides findings with contextual information and offers ideas for analytical methods for extracting behaviors. However, those findings and methods, especially the analytical process, have not been frequently shared and reused in other studies owing to technical constraints due to runnable necessities and context dependencies (Lebis, Lefevre, Luengo, & Guin, 2018). Lebis et al. (2018) proposed a framework called "Capitalization of Analysis Processes," and cited data standardization as related work.

If many people use the same system and do some analysis on that log, the same preprocessing and basic analyses are likely to be performed as the data structure is the same. For example, OpenLA provides common processing as an open-source library for an e-Book log analysis (Murata, Minematsu, & Shimada, 2020). Even if the system is not exactly the same, the general learning tool, such as e-Book or LMSs, have some common actions. Hence, xAPI Profiles have been proposed so that the data of the learning behavior that emerges there can be described in a common format (Usalearning). In other words, for systems such as e-Books or LMSs, there have been several proposals for applying methods and findings for extracting behavior from data from the original system or context to others.

However, Learning Analytics research does not always target behaviors on common systems that many people use, such as e-Books or LMSs. For example, the first author has focused on students' peer evaluation activities and has researched Evaluation Behavior Analysis (EBA), which visualizes students' behaviors during peer evaluation (Horikoshi & Tamura, 2021). EBA uses evaluation process data such as which evaluation items were evaluated by which student in what

order and with how much time. As a result, for example, we found students who evaluate from the evaluation items, or who evaluate in the order of evaluation items in a short time, and it has been clarified that the evaluation behavior varies depending on the student.

This study uses the EBA as an example of an analytical method on systems that are not in general use, unlike e-Books or LMSs. The objective of this study is to clarify the possibilities and challenges in applying a behavioral analysis method to other system contexts. To achieve this objective, we applied EBA to two datasets from different systems, using a situation where one of the authors involved in two different systems targeted the same behavior owing to the transfer of the institution. The research questions are: (1) whether EBA method can be applied to different datasets generated in different peer evaluation activity contexts and systems, and (2) whether the findings in the previous EBA study can be reproduced. We expect that this research will contribute to expanding the reuse of analysis methods for small-scale systems such as EBA, which have not usually been reused, and further development in Learning Analytics fields.

2. Methods

2.1 Research Design

To answer the research questions, we designed this research as shown in Figure 1. We compared the applicability of the method and reproducibility of findings in two different studies: Study A (Horikoshi & Tamura, 2021), which used the original system for EBA, and Study B (Liang, Majumdar & Ogata, 2022), which is a study this paper attempts to apply EBA. This study focuses on four basic steps in EBA: data preprocessing, behavior visualization, feature extraction and distribution, and correlation.

The structure of this paper is as follows. The following part of Section 2 summarizes the differences between the systems and activities in Studies A and B. In section 3, four basic steps in EBA are executed using the data from each system. The results are compared from the perspectives of the applicability of the method and reproducibility of the findings. Where the results in Study B did not match those in Study A (the original), we clarified whether it was because of the difference in the data format generated by the system or in the activity. Section 4 summarizes the results and discusses the challenges and solutions in applying an analysis method to the data of another system. The new data are only for Study B. The results in Study A for comparison were already reported in Horikoshi and Tamura (2021) and others.

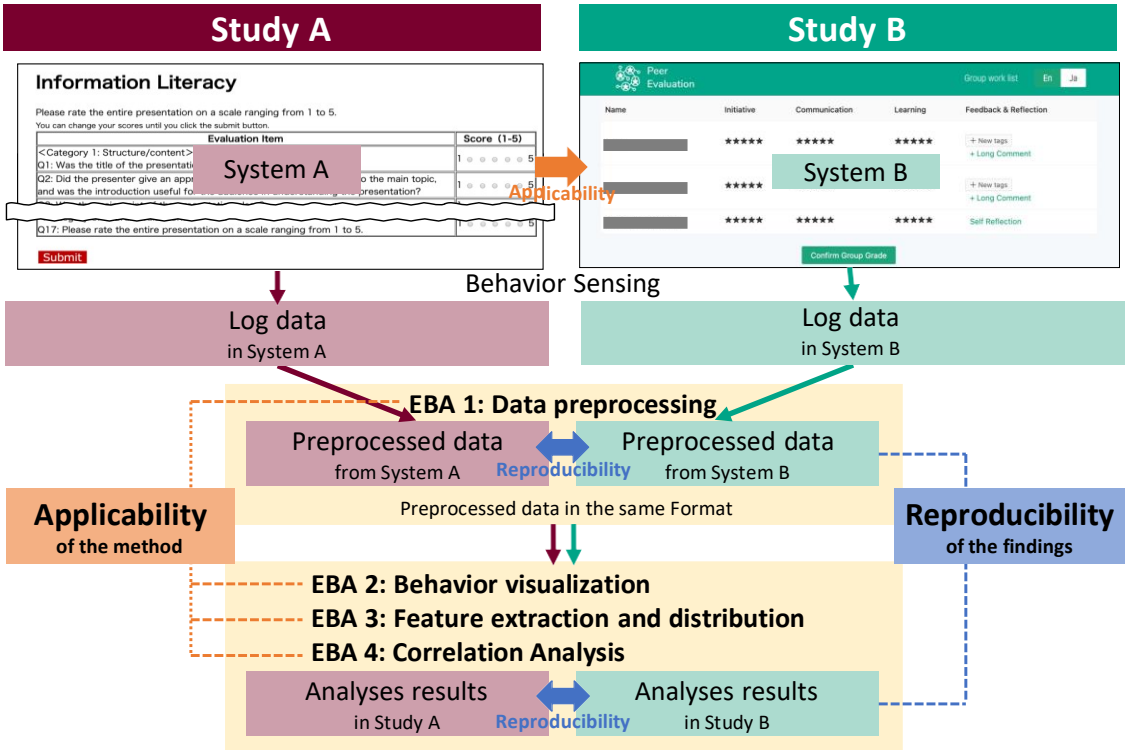


Figure 1. Research Design.

2.2 *Two Students' Peer Evaluation Studies Considered in this Paper*

We considered two contexts of peer evaluation activity. Here, we highlight the system, the peer evaluation activity, and the data collected in both contexts.

2.2.1 *Study A: The Original Evaluation Behavior Analysis (EBA) Study*

(1) System

Figure 2 presents the peer evaluation tool used in Study A. It is developed as a Web-based form to detect students' evaluation process data. The reviewer selects the score for each item and clicks the respective radio button to evaluate the evaluation target student. The reviewer can change their scores at any point in time before clicking the submit button. Evaluation process logs are sent to a server with a timestamp for when the reviewer clicks both the submit and radio buttons.

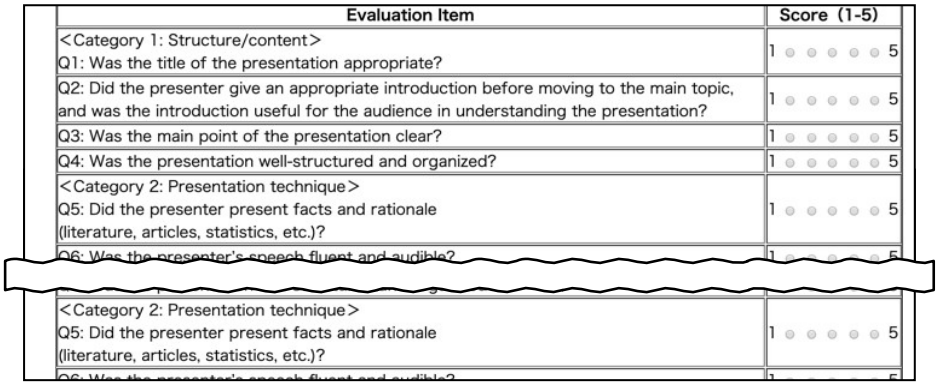


Figure 2. The Peer Evaluation Tool used in Study A (System A).

(2) Activity and Procedure

The target activity was one of the presentation classes at the end of a course at a university in Japan. The presentations were opinion speeches and the evaluations focused on the formal aspects of presentations. The students were divided into groups, each of which was given 15 minutes, which comprised a 10-minute presentation and a 4-minute question and answer (Q&A) session. Peer Assessments were conducted using the Peer Evaluation tool. The log data of the students' evaluations were acquired. The class was 90 minutes long. The students were instructed to submit an evaluation form during class.

(3) Scope of the Evaluation Behavior in Study A

The evaluation process log data used in Study A can only visualize "when the reviewer clicks a button on the evaluation form." As for the process that was not visualized in the log, the research field of questionnaire response behavior helped organize the cognitive process to provide the final response (Horikoshi & Tamura, 2021). This was called the "Cognitive Response Process Model" (Dillman, Smyth, & Christian 2009; Olson & Parkhurst 2013; Tourangeau, Rips, & Rasinski 2000). The components of this model differed slightly across the studies. Study A used the version created by Olson and Parkhurst (2013). Figure 3 illustrates the response process that operated when the respondents answer a form. Steps 1 to 5 and their descriptions were taken from the "Cognitive response process model" in Olson and Parkhurst (2013), and we added the component "Answering" as the sixth step.

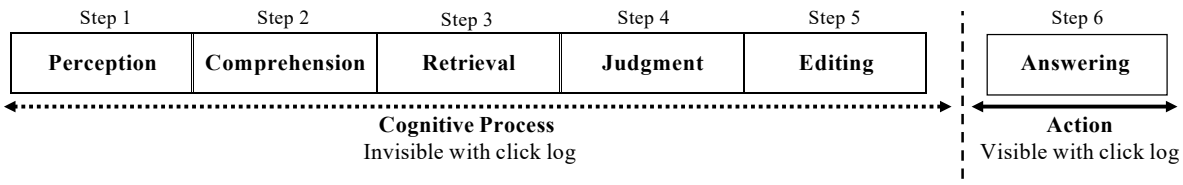


Figure 3. Response Process (Adapted from Fig.4 in Horikoshi and Tamura, 2021)

Only the sixth step can be visualized from the click log. However, as it is assumed that steps 1 to 5 are performed between each click, it is possible to infer the cognitive process until finally pressing submit. For example, if a student clicked from the top of the evaluation items, the cognitive process in steps 1 to 5 may be performed according to the evaluation items. On the other hand, if the student did not click in the order of the evaluation items, the cognitive process may have been performed according to what was noticed in the performance to be evaluated. If the time between clicks were extremely short, the cognitive process may have not taken enough time.

2.2.2 Study B: Where Evaluation Behavior Analysis (EBA) was Applied in this Paper

(1) System

Figure 4 shows the peer evaluation tool used in Study B. This system is one of the functions of the Group Learning Orchestration Based on Evidence (GLOBE: Liang, Toyokawa, Nakanishi, Majumdar, & Ogata, 2021) which constitutes the Learning and Evidence Analytics Framework (LEAF). In this study, we focused on the interaction log generated by the peer evaluation tool. This tool only sends the log when the submission button is clicked, and unlike System A, it does not send the log when the radio buttons (star-shaped) are clicked. In System A, the students had to press the submit button once for each evaluation target. The evaluation items were listed vertically, and the corresponding points were placed next to each item. In System B, evaluation items and targets were lined up horizontally and vertically, respectively. The evaluations of multiple targets were sent together in one submission. As with System A, students can change the scores as many times as they want until they press the submit button, but the process log is not recorded in System B. The log is sent only upon pressing the submit button.

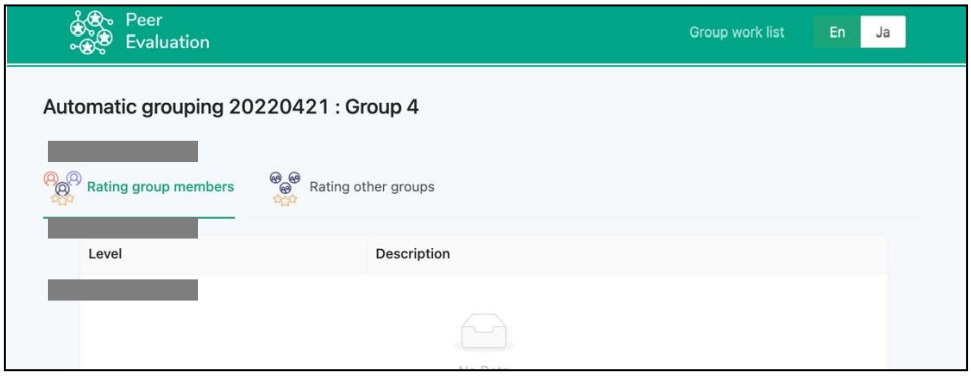


Figure 4. The Peer Evaluation Tool used in Study B (System B).

(2) Activity and Procedure

The target activity was a mini-presentation for sharing the status of assignments within a group at a university in Japan, which was different from Study A. The mini-presentation was instructed to include three elements pertaining to the assignment, and the students evaluated each element in the range of 1 to 5. The evaluation criteria for scoring each evaluation item were provided to the students. The instruction was to give each presentation in about 5 minutes, but as it was a group presentation, it was not possible to know how much time they actually took. The class length was about 90 minutes, of which the target activity was about 30 minutes, and the students were supposed to submit their evaluation by the end of the class.

3. Results

3.1 Analysis 1: Data Preprocessing

Comparing the log data of Systems A and B, the column names and their order were, not surprisingly, different between them. However, although the column names such as the timestamp when the log was stored, the reviewer, the evaluated target, the evaluation item number, and the

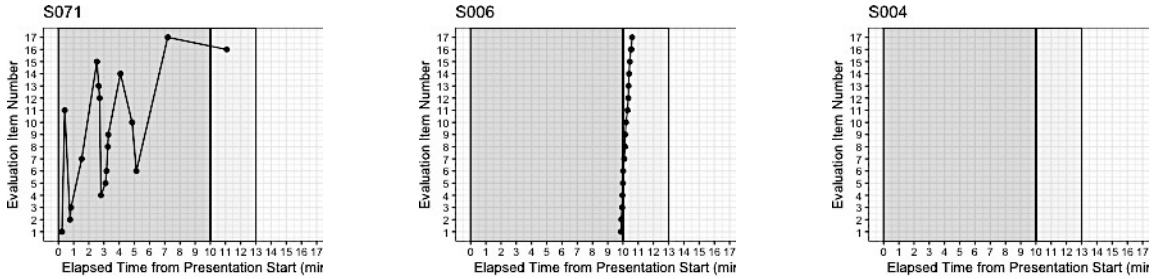
score were different, the contents of the stored information were almost the same. The timestamps in System A differed for each record as the system sent a log each time the radio button was clicked. In System B, the timestamps were the same for multiple lines because the logs were sent together when the submit button was clicked. Nevertheless, even in System B, the submit button can be pressed as many times as necessary, so some students submitted again with some changes.

In System A, these data were preprocessed and shaped into a form of when and who gave what points for what item to whom. Along with shaping the column, the record whose score had not changed was deleted from the previous log, leaving only the data of the change action and its timing behind. We tried to format the System B data into the same shape as the preprocessed System A data. As System B only comprised submission clicks, the number of records was small, and the timestamps of the lines extracted from the same submission became the same. Nevertheless, we could obtain the data in the same format as that in System A.

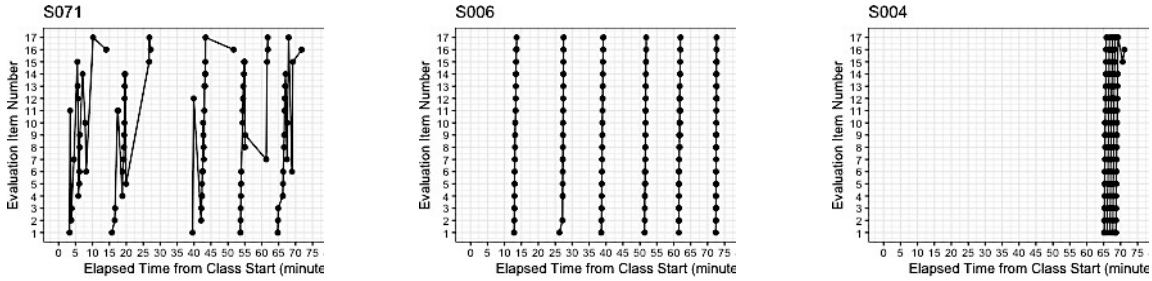
From this, it became clear that if the final goal of the preprocessing was to obtain the data form in basic experience expression of "when-who-did-what", even datasets with completely different columns from completely different systems can be shaped in the same shape. In other words, it can be said that the method of data preprocessing in EBA was applicable.

3.2 Analysis 2: Behavior Visualization

Figures 5 and 6 visualize the evaluation behavior from the data prepared with the preprocessing in Analysis 1. The vertical axis of the plots is the evaluation item number, and the horizontal axes are the elapsed time from the start time of the presentation, class, and activity respectively.



(a) Evaluation Behaviors for One Target



(b) Evaluation Behaviors for All Targets

Figure 5. Evaluation Behaviors from Study A's data. (Source: Fig.7 in Horikoshi and Tamura, 2021)

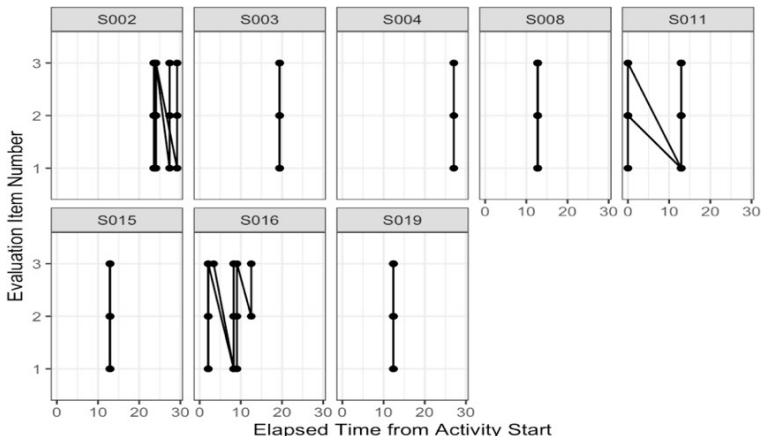


Figure 6. Evaluation Behaviors from Study B's data.

The evaluation behaviors in Study A, shown in Figure 5, can be divided into behaviors for one evaluation target (Figure 5 (a)), and for all evaluation targets performed in that class (Figure 5 (b)). Figure 6 shows the evaluation behaviors visualized from Study B's data. In System B, the radio button-click logs were not sent but the evaluations of multiple evaluation targets were sent together in one submission. Therefore, unlike Study A, it was not possible to separate the behavior for one evaluation target and that for all targets performed in that class.

First, it was possible to visualize evaluation behavior by placing Elapsed Time on the x-axis and Evaluation Item Number on the y-axis. In addition, it was also possible to visualize the timing (some evaluated at the beginning of the activity and some at the end) and modification of the evaluation. On the other hand, the evaluation timing for each evaluation target, evaluation item, and the order of evaluation could not be visualized. Moreover, modifications in evaluation without clicking the submit button could not be visualized. All of these differences are due to the difference in data transmission timing design that System B did not send the log when the radio button was clicked, and the evaluations of multiple evaluation targets were sent together in one submission. In other words, the applicability of the method was not sufficient for behavior visualization, and this was due to differences in data granularity.

In this way, many characteristic behaviors could not even be visualized, and therefore, findings based on behavior visualization were limited. However, the timing of the evaluation was able to be visualized, and the behaviors found in Study A, such as some submitted at the beginning and others at the end, were also visualized in Study B. In addition, the behavior that some students changed their evaluation was reproduced. However, only the submission level was visualized in Study B while this was the click level in the original method. Therefore, there might be more students who made modifications at the click level actually.

3.3 Analysis 3: Feature Extraction and Distribution

In Study A, the six feature variables shown in Table 1 were extracted and used as indicators to quantitatively capture the characteristics of evaluation behavior. Some of the definitions of the feature variables shown in Table 1 could not apply to Study B as they were, so the underlined parts were replaced, as shown in parentheses.

Table 1. Definition of Feature Variables of Evaluation Behaviors Proposed in Study A

Feature Variables	Definition
Evaluation Time (ET)	Time difference between clicking the radio button of the first evaluation item and the last evaluation item. (Replaced it with "the first record" and "the last record")
Click Count (CC)	Total number of times the radio buttons for the evaluation items were clicked. (Replaced it with "the record")

Mean of the Score (sM)	Average score for all the evaluation items scored by the reviewer.
SD of the Score (sSD)	Standard deviation for the scores of all evaluation items scored by the reviewer.
Mean of the Timestamp (tM)	Average elapsed time since the start of the presentation. (Replaced it with "the first record")
SD of the Timestamp (tSD)	Standard deviation of the timestamps for all evaluations.

Figure 7 compares the distribution of the feature variables extracted from the data in Studies A and B. The vertical axis of the plot is the value of Feature Variables; the range of which is different depending on each Feature Variable. As mentioned above, some of the definitions of feature variables could not apply to Study B as they were. ET and CC were because of system differences, and tM was because of activity differences. However, in fact, the program was applicable with almost no changes or modifications. This means that the main problem was that the expressions of the definitions of feature variables were specialized to the context (system or activity) in Study A, and were not sufficiently generalized.

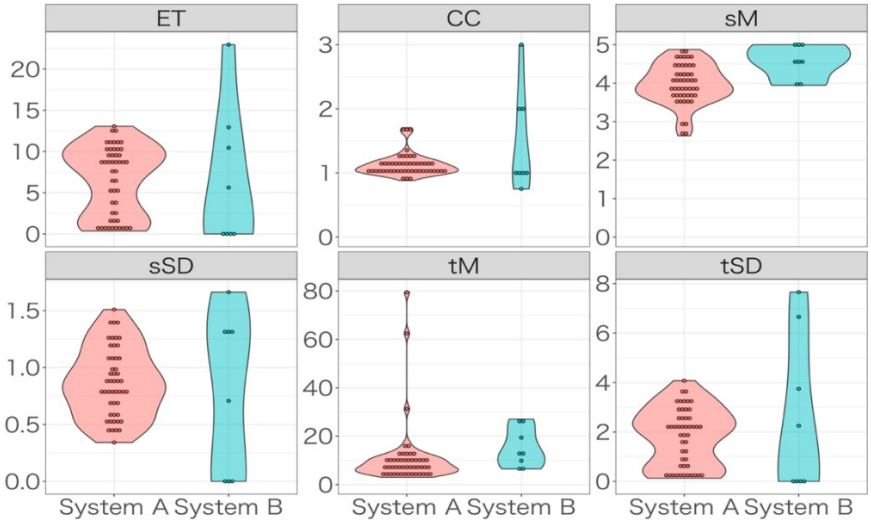


Figure 7. Distribution of Extracted Feature Variables.

All feature variables could be extracted after the above adjustment. However, the extracted variables did not necessarily express the same concept in Studies A and B. For example, score-related features (sM and sSD) were assumed to represent almost the same concept in Studies A and B. On the other hand, other variables related to time and clicks may have been significantly affected by differences in the timing of data transmission. This is because the System B log data did not record the action before the submission button was clicked; for example, ET and tSD will be 0 if there is only one submission, and the value of tM does not always reflect the timing of the evaluation. Also, for CC, the modifications before submission were not captured. Thus, the value may have been a little smaller than the original definition.

As for the visualization of the distribution of the extracted features, CC, sM, and tM showed almost the same distribution in Studies A and B, whereas the distributions in ET, sSD, and tSD were different. Most CCs were one time per evaluation item in both data. Also, regarding sSM, many students tended to use higher scores and sSM became high in both studies. For tM, most tMs were within presentation activity time. On the other hand, ET and tSD had two peaks (Speeder and non-speeder) in Study A, whereas most of them were 0 in Study B. This is because, as mentioned above, if there is only one submission, then ET becomes 0, and this is due to the difference in the system. For sSD, it was 0 to 1.5 and two peaks were observed in Study B while it was around 0.5 to 1.5 in Study A. Regarding sSD,

the extracted variables themselves were considered to express the same concept in Study A and B, but the distribution was different. The interpretation of the reason for this difference is that Study B clearly stated the criteria for each score, and for these reasons, it was easy for students to choose one.

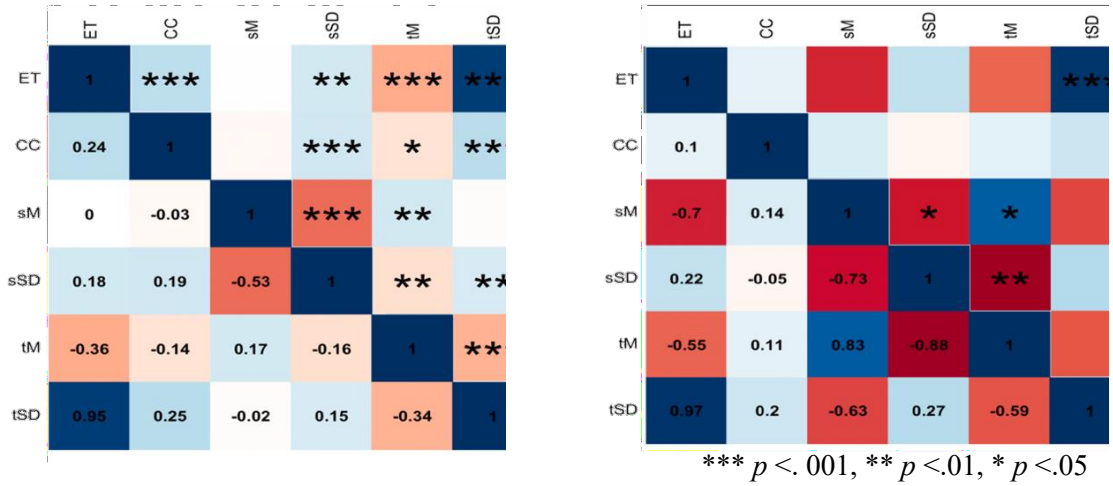
Thus, only three out of the six feature variables reproduced distribution tendencies. Two findings were not reproduced because of differences in data and one was because of differences in activity.

3.4 Analysis 4: Feature Correlation

As the last analysis, correlation analyses between the features were performed to compare the characteristics of the features extracted. In Study A, the correlation shown in Figure 8 (a) was found among the six types of feature variables. For this result, Study A interpreted that there was some common factor, and the factor made ET, CC, sSD, and tSD decreased, and sM and tM increased. This factor was interpreted as the students' motivational state in Study A. In other words, if the students' motivation for evaluation was low, the evaluation might be completed in a short time (small ET, tSD), not be changed (small CC), the same score might be used many times (small sSD), be given many full marks (large sM), and many evaluations might be made at the end of the class (large tM).

This interpretation was based on the findings of the answering behavior Web Survey research. In the Web Survey research field, response time has been used frequently as an indicator of the quality of a survey. According to Yan and Tourangeau (2008, p.64), "respondents tended to answer more quickly as they got closer to the end of the questionnaire." Therefore, we hypothesized that the proposed feature variables could also be influenced by repeatedly conducting peer evaluations in a single class.

As the feature variables were extracted in the same format, there was no problem in applying the correlation analysis. However, there were few significant correlations as the number of subjects and records in Study B were limited. Focusing on the value of the correlation coefficient, which reflects the tendency of the distribution of variables, the results in Studies A and B matched in 12 out of 15 pairs in the positive or negative direction. All three unmatched pairs were related to CC, perhaps because the data in Study B were not click-level ones, and therefore had smaller values than the other features.



*** $p <.001$, ** $p <.01$, * $p <.05$

(a) Study A (b) Study B

Figure 8. Correlation among Feature Variables.

4. Discussion

The research questions were: (1) whether EBA method can be applied to different datasets generated in different peer evaluation activity contexts and systems, and (2) whether the findings

in the previous EBA study can be reproduced. To answer this question, we conducted four analyses from EBA, and compared the two studies conducted with two different systems that collected students' evaluation behavior data. In this section, we summarize the results and discuss the challenges and solutions in applying an analysis method to the data of another system.

4.1 Applicability of the Method

EBA 1 - Data Preprocessing: Even though the original data structures were completely different, the final structure became the same. This shows that if the final goal of the preprocessing was to obtain the data in basic experience expression of "when-who-did what," it is possible to convert datasets with completely different columns of completely different systems into the same shape. In other words, it can be said that the method of data preprocessing in EBA was applicable.

EBA 2 - Behavior visualization: It was possible to visualize the basic evaluation behavior plot with the same axes in both datasets. From these plots, the timing and modification behavior of the evaluation was visualized. The visualization of behavior was significantly affected by the difference in the design of the data transmission timing, and many behaviors such as the order of evaluation and change in evaluation without clicking the submit button could not be visualized. In other words, the applicability of the method was not sufficient for behavior visualization, and this was mainly due to differences in data granularity.

EBA 3 - Feature Extraction and distribution: Some of the definitions of feature variables could not apply to Study B as they were, because of both differences in the systems and activities. Nevertheless, the program was applicable with almost no change or modification. This means that the main problem was that the expressions of the definitions of feature variables were specialized to the context (system or activity) in Study A, and it was not sufficiently generalized. In addition, all feature variables could be extracted with the adjustment. However, the extracted variables did not necessarily express the same concept in Studies A and B. This was mainly caused by system differences.

***EBA 4 - Feature Correlation*:** There was no problem applying the correlation analysis, as the feature variables were extracted in the same format.

4.2 Reproducibility of the Findings

EBA 2 - Behavior Visualization: Many characteristic behaviors could not even be visualized. Therefore, findings based on behavior visualization were limited. However, the timing of the evaluation was visualized, and the behaviors found in Study A (e.g., some submitted at the beginning and others at the end, or some changed their evaluation) were reproduced in Study B.

EBA 3 - Feature Extraction and Distribution: Only three out of six feature variables reproduced distribution tendencies. Two findings were not reproduced because of differences in data and one was because of differences in activity.

EBA 4 - Feature Correlation: The results of Studies A and B matched in 12 out of 15 pairs in terms of the direction of correlation (positive or negative). All three unmatched pairs were related to CC, perhaps because the data in Study B were not click-level, in other words, due to system differences.

4.3 Challenges and Solutions

First, regarding the preprocessing of data, the data structure is usually completely different depending on the system, but it became clear that if preprocessing was performed appropriately and the data were brought into the same format, a considerable part of the later analysis method can be applied. This was considered possible because the result of the preprocessing in the EBA

was the structure of "when-who-did what," which consisted of the components of the description of the basic learning experience. This structure is also similar to the components of xAPI. However, preprocessing for the new dataset was complex and took a lot of time. One of the possible solutions for this is to standardize the data format, but it is unrealistic to standardize the data structure in the database for analysis. Because the data to be analyzed are not only stored just as log data but also usually working as a part of the system to operate. Therefore, storing additional data for analysis in the "when-who-did-what" format using xAPI or other standardized formats into Learning Record Store (LRS) is considered a practical solution for standardizing and reusing preprocessing.

It was also clarified that there are cases where the method can be applied but the findings cannot be reproduced because of the difference in activities. One solution for this problem is to separate the applicability of the method and reproducibility of the findings and the influence of data and activity as done in this paper. However, it was possible to discuss the effects of data and activity in detail because one of the authors was involved in both studies, but such cases are rare. Therefore, it is necessary to disclose data and activity information in papers and leave it thus so that similar discussions can be held even if the same researcher does not participate in subsequent studies.

Finally, the factor that had the most significant impact on both the applicability of the method and reproducibility of the findings in this analysis was the granularity of the stored data. No matter how much pre-processing or interpretation was devised, it was impossible to create unstored information that was supposed to be extracted from process data. The click-level process log is necessary for the behavior analysis related to the cognitive process shown in Figure 3.

As Learning Analytics shifts from the research level to the practical level in the future, it will become more common to apply the same methods and knowledge to log data from different systems. From the results clarified in this paper, it is necessary to acquire similar log data in similar granularity in order to increase the applicability of the method and the reproducibility of the findings. For this reason, when designing learning log data sensing systems, it is important to sense various types of data in as fine a granularity as possible, convert them into a standardized format, and store them with context information. This requires a more detailed discussion on the standardization of data format and stored information.

5. Conclusion

In this paper, we applied Evaluation Behavior Analysis (EBA) to two datasets from different systems and verified the applicability of the method. Not all methods were able to be applied as the data and activity were different, and the results showed that some findings were not reproduced. However, the reproducibility was far higher than we expected.

One limitation of this paper is that the datasets from the two compared systems differed in terms of data granularity and activity. Thus, in this research design, even though the applicability of the method had not been verified, the reproducibility of the findings was discussed. Therefore, in our future work, we would like to discuss the reproducibility of findings again with the same data granularity and in similar activity contexts, while ensuring that the applicability of the method remains satisfied. This time, we only used evaluation behavior data, which were used in Study A. However, the greatest advantage of System B's data is that it can be analyzed by combining behavioral data and other learning logs on the same LEAF platform. Thus, we would like to build further on this study, which was merely a confirmation of applicability and reproducibility, and clarify whether the new interpretation of behavior or knowledge that leads to class improvement can be obtained by combining the evaluation behavior data with other learning logs. Though this study is exploratory and has some limitations, we believe the results present possibilities, difficulties, and solutions with respect to applying similar behavior concepts and methods of behavior analysis to other systems.

Acknowledgments

This study was supported by the following grants: NEDO JPNP20006, JSPS KAKENHI 20H01722

References

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). Internet, Mail and Mixed-Mode Surveys: The Tailored Design Method (3rd edition). Hoboken, NJ: Wiley and Sons, Inc.

Horikoshi, I., & Tamura, Y. (2021). How Do Students Evaluate Each Other during Peer Assessments? An Analysis Using "Evaluation Behavior" Log Data. *Educational Technology Research*, *43*(1), 3-21.

Lebis, A., Lefevre, M., Luengo, V., & Guin, N. (2018, March). Capitalisation of analysis processes: Enabling reproducibility, openness and adaptability thanks to narration. In Proceedings of the 8th international conference on learning analytics and knowledge (pp. 245-254).

Lang, C., Siemens, G., Wise, A., & Gasevic, D. (Eds.). (2022). Handbook of learning analytics – Second edition. New York: SOLAR, Society for Learning Analytics and Research.

Liang, C., Toyokawa, Y., Nakanishi, T., Majumdar, R., & Ogata, H. (2021, August). Supporting Peer Evaluation in a Data-Driven Group Learning Environment. In the proceedings of International Conference on Collaboration Technologies and Social Computing (pp. 93-100). Cham: Springer.

Liang, C., Majumdar, R., & Ogata, H. (2022, June). Continuous Data-Driven Group Learning Support: Case Study of an Asynchronous Online Course, In Proceedings of the 15th International Conference on Computer-Supported Collaborative Learning - CSCL 2022 (pp. 547-548).

McGill, M. M. (2019, November). Discovering empirically-based best practices in computing education through replication, reproducibility, and meta-analysis studies. *In Proceedings of the 19th Koli Calling International Conference on Computing Education Research* (pp. 1-5).

Murata, R., Minematsu, T., & Shimada, A. (2020, November). OpenLA: Library for Efficient E-book Log Analysis and Accelerating Learning Analytics. In 28th International Conference on Computers in Education, ICCE 2020 (pp. 301-306). Asia-Pacific Society for Computers in Education.

National Science Foundation and the Institute of Education Sciences. (2018). Companion Guidelines on Replication & Reproducibility in Education Research: A Supplement to the Common Guidelines for Education Research and Development.
Retrieved from https://www. nsf. gov/pubs/2019/nsf19022/nsf19022. pdf.

Olson, K., & Parkhurst, B. (2013). Collecting paradata for measurement error evaluations. In Kreuter, F. (Eds.), Improving Surveys with Paradata, 43-72. Hoboken, NJ: Wiley

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The Psychology of Survey Response. Cambridge, UK: Cambridge University Press

USAlearning. (n.d.). Actionable Data Book (ADB) Profile
Retrieved from https://profiles.usalearning.net/profile/f24c8d44-4937-44a6-bef7-f5435f57492f

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Association Analysis on Open-Ended Concept Maps using Data Mining

Didik Dwi PRASETYA[a], Setiadi Cahyono PUTRO[a], Muhammad ASHAR[a], Saida ULFA[b], & Tsukasa HIRASHIMA[c]

[a]*Department of Electrical Engineering, Universitas Negeri Malang, Indonesia*
[b]*Department of Educational Technology, Universitas Negeri Malang, Indonesia*
[c]*Department of Information Engineering, Hiroshima University, Japan*
*didikdwi@um.ac.id

Abstract: An open-ended concept map is an appealing approach that allows learners to add concepts and links freely, representing their understanding. This technique reveals differences among students and accurately captures their knowledge structure. However, manually evaluating an open-ended map is challenging and time-consuming, particularly in a large classroom, and involves many propositions. Several works tried to use data mining techniques to generate concept maps from the source text rather than analyzing the knowledge structure. The previous study employed association analysis on provided concept maps. This study focused on association analysis of open-ended concept mapping in revealing hidden students' understanding. A total of 20 open-ended concept maps were involved in this study. The findings suggested that data mining techniques could quickly disclose necessary information on open-ended concept maps. The study identified an exciting topic for further research in educational data mining, particularly concerning knowledge structure.

Keywords: Concept map, open-ended, association analysis, data mining, educational data mining

1. Introduction

Concept maps are beneficial graphical knowledge representation tools to acquire an overview of an individual's knowledge domain. The concept map is well-known for accurately representing individual conceptual knowledge in a given topic. It depicts different ideas at nodes with linking words connecting two concepts describing their relationship and forming a proposition (Cañas and Novak, 2014). Propositions are the minor semantic units and are fundamental elements of a map. Therefore, it can be perceived as a crucial knowledge component (Aleven et al., 2016). The proposition represents a unit's declarative knowledge in forming meaningful statements.

Concept map construction can be divided into open-ended and closed-ended maps (Hirashima, 2019; Prasetya et al., 2019). In an open-ended approach, learners can add concepts and link terms in their diagrams to represent their understanding. The knowledge structure in the open-ended concept map can be analogized as unstructured or indeterministic information. On the other hand, a closed-ended concept mapping style has finite components. Thus, learners must use the defined concepts and links in building their concept maps. The closed-ended concept map allows automatic assessment, provides quick feedback (Hirashima, 2019; Pailai et al., 2017), and motivates learners to reach maximum learning comprehension (Prasetya et al., 2021).

The open-ended approach is more challenging in evaluation than the closed-ended method. So far, the technique used to evaluate students' open-ended concept maps might be using rubrics and assessed by an expert (Taricani and Clariana, 2006). However, rubrics have limitations, especially in the case of various open-ended concept maps. The rubric is time-consuming and cannot comprehensively evaluate students' knowledge structure (Yoo and Cho, 2012). The diversity of concept maps makes it difficult for teachers to determine certain students' mental model patterns. However, judging their knowledge structure is an excellent choice. One approach that can be proposed to manage concept maps automatically is data mining techniques.

Data mining automatically extracts new and vital knowledge hidden in large datasets (Yoo and Cho, 2012). Data mining aims to uncover the databases' latent, previously undisclosed trends and

patterns. Furthermore, the data mining process generates valuable information needed to support decision-making. Data mining has been proven to provide many benefits and is widely used in various fields, one of which is education and is called Educational Data Mining (EDM). EDM is a novel multidisciplinary research field focusing on developing methods for analyzing data generated in educational settings (Romero and Ventura, 2020). EDM is concerned with developing tools for exploring the unique types of data found in educational settings and, by doing so, better understanding students and the environments in which they learn (Baker, 2019).

Until recently, many studies have analyzed the impact of data mining in managing concept maps. For example, Villalon and Calvo (2008) revealed a task of automatically generating concept maps from a text known as Concept Map Mining (CMM). CMM is an automatic or semi-automatic creation of concept maps from text documents. The same concept was also used for the automatic generation of concept maps based on text analysis and association rule mining (Shao et al., 2020). For a similar purpose, a semi-automatic method for constructing a concept map of reading material in English for Foreign Language (EFL) was proposed and yielded positive effects (Pinandito et al., 2021). However, CMM aims to produce a concept map that provides accurate information about students' understanding instead of analyzing the concept map expressed by the students themselves.

The present study utilized data mining techniques to process open-ended concept maps and analyze students' understanding of learning topics. Another study similar to this study was conducted by Yoo and Cho (2012). However, the previous approach was intended for pre-defined concept maps rather than open-ended ones that refer to students' understanding. This study used association rules to reveal valuable hidden information in concept maps. In particular, the following research question guides the study: the extent to which data mining techniques on the concept map could quickly represent students' understanding of a Database topic.

2. Methods

2.1 Research Design

The term mining concept maps in this study refer to using data mining techniques on open-ended concept maps to identify students' understanding quickly. In particular, the association analysis methodology was applied to the concept maps data to obtain valuable hidden information. Since the interpretation of learners' open-ended concept maps cannot be made as easily as the closed method, this technique will make a non-trivial contribution to the learning process.

The association rule is a proper data mining technique for determining the extent of correlation between variables in a database (Ait-Mlouk et al., 2017). It aims to discover regularity in the data by identifying specific patterns. Association rules' main activity is finding essential relationships between items in each transaction. The relationship could reveal the extent of the strength of a rule in the association.

2.2 Data Collection

This study involved 20 open-ended concept maps made by 20 second-year university students majoring in Informatics Engineering. The concept maps were created using the scratch-build concept mapping tool by referring to the "Basisdata Relasional" (Database Relational) topic in the "Basisdata 1" (Database 1) course that was delivered in Indonesia.

The concept map extraction produced 314 concepts with an average of 15.70 (SD = 4.34) that describe the original learners' ideas regarding the Basisdata Relasional topic. Concepts represent key ideas or opinions expressed by learners related to the learning materials. The learners' concepts produced 290 propositions with an average of 14.50 (SD = 4.37).

2.3 Preprocessing

As in general data mining activities, preprocessing was carried out before conducting the main activity. Data preprocessing is an essential task and critical step in data mining (Angiani et al., 2016). It includes the preparation and transformation of data into an appropriate form. The primary purpose of

preprocessing is to normalize data, reduce data size, discover the relations between data, remove outliers, and extract features for data (Alasadi and Baya, 2017). Preprocessing was applied to the pairs of concepts that form the propositions on the concept maps. It transforms the raw ideas and propositions data into a valuable and efficient format. This stage performed the initial processing of the set of propositions on students' concept maps. Initial processing includes reading and parsing concept maps from the database, data cleaning, tokenization, and term weighting.

2.4 Mining Open-Ended Concept Maps

The open-ended mining concept maps process students' knowledge structures to reveal their understanding of related learning materials. The main steps of the mining activity are shown in Figure 1. Each concept map represents a student's knowledge structure. They were treated as a single document and extracted for subsequent operations. There were two main recognition operations: concept recognition and relationship recognition. Based on the recognition operations that have been carried out, association analysis could be performed to reveal common concepts and propositions used by students on their concept maps.

Concepts recognition is the process of finding potential candidates for a student's key ideas. Each concept defined by learners is identified and retrieved from the open-ended database system. Frequent one itemset was applied to determine the emergence of ideas in each concept map. Since the concept is a unique element in the individual concept map, analysis was conducted in the group's structure. Furthermore, the concept recognition operation results could be confirmed to identify what students understand through the knowledge structure that has been formed. The support value for a one itemset X is obtained by dividing the number of transactions containing X by the total number of transactions and is defined as support (X) = PX / P total.

Relationships are propositions that describe the relationship between two concepts that are connected by a link label. Proposition represents the smallest meaningful component on concept maps. It has an important role in revealing students' understanding related to the knowledge they have acquired. In the relationships recognition phase, each concept that is fully connected to another concept is identified as a sub-concept map. The relationships are represented in the form of association rules or sets of frequent items. For example, a frequent itemset, {database, table} indicates that there was a strong relationship between database and table terms in the database material.

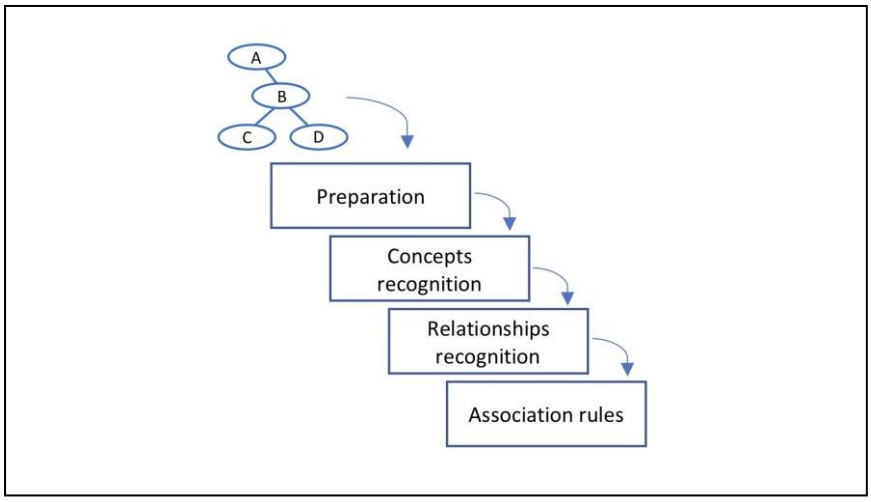


Figure 1. Mining concept map activities

Association rules uncover intriguing connections or relationships among a massive number of data items (Karabatak et al., 2011). It displays attributes value circumstances that frequently occur together in a given dataset. The main purpose of association rules is to recognize the patterns of co-occurrence of attributes in a database. The quality of association analysis results is often measured using support and confidence parameters (Xu, 2016). After receiving a frequent high value, furthermore, the confidence value was examined using associative rules. The present study employed the Apriori

algorithm to analyze frequent association concept map mining. A minimum frequency threshold of 30% was used to analyze concepts, propositions, and association rules.

3. Results and Discussion

Twenty learners' concept maps were extracted from the database and generated 314 unique concepts expressing various ideas. Analysis of pattern creation was carried out on the collection of concepts by using a frequent one itemset. The purpose of this analysis was to identify the ideas that learners most express. Minimum support of 30% was set to reveal frequent concept patterns. The concept "basis data relasional/*relational database*" was the idea that most of the learners sparked. A total of 18 learners (90%) wrote it on their concept maps. Learners' concepts with a minimum support of 50% were: "atribut/*attribute*" (85%), "relasi/*relation*" (85%), "tupel/*tuple*" (85%), "domain/*domain*" (80%), "kardinalitas/*cardinality*" (70%), "tabel dua dimensi/*2 dimension table*" (65%), "derajat/*degree*" (65%), and "istilah/*terms*" (50%). Other frequent itemset of concepts were: "super key/*super key*" (45%), "keuntungan/*benefit*" (40%), "foreign key/*foreign key*" (35%), "candidate key/*candidate key*" (35%), "sederhana/*simple*" (35%), "baris dan kolom/*rows and fields*" (35%), "kolom/*field*" (35%), "primary key/*primary key*" (35%), "alternate key/*alternate key*" (30%), and "baris/*rows*" (30%).

Further analysis was carried out to reveal frequent relationship patterns that expressed learners' understanding based on the propositions they made. Table 1. shows the frequent itemset of relationships with minimum support of 30%. The pattern of frequent relationships looks constant with the findings of the frequent itemset of concepts.

Table 1. *Frequent itemset of relationships*

Relationships	Frequency	Support (%)
basis data relasional – tabel dua dimensi	10	50
basis data relasional – istilah	8	40
istilah – atribut	8	40
istilah – relasi	8	40
istilah – tupel	8	40
istilah – domain	7	35
basis data relasional – keuntungan	7	35
istilah – derajat	6	30
istilah – cardinality	6	30

The frequent sub-concept map analysis was conducted to reveal the possibility of the appearance of new hidden patterns. Frequent sub-concept maps explain the association of proposition a and proposition b on concept maps. By setting minimum support of 30%, three sub-concept map patterns consisting of 2 propositions were obtained. The association {basis data – degree, basis data – domain} having a support value of 35%, {basis data – derajat, basis data – relasi} was 30%, and {basis data - domain, basis data – relasi} was 30%.

Further analysis was carried out to find association rules that were formed by involving the minimum confidence requirements. By setting the 30% support threshold, association rules were obtained, as shown in Table 2. Six association rules defined the best rules with a confidence value of 100% and support of 40%. Furthermore, the second-best rule with a confidence value of 87.5% and support of 35% formed nine association rules. The following best rule produced seven association rules with a confidence value of 85.71% and support of 30%. Associative rules that were formed provide new information for teachers for decision-making.

Table 2. *Association rules of propositions*

Association rules	Support (%)	Confidence (%)
istilah – atribut => istilah – relasi	40	100
istilah – atribut => istilah – tupel	40	100
istilah – relasi => istilah – atribut	40	100

istilah – relasi => istilah – tupel	40	100
istilah – tupel => istilah – relasi	40	100
istilah – tupel => istilah – atribut	40	100
istilah – atribut => istilah – domain	35	87.5
istilah – atribut => basis data relasional – istilah	35	87.5
istilah – relasi => istilah – domain	35	87.5
istilah – relasi => basis data relasional – istilah	35	87.5
istilah – tupel => istilah – domain	35	87.5
istilah – tupel => basis data relasional – istilah	35	87.5
basis data relasional – istilah => istilah – atribut	35	87.5
basis data relasional – istilah => istilah – tupel	35	87.5
basis data relasional – istilah => istilah – relasi	35	87.5
istilah – domain => istilah – derajat	30	85.71
istilah – domain => istilah – cardinality	30	85.71
istilah – domain => basis data relasional – istilah	30	85.71
basis data relasional – keuntungan => istilah – relasi	30	85.71
basis data relasional – keuntungan => istilah – atribut	30	85.71
basis data relasional – keuntungan => istilah – tupel	30	85.71
basis data relasional – keuntungan => basis data relasional – istilah	30	85.71

In line with previous research, data mining techniques have the potential to be beneficial for uncovering hidden information in the educational context (Yoo and Cho, 2012; Adeitan and Salau, 2019). This study found that data mining techniques could reveal students' understanding of an open-ended concept map. In particular, teachers could quickly find out frequent concepts, frequent relationships, and association rules formed by learners through the operation of association rules.

The teacher received many benefits after revealing the frequent itemset of relationships, as it emphasized that relationships are representations of propositions on the concept map (Prasetya et al., 2021). Each occurrence of a proposition was given a weight to reveal its role in the concept map. Finally, through the forming of association rules, teachers could capture the understanding patterns of their students. This information has an important meaning in revealing what they think after gaining new knowledge.

4. Conclusion and Future Work

The present study utilized data mining techniques to find hidden information on open-ended concept maps. This initial study found that data mining techniques have the positive potential for interpreting large sets of open-ended concept maps. The analysis results suggested that the learners' understanding could be revealed through frequent itemset of concepts, frequent itemset of relationships, and patterns of association rules that are formed. Compared to manual evaluation, which is time-consuming and challenging, association analysis finds valuable information shortly. Furthermore, the obtained associative rules could be used to identify the extent to which the learners understand the material that the teacher has delivered.

The present study has several limitations that should be considered. First, the number of concept maps as data sets involved in this experiment was relatively small and needed to be increased to achieve more reliable results. Second, this study employed the Apriori algorithm, which tends to be less efficient in handling large data sets. Therefore, other algorithms such as FP-Growth, Generalized Rules Induction, or hash-based can be tested. In addition, the present study focused on association rules which simply reveal the percentage of proposition combinations. Other operations such as similarity analysis of learners' concept maps and teacher's maps are fascinating to be applied.

Acknowledgments

The first author would like to acknowledge the Universitas Negeri Malang (UM) for supporting this research.

References

Adekitan A. I & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining, Heliyon, 5(2), e01250, 2019.

Ait-Mlouk, F. Gharnati, and T. Agouti. (2017). An improved approach for association rule mining using a multi-criteria decision support system: a case study in road safety. European transport research review, 9(3), 1-13.

Alasadi and W. S. Bhaya, Review of data preprocessing techniques in data mining. Journal of Engineering and Applied Sciences, 12 (16), 4102-4107, 2017.

Aleven, V., McLaren, B. M., Sewall, J., Van Velsen, M., Popescu, O., Demi, S. & Koedinger, K. R. (2016). Example-tracing tutors: Intelligent tutor development for non-programmers, International Journal of Artificial Intelligence in Education, 26(1), pp. 224-269.

Angiani, L. Ferrari, T. Fontanini, P. Fornacciari, E. Iotti, F. Magliani, and S. Manicardi. (2016). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. In KDWeb, 2016.

Baker. (2019). Challenges for the future of educational data mining: The Baker learning analytics prizes. Journal of Educational DataMining,11(1), 1–17.

Cañas, A. J & Novak, J. D. (2014). Concept mapping using CmapTools to enhance meaningful learning, In Knowledge cartography, pp. 23-45. Springer, London.

Hirashima, T. (2019). Reconstructional concept map: automatic Assessment and reciprocal reconstruction, International Journal of Innovation, Creativity and Change, Vol. 5, pp. 669-682.

Karabatak, M. C. Ince, and A. Sengur. (2017). Wavelet domain association rules for efficient texture classification. Applied Soft Computing, 11(1), 32-38, 2011.

Pailai, J., W. Wunnasri, K. Yoshida, Y. Hayashi, and T. Hirashima. (2017). The practical use of Kit-Build concept map on formative assessment. Research and Practice in Technology Enhanced Learning, 12(1), pp. 1-23, 2017.

Pinandito, A., Prasetya, D., Hayashi, Y and Hirashima, T. (2021). Design and development of semi-automatic concept map authoring support tool. Research and Practice in Technology Enhanced Learning, 16(1), 1-19.

Prasetya, D. D., T. Widiyaningtyas, T. Hirashima, and Y. Hayashi. (2019). Reconstruction of Concept Map to Promote Learners' Comprehension on New Knowledge, In The 27th International Conference on Computers in Education (ICCE 2019), volume 1, pages 369-371, 2019.

Prasetya, D. D., T. Hirashima, and Y. Hayashi. (2021) Comparing Two Extended Concept Mapping Approaches to Investigate the Distribution of Students' Achievements, IEICE TRANSACTIONS on Information and System. Vol. 104 (2), pp. 337-340.

Romero and S. Ventura. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355.

Shao, Y. Li, X. Wang, X. Zhao, and Y. Guo. (2020). Research on a new automatic generation algorithm of concept map based on text analysis and association rules mining. Journal of ambient intelligence and humanized computing, 11 (2), 539-551.

Taricani, E. M. and R. B. Clariana. (2006). A technique for automatically scoring open-ended concept maps, Educational Technology Research & Development, 54: pp.65-82.

Villalon and Calvo, R. A. (2008). Concept map mining: A definition and a framework for its evaluation. In 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (Vol. 3, pp. 357-360), IEEE.

Xu, Y. (2016). Research of Association Rules Algorithm in Data Mining, International Journal of Database Theory and Application 2016; Jun 30;9(6): 119-30, 2016.

Yoo, J.S. & M. H. Cho. (2012). Mining Concept Maps to Understand University Students' Learning. International Educational Data Mining Society.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Comparing Short-Term and Long-Term Online Courses Using the Kano Model and Neural Network Language Models

Daniel Moritz MARUTSCHKE[a*] **& Yugo HAYASHI**[b]
[a]*Ritsumeikan University, College of Global Liberal Arts, Osaka, Japan*
[b]*Ritsumeikan University, Department of Comprehensive Psychology, Osaka, Japan*
[*]moritz@fc.ritsumei.ac.jp

Abstract: Evaluating rich data from online courses poses challenges, even with established methodologies. A mixed methods approach can help in addressing research questions. One of the difficulties is to understand users' expectations before taking a course and their consumption experience after completing a course. The focus of this study is on the difference in short-term courses, which can be taken within a day, and long-term courses, which span several weeks or months. This paper compares two online courses for undergraduate students in Japan. The courses were hosted on the same system, one that was complete within 90 minutes (introduction to photography) and one that was conducted for a semester totaling 15 weeks (introduction to programming). Students were surveyed on the same 12 features related to online course satisfaction before and after each course. Textual comments were also gathered. The Kano model from customer satisfaction research was used to perform an ex-ante and ex-post comparative analysis for the 12 features of both short-term and long-term courses. A simple neural network was trained on freeform comments for both courses to create language models (Word2Vec) and compare the findings with the Kano model results.

Keywords: Online course, higher education, e-learning, kano model, machine learning, word2vec, language model

1. Introduction

Online courses and e-learning have gained popularity not least due to the current Covid-19 pandemic. While online-supported and digital education has uncovered shortcomings in many areas throughout 2020 and 2021, a broader challenge in research is the handling of large data collected from online courses. Most online courses that are in form of Massive Open Online Courses (MOOC) and students can enroll in a course and take lessons at their own pace. Famous universities offer courses for free with an option to purchase a certificate of completion, or companies provide fee-charging content for specific skills. Courses can last a whole semester or can be finished in one session.

As learning outcome can differ, depending student expectation and consumption experience and course duration, this research looked at two different courses provided on the same platform, one short-term introduction to photography class and one long-term introduction to programming class. Undergraduate students in Japan from different faculties (psychology for the short-term and liberal arts for the long-term) were surveyed before and after taking the online course. The expectations and consumption experience of both courses are evaluated by the Kano model from customer satisfaction research (Kano et al., 1984) and by creating an AI language model from open-ended questions.

The rest of the paper is structured as follows. Section 2 lists previous works regarding the Kano model in e-learning and language models using Word2Vec. Section 3 describes the research method using the Kano model and Word2Vec. Results and discussion is covered in Section 4. Limitations of this study are considered in Section 5. The paper concludes in Section 6.

2. Previous Work

2.1. Kano Model in E-Learning

While the Kano model is mostly used in business or customer satisfaction research, other previous research in e-learning include learning motivation and exploring multiple hypothesis that tie into student satisfaction (Sun et al., 2008). The Kano method was used to poll students on several e-learning factors in a previous study by Dominici and Palumbo (Dominici and Palumbo, 2013) to construct a theoretical framework of online course design (ex-ante questionnaire without implementation and testing the ex-post consumption experience). Wang et al. investigated a hybrid online/face-to-face course using blended learning (Wang et al., 2014).

2.2. Uses of Language Models Trained by Word2Vec

Large language models have gained popularity, but also garnered discussions about implicit biases in large models (Nissim et al., 2020; Zou and Schiebinger, 2018). This should be taken into account when using blackbox approaches, such as neural network training. Altszyler et al. investigate small text corpora for word associations and find that Latent Semantic Analysis (LSA) can provide superior results for low-frequency words compared with Word2Vec (Altszyler et al., 2017). They apply their algorithms to dream journals. A large corpus of abstracts and papers related to biomedical research was analyzed by Zhu et al. (Zhu et al., 2017). They conclude that increasingly large datasets can result in more relations in biomedical terms, but do not guarantee better precision. This ties into the strength of word associations, but limited semantic properties of Word2Vec. Kazama et al. propose an ingredient substitution method based on the vector arithmetic allowed by the Word2Vec algorithm (Kazama et al., 2018). One of their proposals is to allows a change in cuisine style from one cultural background to another, but also to adjust meals to a persons's individual dietary preferences.

Other limitations and strengths of Word2Vec were investigated by Di Gennaro et al. (Di Gennaro et al., 2021). Overfitting was mostly ruled out for the training process of large datasets and they list analogy models as a strength. Syntactic relationships were not able to be achieved by Word2Vec.

3. Research Method

3.1. Online Course and Questionnaire Design

The online course used for this research was built with WordPress, one of the most popular blogging and content management systems (CMS), and the premium plugin LearnDash (WordPress, 2022; Learndash, 2022). WordPress has also been used for educational purposes (Scott, 2012; Quesenberry et al., 2014). The course design followed the same format in both courses, with same functionality, user-interface, and lesson-topic-quiz structure.

To perform an ex-ante and ex-post analysis, students were surveyed before and after taking the online course. The questionnaires were designed with three types of questions in mind—to gather statistical information about age and English proficiency, functional and dysfunctional questions for online course features to assign the Kano categories, and freeform textual comments. The functional and dysfunctional questions were asked in the same form before and after the online course was taken. 12 online course features were surveyed, detailed in Section 4. Three additional questions about what students liked, disliked, and wished for were asked ex-post. Questions were selected based on previous research and course specific insights (Selvi, 2010; Dominici and Palumbo, 2013).

3.2. Kano Model

The Kano model was first introduced by Kano et al in 1984 and is a tool to conduct customer satisfaction research (Kano et al., 1984). At the core is a questionnaire that asks (prospective) customers to answer functional and dysfunctional questions regarding a product's feature. The *functional* question asks customers how they feel if a feature is present (implemented) and the *dysfunctional* question asks

customers how they feel if a feature is missing. Combining these answers results in one of six Kano categories. Participants have to rate a feature in five levels for both these questions—highly satisfied, as expected, neutral, can live with it, and highly dissatisfied.

The list below details the categories that the Kano model produces. As the original publication describing the categories is written in Japanese, the terminology has been adjusted. *Quality* has been replaced by *requirement* to reflect terminology from software engineering and the term *must-be* was replaced by *basic*. For this research, the term *customer* is replaced by *user*, *student*, or *participant*.
B (basic requirement), O (one-dimensional requirement), I (indifferent requirement), Q (questionable requirement), A (attractive requirement), R (reverse requirement).

To illustrate the meaning of the Kano model categories covered in the case of an e-learning platform or online course, the following examples are explained for each requirement. A working website, consistent URLs, and mobile-friendly content can be considered basic requirements (B). They do not increase the satisfaction if properly included, but would reduce the satisfaction if missing or implemented poorly. An attractive feature (A) could be a live note-taking function or engaging implementation of gamification. Over time, these features can become one-dimensional (O) or basic requirements (B). One-dimensional requirements (O) increase the user's satisfaction proportional to the degree of implementation. Factors that result in a good User Experience (UX) increase the satisfaction proportionally to the degree of implementation, as well as in the opposite direction. The Kano model includes states that cannot be captured with other methods in this field. Reverse requirements (R) have the opposite effect of one-dimensional requirements (O) or in some cases of attractive requirements (A), i.e., they decrease (proportionally) the satisfaction when implemented. An example for web-based applications would be pop-up dialogue boxes. This is one of the reasons for choosing the Kano over other commonly used ones, such as SERVQUAL, E-Learning Satisfaction (ELS), or e-SERVQUAL (Wang et al., 2007; Parasuraman et al., 1988, 2005). Dominici and Palumbo also detail two more reasons for this methodology: the first is crucial to this research and allows ex-ante and ex-post analysis, the second is that this model does not assume a linear relationship between the service performance and satisfaction (Dominici and Palumbo, 2013; Chaudha et al., 2011).

3.3. Language Model with Word2Vec

Machine learning algorithms to create language models can help analyze unstructured textual data by converting words into numerical vector values called word embeddings. By training Word2Vec on a text corpus, words can be compared in vector space. The algorithm allows to create a model that returns a numerically close word, given a context, or return a numerically close context, given a word. The former is called continuous bag-of-words (CBOW) and the latter a skip-gram model and used in this research. As words are represented as multidimensional vectors, the algorithms also allows to perform vector arithmetic (Gittens et al., 2017). Word embeddings can be added (context positive) to each other or subtracted from each other (context negative), which numerically shifts the predicted context or target word in the word-vector space. Stanza, a natural language processing (NLP) toolkit developed by the Stanford NLP group was used to tokenize and lemmatize the words (Qi et al., 2020). Stop-words were removed from the sentences and all words were converted into lowercase.

Word2Vec was trained on the dataset with a vector size of 60 to avoid overfitting (observed with smaller datasets and default vector size of 100). Ex-ante comments were primed with the word need, ex-post comments with like, dislike, and wish, respectively, to allow targeted context and clearer context shifts via word-vector arithmetic. For higher accuracy, negative sampling is used on contexts (Aggarwal, 2018). The model was trained using the Word2Vec implementation of Gensim for Python (Rehurek and Sojka, 2010) with the following parameters: Skip-Gram mode, learning rate $\alpha = 0.001$, window size 5 for short-term and 7 for long-term, vector size of 60, and minimum word-count of 1.

4. Results and Discussion

In the long-term online course, 36 undergraduate students participated, 18 male, 18 female, and 0 non-binary. The average age was 19.00 ($\sigma = 1.18$), ranging from 18 to 23. The course was conducted over a period of 15 weeks. Students could access new material on the online course each week after a preset

date. For the short-term online course, 16 undergraduate students were surveyed, 5 male, 11 female, and 0 non-binary. The average age was 19.82 (σ = 1.70), ranging from 18 to 24. The experiment, including introduction, questionnaire conduct, and taking the online course was completed in 90 minutes. The following 12 online course features were surveyed before and after the online course (abbreviations corresponding to Figs. 1): User-friendly platform (Ufp) • Certificate of completion (Coc) • Download of course material (Docm) • Own profile and account page (Opaap) • Quizzes and exercises (Qae) • Interactive quizzes and exercises (Iqae) • Comment function (Cf) • Personal tutor (Pt) • User manual for the platform (Umftp) • Videos (V) • Photos/Graphics (PG) • Text (T). A satisfaction index (CS) and dissatisfaction index (CD) was calculated (Eq. (1) and Eq. (2)). Each feature has a pair of CS and CD values. C(k) denotes the frequency of each Kano category.

$$\mathrm{CS}_f = \frac{C(\mathrm{A})+C(\mathrm{O})}{C(\mathrm{B})+C(\mathrm{O})+C(\mathrm{A})+C(\mathrm{I})} \tag{1}$$

$$\mathrm{CD}_f = \frac{C(\mathrm{B})+C(\mathrm{O})}{C(\mathrm{B})+C(\mathrm{O})+C(\mathrm{A})+C(\mathrm{I})} \cdot -1 \tag{2}$$

$$C(k) = \sum_{i=1}^{n} [f_i = k], \tag{3}$$

where k ∈ {A,B,O,I}, and f is the feature variable of each of the 12 Kano features f ∈ {Ufp, Coc, Docm, Opaap, Qae, Iqae, Cf, Pt, Umftp, V, PG, T}. The two equations allow a cost-benefit tradeoff of gaining satisfaction and preventing dissatisfaction (Berger et al., 1993; Dominici and Palumbo, 2013). Values for CS range from 0 to 1 and higher numbers indicating higher satisfaction, CD ranges from −1 to 0 and lower numbers indicating higher dissatisfaction. Values of CS and CD are plotted both for the short and long-term course in Fig. 1. The values are connected by a vector from ex-ante to ex-post. The length (magnitude) of the vectors is reflected by the line thickness. The two language models trained by Word2Vec resulted in the following list, word-embeddings (in bold) and context words (with similarity scores). Due to the small corpora sizes, words with a similarity scores larger than 0.2 are listed.

Short-term

- **need**: video (0.35), understand (0.29), test (0.26), system (0.23), content (0.22), teacher (0.21)
- **like**: professional (0.34), English (0.3), understand (0.28), photography (0.24), study (0.23), beautiful (0.21), restart (0.21), test (0.21)
- **dislike**: manual (0.39), complex (0.28), frustration (0.25), want (0.21)
- **wish**: explanation (0.37), professional (0.31), select (0.26), hint (0.24), understand (0.23), feature (0.22), video (0.2)
- **need − wish**: explain (0.22), weakness (0.22), system (0.22), content (0.21)
- **wish − need**: explanation (0.43), select (0.34), professional (0.27)

Long-term

- **need**: daily (0.41), assignment (0.31), platform (0.3), anonymously (0.29), concentrate (0.28), detailed (0.27), chat (0.25), motivative (0.23), textbook (0.23)
- **like**: professor (0.38), compare (0.38), presentation (0.37), work (0.34), clearly (0.31), lecture (0.3), discussion (0.27), platform (0.27), intuitive (0.26), motivative (0.23)
- **dislike**: participation (0.45), video (0.38), unmotivated (0.35), material (0.32), download (0.32), bad (0.3), vague (0.29), cheating (0.27), missing (0.26), concentrate (0.21)
- **wish**: time (0.38), download (0.36), miss (0.35), study (0.32), consult (0.31), find (0.31), anonymously (0.29), flexibility (0.27), q&a (0.26), concentrate (0.25), understanding (0.25), explain (0.24), motivative (0.24), detailed (0.24), data (0.23)
- **need − wish**: interaction (0.3), member (0.29), important (0.29), helpful (0.28), anonymously (0.25), contact (0.23)
- **wish − need**: time (0.41), post (0.34), question (0.33), learn (0.3), interesting (0.3), study (0.3), proceed (0.3), description (0.28), data (0.27), example (0.26)

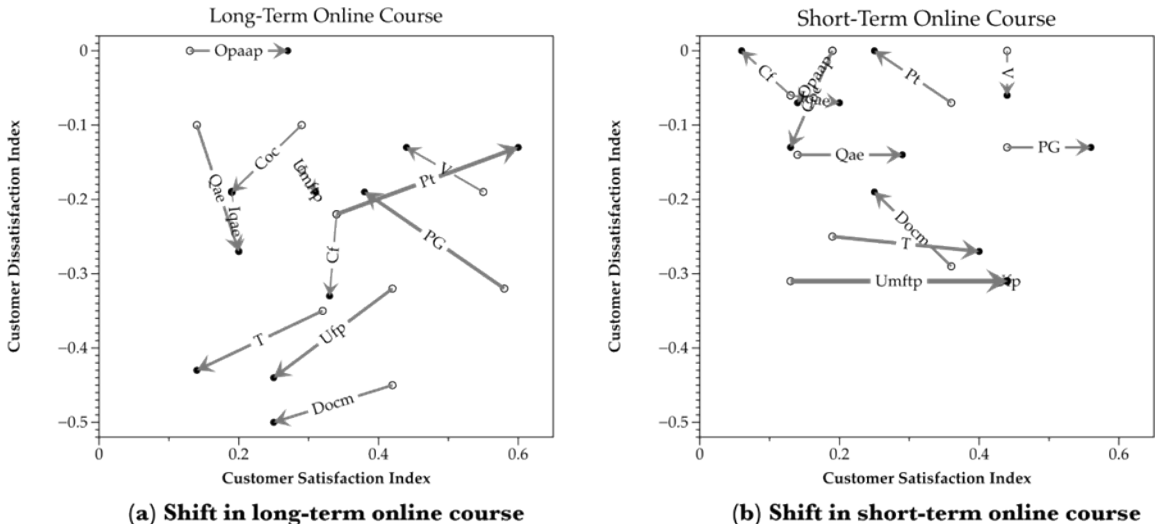


Figure 1. Shifts of Satisfaction and Dissatisfaction Indices in online course of different duration.

From Fig. 1 and the context word list, it can be seen that the students from the short term online course struggled to simultaneously work through the course while figuring out how the site works. The feature *User manual for the platform* had the strongest increase in satisfaction index, while the language model for **dislike** lists *manual*, *complex* and *frustration* in combination with the word **wish** and *download*. Photos and graphics were regarded as positive in the freeform comments and a raise in user satisfaction index can also be observed. The Kano feature *Text* also experienced an increased satisfaction index. The authors suggest the ex-ante expectation to be more on multimedia, with an ex-post consumption experience that text helped in explaining concepts, supported by the word-vector arithmetic of **wish** − **need** and *explanation*. During the long-term online course, the feature *Personal tutor* experienced the largest increase in the satisfaction index (Fig. 1). Students felt the necessity to consult with the professor and tutor throughout the semester, supported by **like** and *professor*, *lecture*, and *discussion*. The largest decrease in satisfaction index could be observed for *Photos/Graphics*, *Text*, *User friendly platform*, and *Download of course material*. The first experienced a decrease in dissatisfaction index, whereas in the latter three, a decrease in CD could be observed. Features that simultaneously have a lower CS and CD indicate a drifting towards basic or one-dimensional requirements, that do not increase the satisfaction when implemented correctly, but suffer heightened dissatisfaction when missing. Some students wished for more detailed materials, as can be seen from **dislike** and *material*, *download* and *vague*.

5. Limitations

The content of the courses were different, with a bias towards visual content for the short-term introduction to photography course. Students that took part in this course were from a Japanese psychology college. Students from the long-term introduction to programming course were from a liberal arts college with a mix of Japanese and international (mostly Asian) students. A cultural component could also influence the results. The sample sizes, especially for the short-term online course is small and future work is planned to increase participants. However, while the absolute number of participants is in general terms statistically low, the homogeneous nature of Japanese students and the rarity of such data should also be taken into consideration. The limited availability of conducting comprehensive academic online courses should however be taken into account as well.

6. Conclusions

This study presents a combination of the Kano model with insights from language models to compare short-term and long-term online courses. The use of word-embeddings can be used to further quantify results obtained by the Kano method. The results show that needs differ depending on the format a

course is offered. An ex-ante and ex-post analysis can deepen the understanding of expectation versus consumption experience. Short-term online courses could experience an increase in user satisfaction when including a short introduction video to explain how to navigate the platform. Long-term online courses benefit from a personal tutor and instructors can avoid dissatisfaction by continuously extending the online material, a well as investing in adequate web-hosting.

References

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. Springer International Publishing AG.

Altszyler, E., Ribeiro, S., Sigman, M., and Fernandez Slezak, D. (2017). The interpretation of dream meaning: Resolving ambiguity using latent semantic analysis in a small corpus of text. Consciousness and Cognition, 56:178–187.

Berger, C., Blauth, R. E., and Boger, D. (1993). Kano's methods for understanding customer-defined quality. Center for Quality Management Journal, 2:3–36.

Chaudha, A., Jain, R., Singh, A. R., and Mishra, P. K. (2011). Integration of kano's model into quality function deployment (qfd). The International Journal of Advanced Manufacturing Technology, 53(5):689–698.

Di Gennaro, G., Buonanno, A., and Palmieri, F. A. N. (2021). Considerations about learning Word2Vec. The Journal of Supercomputing.

Dominici, G. and Palumbo, F. (2013). How to build an e-learning product: Factors for student/customer satisfaction. Business Horizons, 56(1):87–96.

Gittens, A., Achlioptas, D., and Mahoney, M. W. (2017). Skip-gram – Zipf + uniform = vector additivity. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 69–76, Vancouver, Canada. Association for Computational Linguistics.

Kano, N., Seraku, N., Takahashi, F., and ichi Tsuji, S. (1984). Attractive quality and must-be quality. Journal of The Japanese Society for Quality Control, 14(2):147–156.

Kazama, M., Sugimoto, M., Hosokawa, C., Matsushima, K., Varshney, L. R., and Ishikawa, Y. (2018). A neural network system for transformation of regional cuisine style. Frontiers in ICT, 5.

Learndash, (2022). Learndash.com

Nissim, M., van Noord, R., and van der Goot, R. (2020). Fair Is Better than Sensational: Man Is to Doctor as Woman Is to Doctor. Computational Linguistics, 46(2):487–497.

Parasuraman, A., Zeithaml, V. A., and Berry, L. (1988). Servqual: A multiple-item scale for measuring consumer perceptions of service quality. Journal of Retailing, 64:12–40.

Parasuraman, A., Zeithaml, V. A., and Malhotra, A. (2005). E-s-qual: A multiple-item scale for assessing electronic service quality. Journal of Service Research, 7(3):213–233.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.

Quesenberry, K. A., Saewitz, D., and Kantrowitz, S. (2014). Blogging in the classroom: Using wordpress blogs with buddypress plugin as a learning tool. Journal of Advertising Education, 18(2):5–17.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta.

Scott, A. D. (2012). WordPress for Education. Packt Publishing.

Selvi, K. (2010). Motivating factors in online courses. Procedia - Social and Behavioral Sciences, 2(2):819–824. Innovation and Creativity in Education.

Sun, P.-C., Tsai, R. J., Finger, G., Chen, Y.-Y., and Yeh, D. (2008). What drives a successful e-learning? an empirical investigation of the critical factors influencing learner satisfaction. Computers & Education, 50(4):1183–1202.

Wang, Y.-S., Bauk, S., Scepanovic, S., and Kopp, M. (2014). Estimating students' satisfaction with web based learning system in blended learning environment. Education Research International, 2014.

Wang, Y.-S., Wang, H.-Y., and Shee, D. Y. (2007). Measuring e-learning systems success in an organizational context: Scale development and validation. Computers in Human Behavior, 23(4):1792–1808.

Wordpress, (2022). Wordpress.org

Zhu, Y., Yan, E., and Wang, F. (2017). Semantic relatedness and similarity of biomedical terms: examining the effects of recency, size, and section of biomedical publications on the performance of word2vec. BMC Medical Informatics and Decision Making, 17(1):95.

Zou, J. and Schiebinger, L. (2018). Ai can be sexist and racist—it's time to make it fair. Nature, 559(7714):324–326.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

A Curriculum Package of Social and Ethical Concerns in Metaverse Ecosystem in School Education Based on Bloom's Taxonomy

Percy Lai-Yin KWOK [a*], Joe Kwong-Ngai CHEN [b] & Jordan Tsz-Chun FUNG [c]
[a] *The Education University of Hong Kong*
[b] *ICT-In-Physical Education Foundation Ltd.*
[c] *Jordan Fung & Company (Hong Kong) Ltd.*
*lykwok@eduhk.hk

Abstract: Facing the newly emerged blockchain technology in metaverse construction, the instructional design, pedagogy, and curriculum package of the metaverse ecosystem is underdeveloped or under-researched in education locally and globally. Proponents of metaverse education mostly focus on its application in the virtual world without thoroughly thinking about its connections between the physical world and virtual world. Metaverse construction not only provides lots of work-from-home business and studying opportunities during the considerable impacts of COVID-19 but also raises lots of social and ethical issues such as data privacy, virtual identity, investment/security risks, system integrity, social media toxicity, and corporate control problems. There is an urgent need to address such social and ethical concerns in some curriculum packages of metaverse education. Based on this perspective, this paper aims to develop a curriculum package of addressing some social and ethical concerns in daily-life applications of blockchain technology in upper secondary (K10-K11) education. Through the six levels of Bloom's taxonomy framework, course materials have been developed to provide the value-laden ecosystem concerning decentralized properties and the digitalization of IoT and STEM metaverses in some lessons. This paper also endeavors to help those participating K10-K11students raise their social and ethical awareness during the course development of a curriculum package related to metaverse literacy.

Keywords: Metaverse literacy, social & ethical concerns, curriculum package, Bloom's Taxonomy

1. Introduction

Metaverse is a current and future integrated totality of the physical world and virtual world connecting internet-of-things (IoT) and other aspects of human lives with a wide range of applications of (wired / wireless) computer networks, handheld mobile devices, augmented reality / virtual reality / substitutional reality / mixed reality / extended reality, artificial intelligence (AI) and robotics. Currently, metaverses provide a range of human imaginations and potential functionalities that enable new modes of e-business, e-trades, e-digital currencies and e-learning, and lots of immersive social and virtual experiences in STEM education and game-based learning (GBL). IoT and STEM metaverses refer to focused uses of internet-of-things (IoT) and STEM applications respectively in daily lives connecting the physical world and virtual world in metaverses.

The newly evolved Blockchain technology is a peer-to-peer decentralized, distributed ledger that keeps digital assets transparent and unchangeable without the involvement of any third-party intermediary. In particular, bitcoin is a cryptocurrency and one application of blockchain technology. Notably, there are some system integrity, privacy threat, and (monetized) data insecurity problems found in the uses of blockchain technology such as immoral and illegal withdrawal of e-cash or e-currencies in bitcoins, causing the possible financial crisis in national, regional, and global economies (Fromkin, 2022; Longstaff, 2019). Recently, an increasing number of ICT educators have paid heed to ethical and legal concerns of blockchain technology such as privacy, security, environmental impacts, loss of human jobs, and investment risks in metaverse education (Houlding, 2019). Students and teachers ought to be cautious with those underlying ethical and legal issues, and some social inequity problems like potential bad impacts of blockchain technology upon those elderly and disadvantaged ICT-illiterate persons. Owing to the threats posed by the pandemic, metaverse construction not only provides lots of work-from-home business and studying opportunities but also raises lots of social and ethical issues such as virtual identity, data privacy, social media toxicity, and corporate control

problems (Djordjevic, 2022). There is an urgent need to address such social and ethical concerns in metaverse education (Gurmen, 2022) and to develop suitable curriculum packages locally and globally.

2. Metaverse Ecosystem

In a balanced metaverse ecosystem, arts/sports/Green STEM activities in the physical world need to be involved whilst game-based learning /AI & robotics / digitalized STEM activities in the virtual world also need to take place through industry-market collaboration. Participating students and teachers in the earlier stages of the project have enriched and connected their real-world and virtual-world learning experiences through the IoT bridge. Mere one-sided physical or virtual world explorations might limit their learning experiences without inter-world connections.

3. Research Project

Facing the newly emerged blockchain technology and metaverse construction, the instructional design, pedagogy, and curriculum packages of the ecosystem of metaverses using blockchain technology have been under-developed and under-researched in education locally and globally (c.f. Hirsh-Pasek, *et al.* 2022). And the main reason for using the six-level Bloom's Taxonomy in the project is to use a widely accepted academic standard to evaluate the quality of the curriculum package and measure participating students' cognitive gain in the ecosystem of the mentioned metaverses. In this project, two key features of metaverses related to KT are investigated in table 1. The main reason why a group of upper secondary school students will be selected is that lower secondary and primary students cannot master blockchain technology without any fundamental concepts of shareability and scalability of databases and advanced programming techniques and tertiary students are not selected properly due to their tight studying schedules and short school semesters, in comparison with upper secondary ones.

Table 1. *Two key features of metaverses and socio-cognitive functions in student activities*

Key Features of Metaverses	**Socio-cognitive Functions in Two-way Knowledge Transfer (KT) model**	**Students' exemplified activities in IoT and STEM metaverses**
Decentralized	Communal ownership enhances socialization and externalization in KT	Students learn blockchain to understand the importance of communal ownership for e-trade transactions, tokens, e-marketing prices & bitcoins
Digitalization	Monetization in digital currencies to build value systems of buying goods and consuming services in combination (cognitive synthesis) in KT	In digital currency ecosystems using blockchain technology, students use e-cash coupons in AI shops to buy / sell products and services in metaverses

The objectives of the current project are:
a. To develop and fine-tune an innovative curriculum package of the value-laden ecosystem (concerning decentralized properties and digitalization) of IoT and STEM metaverses (using blockchain applications) in education using Bloom's Taxonomy framework
b. To help upper secondary students raise ethical awareness during daily life applications of Blockchain technology using Bloom's Taxonomy framework during the course development of curriculum package in catering for their learning diversity in some studying classes

Notably, some value-laden parameters in the objective **a** like core humanistic values of blockchain uses in metaverses, moral conscience of investors in bitcoins, and ethical and equity concerns in metaverses ecosystem have been added to the curriculum package. Research subjects are 60 upper secondary (K10-K11) students (in three classes of 20 students) in Hong Kong and 6 employed teaching mentors are some undergraduate students in universities or come from some NGOs (with undergraduate major in AI, educational technology, computer engineering or computer science) during the two testing periods.

4. Curriculum Package

The curriculum package in normal classroom or after-school settings contains beginners / intermediate / advanced learner levels of learning topics / studying themes related to social and ethical concerns of ten lessons per class, based on Bloom's Taxonomy; and some exemplars of individual course assignments and group project involving coding design on some social and ethical concerns related to blockchain applications in some IoT and STEM learning activities with continuous, summative and formative assessment rubrics. Initial lesson contents (two lessons per topic) on blockchain technology in the metaverse ecosystem include subject knowledge on blockchain technology and conditional knowledge in the metaverse literacy framework and applied knowledge related to social and ethical concerns in the following table 2.

Table 2. *Types of Knowledge on Blockchain Technology in School Education*

Subject (propositional) knowledge	**Conditional (meta-cognitive) knowledge in metaverse literacy framework**	**Applied (procedural) knowledge in social and ethical concerns**
1. Historical development of different blockchains in local and global contexts	Students know how to: ● use, provide and communicate information on blockchain technology ethically and responsibly ● have effective and ethical uses of information for their lifelong learning (EDB, 2018)	Students are instructed to do group presentation projects on raising their moral awareness and categorizing types of ethical issues (data privacy, virtual identity, investment / security risks, system integrity and social media toxicity problems) towards daily-life uses of blockchain technology
2. Ethereum, EVM and solidity coding 3. Various applications of EVM blockchain 4. NFT metadata structure and storage 5. Future development, Proof-of-work vs Proof-of-stake	Students know how to: ● use technology to improve use of digital information ● research effectively and responsibly (The Ministry of Education and Child Care, British Columbia, 2022)	Students are instructed to do literature review in group projects on using blockchain to address problems about its social and ethical impacts on information societies and endeavor to find out feasible solutions or alleviate those problems using their subject (proposition) knowledge

Three normal or after-school classes of mixed-ability groups of upper secondary students in 2022-23 have been instructed to do various learning tasks of the beginners / intermediate / advanced levels in class activities and group projects works in the two testing periods. Students and mentors of the first two classes will help refine the curriculum package with some studying topics / learning tasks using the six levels of Bloom's Taxonomy in their individual course assignments and group project works. The third class in later stages will further help refine the topics / tasks of the revised version of the curriculum package after interim evaluation. Table 3 provides a six-level of seeking explanations and guided studying tasks and table 4 provides a three-level structure of learners in the instructional design of the curriculum package on social and ethical concerns of blockchain technology.

Table 3. *A Six-level Structure of Seeking Explanations and Guided Studying Tasks / Questions*

Levels of BT	Seeking-explanations in instructional design of the curriculum package	Guided students' studying tasks / thinking questions in curriculum package
Remembering / Recalling	Students is instructed to recall and recite the concept of blockchain learned from relevant digital data sources	'What is blockchain'?
Understanding	Students can explain the relationships between the two concepts in their own words	'How are bitcoins related to blockchain technology?'
Applying	Students can give particular examples of some concepts in their daily encounter	'How is blockchain used in digitalization of metaverses? Please give some examples.'
Analyzing	Students can analyze the whole organizational structure of the codes / programs in terms of its individual components and finding their interrelationships in given codes / programs	'Please find out some / all logical errors in the given codes / programs involving the decentralized properties of metaverses.'
Extending/ Synthesizing	Students can extend the given codes to other learning scenarios and put parts into a whole to work out a flow chart / coding system	'Please write a flowchart / a simple system to perform some decentral features of database using the program codes.'
Creating / Evaluatin g	Students can make value judgments by comparing the advantages and disadvantages of both 'bottom-up' and 'top-down' approaches in databases in blockchain applications. Some even crease their own blockchain coding system.	'Given the two 'bottom-up' an 'top-down' approaches i databases, which is the best suitabl one for blockchain applications And why?' 'Can you create a ne blockchain coding system?'

(modified from. Starr, Manaris, & Stalvey, 2008)

The whole curriculum package in daily-life applications of blockchain technology has a wide range of learning topics / tasks at beginners / intermediate / advanced levels using Bloom's Taxonomy (BT) framework in the following table 4.

Table 4. *A Three-level Structure of learners Using Bloom's Taxonomy in the instructional design*

Levels of learners	Production of knowledge in individual assignments and group projects	Seeking-Explanations Aspects for production of knowledge
Beginners	Recalling / remembering	Understanding
Intermediate	Applying	Analyzing
Advanced	Extending / Synthesizing	Creating / Evaluating

(modified from. Starr, Manaris, & Stalvey, 2008)

5. Potential Contributions of the Curriculum Package

Owing to its limited length, this paper merely highlights some social and ethical concerns in daily-life applications of blockchain technology in a metaverse ecosystem connecting the physical and virtual worlds. The curriculum package enriches local and global curricula in metaverse education, especially curriculum development in the metaverse ecosystem on its digitalization and decentralized properties, and adds more value-added parameters to social and ethical concerns in local and international secondary curricula related to metaverse literacy.

References

Djordjevic, I. (2022). *Ethical implications of the metaverse and the role of blockchain tech within.* https://www.coinstacknews.com/news/morality-and-ethical-concerns-of-the-metaverse/

Education Bureau (EDB), The HKSAR Government (2018). *Information literacy for Hong Kong students' learning framework.*

https://www.edb.gov.hk/attachment/en/edu-system/primary-secondary/applicable-to-primary-secondary/it-in-edu/Information-Literacy/IL20180516E.pdf

Fromkin, S. (2022). *The metaverse can provide a whole new opportunity for education. Here's what to consider.* https://www.fastcompany.com/90718919/the-metaverse-can-provide-a-whole-new-opportunity-for-educati on-heres-what-to-consider

Gurmen, I. (2022). *Socio-ethical challenges of the metaverse*. https://thenextcartel.com/discover/socio-ethical-challenges-of-the-metaverse

Houlding, D. (2019). *Blockchain: 6 key ethical considerations:* https://lifeboat.com/blog/2019/01/blockchain-6-key-ethical-considerations

Longstaff, S. (2019). *Blockchain: Some ethical considerations.* https://ethics.org.au/blockchain-some-considerations/

Starr, C. W., Manaris, B., & Stalvey, R. H. (2008). Bloom's Taxonomy revisited: Specifying assessable learning objectives in Computer Science. *Conference proceedings of SIGCSE'08*, Portland, Oregon, USA.

The Ministry of Education and Child Care, British Columbia (BC) (2022). *BC's digital literacy framework.* https://www2.gov.bc.ca/assets/gov/education/kindergarten-to-grade-12/teach/teaching-tools/digital-literacy-framework.pdf

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

PyGuru: A Programming Environment to Facilitate Measurement of Cognitive Engagement

Daevesh SINGH[a*], Hema SUBRAMANIAM[b], Ramkumar RAJENDRAN[a]
[a]*IDPET, Indian Institute of Technology Bombay, India*
[b]*Department of Software Engineering, Universiti Malaya, Malaysia*
*daeveshsingh@iitb.ac.in

Abstract: There is a vital link between students' engagement and their performance. Despite this, a dearth of studies exploring student engagement (especially cognitive) in computer programming, where high failure rates are a norm. Interestingly, online environments capture fine-grained interactions; this data has the potential to detect students' cognitive engagement. The contribution of this paper is two-fold. Firstly, it introduces an online learning environment to learn and practice python programming called PyGuru and the data logged in the system. Secondly, we provide a preliminary analysis of students' interaction with the system to reveal their level of cognitive engagement in different system components. The results of this study conducted for two weeks and 21 students demonstrate that students perform more actions that correspond to passive engagement than other levels of engagement.

Keywords: Programming, Learning Environment, log data, PyGuru, Cognitive Engagement

1. Introduction

The significance of Student Engagement(SE) in learning is well established in the literature (Duderstadt et al., 2002). Empirical studies demonstrate that SE is linked significantly with critical thinking, problem-solving and academic success (Carini et al., 2006; Wang & Holcombe, 2010; Zhao et al., 2021). On the other hand, introductory programming courses suffer a fate of high failure and withdrawal rates (Watson & Li, 2014). Due to the strong association between SE and academic success, there is a keen interest among researchers in identifying SE levels to assist at-risk students.

SE is "*participation in educationally effective practices, both inside and outside the classroom, which leads to a range of measurable outcomes*" (Kuh 2007). In literature, various dimensions of engagement have been defined, including behavioural, affective, cognitive etc. Our work focuses on Cognitive Engagement (CE) as it is not directly observable, making its measurement challenging. There is also a paucity of literature that studies CE in the context of computer programming.

Although some studies focus on CE in online learning in other domains, they mainly focus on only one of the online learning components like video watching, discussion forums, etc. Hence, there is a need to analyse all the actions students' perform in a Computer-Based Learning Environment (CBLE) to understand CE better. This analysis will enable the instructors to identify at-risk students and provide interventions to prevent failure and dropouts.

This study proposes a CBLE for teaching-learning computer programming called PyGuru based on the ICAP framework. PyGuru captures students' interaction in four online components: book-reader, video player, discussion forum, and IDE. This paper provides details about the design and the data captured by the system. Further, we present learners' interaction data to demonstrate how they interact with PyGuru and the different modes of engagement they involve. This study was conducted for two weeks in a public university in Malaysia in the research methods classroom with 21 postgraduate students. The results obtained after analysis of log data reveal that students perform more actions that correspond to lower levels of engagement. This lower engagement could be one of the reasons for lower learning levels.

2. Background and Related Work

CE is the thinking that students get involved in while learning (Helme & Clarke, 2001). As per Fredricks et al. (2004), "*cognitively engaged students would be invested in their learning, would seek to go beyond the requirements, and would relish challenge*" (Fredricks, Blumenfeld, & Paris, 2004). CE is the most misunderstood and ill-defined construct (Greene, 2015) and thus making its operationalisation and measurement difficult and challenging (Chi et al., 2018). To overcome this gap, Chi and Wylie (2014) proposed a hierarchical framework called ICAP to understand CE (Chi & Wylie, 2014). This framework tries to classify students' overt behaviours into different modes of engagement: interactive, constructive, active, and passive.

The passive mode of engagement is when "*learners receive information without overtly doing anything related to learning*" (Chi & Wylie, 2014, p. 221). In this mode, learners attend to the information (without performing any actions like note-taking) and store it in episodic form rather than integrating it with prior knowledge. Active engagement occurs when the learner's information acquisition is accompanied by specific physical or motoric actions that support their learning. This engagement includes taking notes in the classroom, pausing and recapping the videos or highlighting the text while reading. The cognitive process involved during active engagement demands the activation of prior knowledge and integration of the new information into the existing one. Constructive engagement happens when the learners attempt to produce artefacts using information that goes beyond the information available in the environment. Constructive engagement includes "*elaborating, comparing and contrasting, generalising, reflecting on, and explaining how something works*" (Chi & Wylie, 2014, p. 228). Interactive engagement occurs when learners, during interactions with partners, are constructively engaged, and there is sufficient turn-taking. These partners could be peers, instructors or even computer agents. ICAP framework also informs that the highest learning occurs at the interactive level followed by constructive, active, and the least when engagement is passive.

ICAP model was initially developed for classroom learning; however, this framework was later extended to online mode as well. Yogev et al. (2018) examined students' CE in reading material using the Nota Bene annotation platform. They firstly analysed CE anchored in the text by manually labelling students' annotations and then developed an interactive decision tree to automate this process. Similarly, Dodson et al. (2018) developed a framework to classify students' different video-watching behaviours as per the ICAP framework. They classified behaviours as passive(playing video content), active(replay, pause, seek specific information), constructive(taking notes, highlighting), and interactive (cooperating and collaborating with others). The study by Atapattu et al. (2019) tried to automate the process of classifying the post in the discussion forum into active and constructive modes of engagement. The posts that were significantly different from the learning materials were classified as constructive, and the more similar ones were labelled as active.

To summarise, we have presented how ICAP is used in book reader, video-player and discussion forums. These studies are crucial in understanding students' CE in different online learning components. However, understanding students' overall CE is also crucial and to fill this gap, we have designed our system PyGuru, which combines these three components, namely book reader, video-player and discussion forum, along with an IDE. This is done to investigate how overall CE (in book reader, video player etc.) impacts students' performance in Python programming. To understand this, we present PyGuru- a learning environment for learning Python programming capable of logging the user actions.

3. PyGuru: Learning Environment

PyGuru (https://pyguru.personaltutoring.in/) is a computer-based learning environment developed to teach and learn Python programming skills. PyGuru has four components: book reader, video player, code editor, and discussion forum. This section describes each of these components.

The book reader (shown in Fig 1) in PyGuru contains textual information readers can highlight and annotate. The annotating feature in the learning environment comprises selecting a text, commenting

on that text, and providing a tag to that text.

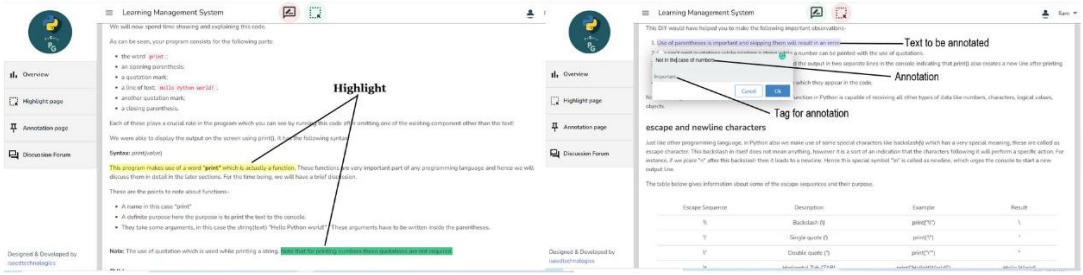


Figure 1. The figure shows the features of the book reader. The highlight feature is shown in the left image. The right image shows the annotation feature.

PyGuru has an interactive video-watching platform (Fig. 2) that allows learners to perform basic actions like enhancing the speed of the video, playing, pausing, and seeking. Along with this, more advanced interactive features are embedded into the system that allows the instructor to add questions within the video. The video automatically stops and waits for the learner's response.

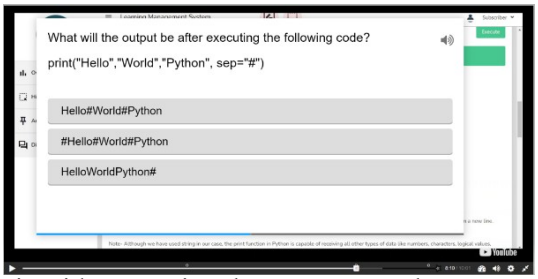


Figure 2. The figure shows the in-video question that appears on the screen. The video player has a panel below that allows learners to adjust the speed and volume. Also, the size of the window can also be adjusted.

Learning programming requires a code editor where learners can practise coding. PyGuru offers two kinds of code editors (Fig 3). The first kind of code editor is embedded into the book reader to facilitate learners to practice codes immediately after learning about the concept. As shown in (Fig 3 left), this code editor has a coding window and an execute button.

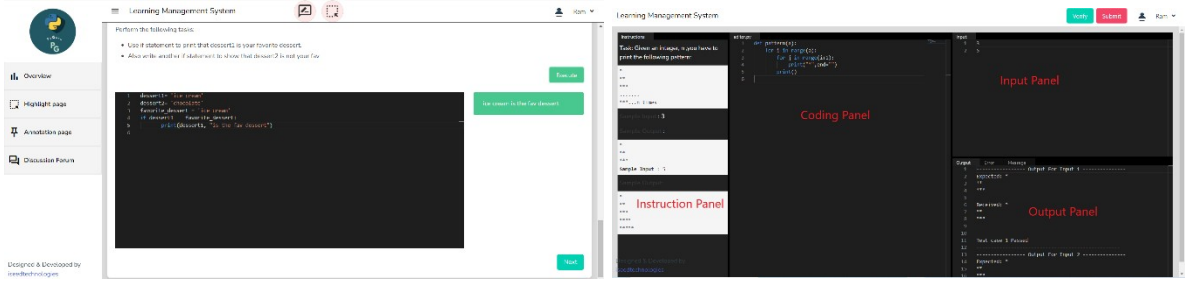


Figure 3. The figure on the left shows the code editor embedded in the book reader. The figure on the right shows the code editor with four panels.

The second kind of code editor shown in (Fig 3, right) is more advanced and used to assign programming questions to students. It evaluates the learners' code against the test cases. This code editor consists of four panels, a) Instruction panel provide details about the problem, b) Input panel contains the test cases for the problem, c) Coding panel is where the student is expected to write the code, d) Output panel displays the output once the program is run. It will also provide information about the number of test cases passed and failed.

PyGuru also has a discussion forum. The learners can use this discussion forum to put forth their queries which the instructor or their peers can then answer. The students can also like the post to show that their question has been answered.

4. Methods

This section describes the methodology. The first subsection informs about the participants. The second subsection provides details about the data captured.

4.1 Participants

The study was conducted with 21 students who used PyGuru to learn Python programming in 2022. These students were enrolled in a Master's program at a public university in Malaysia in the research methods course. Informed consent was obtained from the student, and the study was cleared by the Institute Ethics Committee (IEC). No monetary compensation was given to the students. The course instructor provided students with a demographic survey that involved questions like their name, age, gender etc. at the beginning of the course. The duration of this study was two weeks. During this time, the students interacted with PyGuru. Their interaction with the system (clickstream) and the time stamp was captured (more details about log data are provided in the following subsection). In addition to this, the students took a pre and post-test containing 12 multiple-choice questions along with two program writing questions from Python basics (variables, operators, conditional statements etc.).

4.2 Data

Log data from PyGuru was collected for all the 21 students while they interacted with the system for two weeks. In this subsection, we provide the details of the log data generated in each component of PyGuru. Table 1 provides the details of the interaction data. We will now explain the data presented in table 1.

Table 1. *The table shows the data generated when students interact with PyGuru.*

S_id	Time_Stamp	Action	Context	Context_Significance	Page
82	5/4/22 19:13	Log In			Home_Page
82	5/4/22 19:32	Log Out			
82	5/4/22 12:13	Reading			creating-a-varible
82	6/4/22 16:55	Highlight	\n		writing-first-program
82	6/4/22 17:15	Annotation	escape sequence	Error due to absence of \n	writing-first-program
82	7/4/22 20:11	Execute	success		writing-first-program
82	7/4/22 20:11	Watching			demo-on-print
82	7/4/22 20:19	Played	0:00		demo-on-print
82	7/4/22 20:22	Seek	8:01		demo-on-print
82	7/4/22 20:22	Paused	8:01		demo-on-print
82	7/4/22 20:25	VQ_opt_select	8:01	Correct	demo-on-print
82	9/4/22 10:25	VQ_retry	3:02		demo-on-input
82	10/4/22 8:30	Assessment			problem-on-if-else
82	10/4/22 8:37	Submitted	3		problem-on-if-else
82	11/4/22 13:25	Post viewed			Discussion Forum
82	11/4/22 13:25	Posted			Discussion Forum
82	12/4/22 9:30	Liked			Discussion Forum
82	12/4/22 9:32	Replied			Discussion Forum

The first and second columns provide the unique id of the learners and the timestamp of the action. The third column displays the actions in four different components. In book reader, the actions performed are reading, highlight, annotation and execution. The context column provides details about the text selected for highlight and annotation, and for execution, it informs whether there are errors. The Context_Significance for annotation offers information about the annotated text. The last column, named Page, gives the page's name on which the learner is currently present.

In the video player, the actions performed are played, paused, seek and the context provides the video time at which these actions were performed. Since the videos have in-video questions, the action VQ_opt_select correspond to the selection of one of the options in the in-video question, and the Context_Significance tells whether the correct option was chosen or not. If the option selected is incorrect, learners can retry the question and the action VQ_retry corresponds to this.

As mentioned earlier, PyGuru has two code editors. The second kind of code editor is more advanced and the action "Assessment" corresponds to using or accessing this code editor. The learners

can also click the submit or verify button to check if their code is thriving on the given test cases, and the number provided in the context informs the number of test cases successful.
The learners can view the discussion forum, write a post, like an existing post, and reply to the post in the discussion forum.

5. Results and Discussions

We analysed the log data from PyGuru to understand how learners interact with the system. As one of the main goals for developing this system is to measure the learners' CE. As a result, we looked into more granular data, i.e. what specific actions learners' perform in each component. Figure 4 presents the composition of different actions learners' perform in each component of the system. From these charts, it can be seen that the actions that correspond to lower levels of CE are more prominent. For instance, in book reader, the action of highlight and annotation indicate active and constructive engagement as per ICAP but simply reading without doing anything is passive engagement. We see that reading comprises 90% of the total activities done in book reader, indicating that very learners rarely have higher levels of engagement.

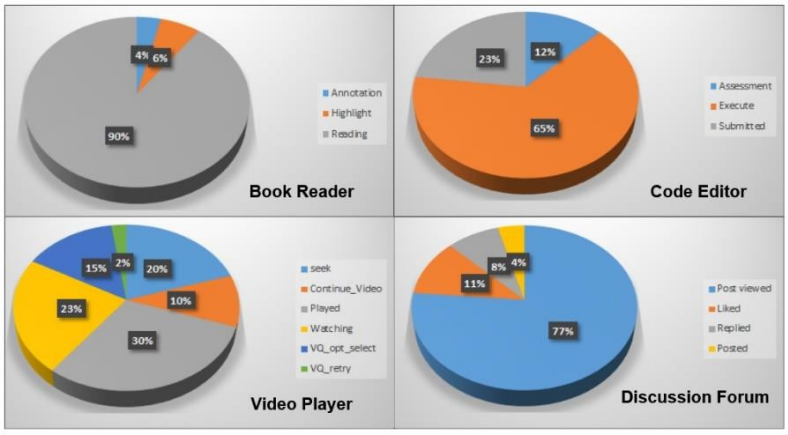


Figure 4. The figure shows the composition of different actions in the four components of PyGuru, namely book reader, video player, discussion forum and code editor.

In video player also, the actions that involve constructive engagement, like answering in-video questions, have a lower frequency than passive watching of videos. Also, the frequency of actions corresponding to active engagement that includes pause and seek behaviour is less. The same behaviour is seen in the case of discussion forum and code editor.

6. Conclusion and Future Work

This paper presents a learning environment called PyGuru for teaching-learning Python programming. PyGuru is based on the ICAP framework and consists of four components: book reader, video player, discussion forum, and IDE. The system captures the click-action of the learners along with the time stamp. This data was analysed to understand student interaction in the system. We found that students mainly interact with the system in passive mode. As suggested from the literature, lower levels of engagement affect students' performance. Hence, it is crucial to either motivate students to engage actively and constructively or design learning activities that promote higher engagement levels.

Since all the click-action of the learners along with the timestamp are collected. This fine-grained data has the potential to capture students' CE using the ICAP framework. Also, this data can be used to provide feedback to the students. For instance, in the case of in-video questions, each incorrect option will provide feedback on why a particular option is incorrect. Similarly, in IDE, learners will be informed about how many test cases are correct. Although the learning environment currently supports all basic functionalities, there are some limitations. For instance, it currently does not have any timer to

keep track of time spent solving a programming problem.

In future, we propose to validate these log data by doing correlation analysis with self-report data. We also plan to investigate learning strategies using sequences of behaviours corresponding to higher modes of engagement. For example, what behaviours do students engage in before and after highlighting? Identifying common video-based learning strategies etc. Methods for measuring learning processes and outcomes in relation to ICAP will also be investigated.

References

Atapattu, T., Thilakaratne, M., Vivian, R., & Falkner, K. (2019). Detecting cognitive engagement using word embeddings within an online teacher professional development community. Computers & Education, 140, 103594.

Carini, R. M., Kuh, G. D., & Klein, S. P. (2006). Student engagement and student learning: Testing the linkages*. Research in Higher Education, 47(1), 1–32. https://doi.org/10.1007/s11162-005-8150-9

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational psychologist, 49(4), 219-243.

Chi, M. T., Adams, J., Bogusch, E. B., Bruchok, C., Kang, S., Lancaster, M., ... & Yaghmourian, D. L. (2018). Translating the ICAP theory of cognitive engagement into practice. Cognitive science, 42(6), 1777-1832.

Dodson, S., Roll, I., Fong, M., Yoon, D., Harandi, N. M., & Fels, S. (2018). An active viewing framework for video-based learning. Proceedings of the Fifth Annual ACM Conference on Learning at Scale. New York, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/3231644.3231682

Duderstadt, J. J., Atkins, D. E., Van Houweling, D. E., & Van Houweling, D. (2002). Higher education in the digital age: Technology issues and strategies for American colleges and universities. Greenwood Publishing Group.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. Review of Educational Research, 74(1), 59–109. https://doi.org/10.3102/00346543074001059

Gorgun, G., Yildirim-Erbasli, S. N., & Epp, C. D. (2022). Predicting Cognitive Engagement in Online Course Discussion Forums. learning, 53(60), 61-62.

Greene, B. A. (2015). Measuring cognitive engagement with self-report scales: Reflections from over 20 years of research. Educational Psychologist, 50(1), 14-30.

Helme, S., & Clarke, D. (2001). Identifying cognitive engagement in the mathematics classroom. Mathematics Education Research Journal, 13(2), 133-153.

Kuh, G. D. (2007). What student engagement data tell us about college readiness.

Wang, M.-T., & Holcombe, R. (2010). Adolescents' perceptions of school environment, engagement, and academic achievement in middle school. American Educational Research Journal, 47(3), 633–662. https://doi.org/10.3102/0002831209361209

Watson, C., & Li, F. W. B. (2014). Failure rates in introductory programming revisited. Proceedings of the 2014 Conference on Innovation & Technology in Computer Science Education - ITiCSE '14. New York, New York, USA: ACM Press. Retrieved from http://dx.doi.org/10.1145/2591708.2591749

Yogev, E., Gal, K., Karger, D., Facciotti, M. T., & Igo, M. (2018). Classifying and visualising students' cognitive engagement in course readings. Proceedings of the Fifth Annual ACM Conference on Learning at Scale. New York, NY, USA: ACM. Retrieved from http://dx.doi.org/10.1145/3231644.3231648

Zhao, Y., Lin, S., Liu, J., Zhang, J., & Yu, Q. (2021). Learning contextual factors, student engagement, and problem-solving skills: A Chinese perspective. Social Behavior and Personality: An International Journal, 49(2), 1–18. https://doi.org/10.2224/sbp.9796

Claudio, D. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Modelling Physical Activity Behaviour Changes for Personalised Feedback in a Health Education Application

Claudio DIAZ [a*], Olivier GALY[b] & Corinne CAILLAUD[c] & Kalina YACEF[a]
[a] *School of Computer Science, The University of Sydney, Australia*
[b] *Interdisciplinary Laboratory for Research in Education, University of New Caledonia, New Caledonia*
[c] *Discipline of Biomedical Informatics and Digital Health, Charles Perkins Centre, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Australia*
*cdia0348@uni.sydney.edu.au

Abstract:
Open-ended domains, where the focus is not about learning specific expert movements but about adopting healthy physical activity behaviours, require the use of unsupervised algorithms and artificial intelligence in education techniques for modelling evolving patterns from physical activity sensor data to enable feedback and personalisation. We present a suite of unsupervised window-based algorithms that detect physical activity changes aligned with learning objectives from accelerometer data. These are translated into learner model attributes and used to generate timely feedback. We illustrate our method in the context of a health education program that teaches adolescents about healthy physical activity behaviours through an application connected to a wrist-worn activity tracker. We present the feedback generated by our algorithms and report on the qualitative evaluation with four experts. We conclude that the automated feedback is useful, important and timely to leverage adolescents' physical activity learning.

Keywords: Physical Activity Tracker, Behaviour Change, Health Education, Personalised Feedback, Learner Model.

1. Introduction

Artificial Intelligence in Education (AIED) systems have focused on the students' interactions with traditional computer interfaces; however, the recent availability of affordable sensors has opened opportunities to capture new interaction types in physical and motor skill domains (Martinez-Maldonado et al., 2017; Santos, 2016), creating the need of novel learner modelling techniques to drive personalisation. In a traditional AIED system, feedback may be based on the knowledge and mistakes made by the student who executes some tasks using a computer. On the other hand, in physical learning systems, feedback may focus on the physical activity (PA) behaviours detected by the sensors. This means that the system needs to extract and assess relevant information from all sensor data in real-time, in addition to the more standard forms of computer-based interactions. Our goal is to create techniques for extracting relevant PA behaviour change data from activity tracker time-series to determine whether students learn and improve their PA behaviours in the context of a health education programme dedicated to 10-12-year-old adolescents called iEngage. iEngage teaches health knowledge and skills related to PA, goal setting and self-assessment of achievements, and aims to promote participation in PA (Yacef, Caillaud, & Galy, 2018). The learning activities are delivered through an application connected to a wrist-worn activity tracker that records steps continuously throughout the programme running period. The activity tracker helps adolescents to monitor their PA towards personal step goals, and supports knowledge acquisition through experiential learning activities. The learning programme delivers evidence-based content about PA and health, complemented by practical activities and quizzes for knowledge consolidation. The activity tracker captures steps and feeds them back to the application. This information is then presented to the adolescents in the application, thus allowing them to self-assess their achievements against their

individual goals. Indeed, learning is not just about how well students know the theoretical health facts, but also about how they translate this knowledge into actual PA behaviours in their daily life.

As the detected patterns will inform the feedback, it is important to capture, model and analyse patterns of PA changes in response to the health education programme. This allows determining whether new PA behaviours are observed in response to what they have learnt, whether these new behaviours are positive or negative, when they occur, and, importantly whether these new behaviours are maintained over time. These data are crucial to provide individualised feedback to the students.

The task is three-fold: (1) to detect changes in PA behaviours using the student's step data continuously recorded by the activity tracker, (2) to model and integrate them into the learner model, and (3) to update these data periodically to give timely personalised feedback. To this aim, we propose a framework and algorithms that we tested using step data and personal PA goals extracted from the e-learning platform from students who previously participated in the iEngage programme. We also validated the value of the automated feedback with experts.

2. Related Work

The emerging use of sensors to support PA and psycho-motor learning has led to the development of specific data analysis techniques for tracking physical behaviour patterns, modelling the learners' behaviour, and providing personalised feedback, for instance, for dance learning (Dias Pereira dos Santos, Yacef, & Martinez-Maldonado, 2017), Aikido training (Santos, 2015), and healthcare training (Echeverria, Martinez-Maldonado, Power, Hayes, & Shum, 2018).

On the other hand, learning about healthy PA behaviours does not involve learning prescribed postures or movements patterns, but achieving an overall goal represented by the recommended PA level (Organization et al., 2020). Moreover, PA changes are specific to each learner and are influenced by their fitness levels and daily schedule. Importantly, PA does not follow a particular pattern, occurs at various times of the day, with various characteristics of frequency, speed, duration and intensity. This implies using unsupervised techniques to capture PA changes during the intervention programme instead of specific activity recognition supervised techniques.

Various unsupervised techniques exist for detecting PA behaviour changes in cohorts or in individuals after an intervention (Lee, Yu, McDowell, Leung, & Lam, 2013; Dobbins & Rawassizadeh, 2015; Diaz, Galy, Caillaud, & Yacef, 2020). However, these methods are not suitable for monitoring PA behaviour changes in real-time. To provide feedback during a health education programme or intervention, live streaming data need to be collected and analysed using window-based techniques. For instance, (Dawadi, Cook, & Schmitter-Edgecombe, 2016) used semi-supervised algorithms to detect PA changes in behavioural routines from smart home sensors data during functional health assessments. Further expanding this work, (Sprint, Cook, & Schmitter-Edgecombe, 2016) created an unsupervised PA change detection framework to track behaviour changes from wearable sensor data. Our work extends these ideas by capturing the characteristics of the students' PA behaviour changes that consolidate their knowledge acquisition.

3. A Learner Model for iEngage

A learner model for physical learning environments needs to include knowledge-specific attributes (i.e. what the student knows) and PA behaviour-specific attributes (i.e. what the student actually does). In iEngage, the knowledge-specific attributes rely on the interactions with the application (quizzes, answers, and set goals), while the PA behaviour-specific attributes are constructed from the data collected by the activity tracker continuously worn by the student.

Using the PA data from the activity trackers and the knowledge-specific data from the e-learning application, we generated learning model attributes using unsupervised algorithms based on rolling window procedures. Then, we used these attributes to produce personalised summative and formative feedback content. Our method is illustrated in Figure 1.

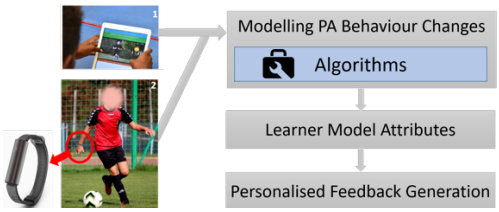


Figure 1. Schematic illustration of the method. Quizzes, answers, and goals are captured by the application (top-left photograph), and real-life PA by the wrist activity tracker (bottom-left photograph). These data are processed by algorithms and stored in the learner model as attributes to generate personalised feedback.

3.1 Algorithms for Modelling PA Behaviour Changes

We built four unsupervised algorithms to detect PA behaviour changes from step data. Detecting PA behaviour changes requires comparing present and past behaviours to identify significant differences. As PA behaviours are individual and dynamic, the algorithms use moving time windows that iteratively goes through time frames of each student's PA behaviour and analyses the presence of changes and their nature.

Difference Scanner (DS): Detects whether a student performed an hourly PA behaviour change by comparing it with the cohort's hourly PA behaviour (baseline). Using the hourly steps, it indicates when students do less PA than their peers. Specifically, the student's number of steps per hour is compared with the cohort's median number of steps per hour. This allows identifying the hours when the student does fewer steps than the cohort. The algorithm returns a data frame with two vectors: the day and the time when the difference is detected.

Rolling Window Linear Regression (RWLR): Detects recent daily PA behaviour changes. Using the number of steps per day, the algorithm measures the student's PA trend during the previous days. Specifically, it uses a 3-day window to scan the students' time series of daily steps, and calculates iteratively the linear regression for each rolling window. It returns a data frame with two vectors: the day where the window is located and the resulting linear regression coefficients.

Window-Based Interquartile Range (WBIQR): Detects recent hourly PA behaviour changes. Using the hourly steps, this algorithm labels the student's latest significant PA differences. Specifically, a time window rolls through the number of steps and time spent doing moderate to vigorous PA (MVPA) per hour. For each rolling window, the algorithm calculates the mean step number and MVPA per hour. Then, it subtracts the present-day hourly steps and MVPA from the hourly mean step number and MVPA of the windows, to generate the hourly differences of step number and MVPA. These differences are boxplotted and the IQR limits are used to detect outliers that are labelled as significant behaviour changes. It returns a data frame with eleven vectors: the date and time when behaviour changes were detected, the mean hourly step number and MVPA, the difference between the detected behaviour changes and the hourly mean step number and MVPA, the upper and lower IQR limits, and the length of the rolling window used.

Multi-Shifted Window-Based IQR (MSWBIQR): Detects PA behaviour changes that are sustained over time. Using the previously detected hourly behaviour changes, it labels as habits the changes that are repeated consecutively. Briefly, it scans the upper and lower limits of the previously detected behaviour changes from the WBIQR algorithm through the time series of the hourly step difference. If the hourly steps difference is outside the limits of the previously detected behaviours, then the behaviour change is sustained. It returns a data frame with three vectors: the sustained behaviour change date, time, and the number of days of uninterrupted behaviour change.

3.2 Learner Model Attributes

The main topics of iEngage were related to the learning model attributes to capture knowledge and specific patterns of PA behaviour. These attributes use outputs from the algorithms described in the previous section. Figure 2 lists the attributes and maps them to the iEngage topics and to the algorithms.

ID	Attribute	Attribute Description	iEngage Topic	Algorithm Used
1	Step Goals Achieved	Number of days when the step sum ≥ daily step goal	Recommended Physical Activity	DS
2	MVPA Goals Achieved	Number of days when the sum of minutes spent in MVPA per day > daily MVPA goal	Recommended Physical Activity	DS
3	Step Change	Indicates whether the daily number of steps has increased since the previous module	Be Physically Active	DS
4	Step Trend	Trend (β1) of the step number	Be Physically Active	RWLR
5	Positive Step Behaviour Changes	Number of days in which positive step behaviour changes are detected	Healthy and Sedentary Behaviours	WBIQR
6	Negative Step Behaviour Changes	Number of days in which negative step behaviour changes are detected	Healthy and Sedentary Behaviours	WBIQR
7	Positive Behaviours, Step Increase	Difference between the step number in positive steps behaviour changes and the mean step number	Healthy and Sedentary Behaviours	WBIQR
8	Negative Behaviours, Step Loss	Difference between the step number in negative step behaviour changes and the mean step number	Healthy and Sedentary Behaviours	WBIQR
9	Daily MVPA Change	Indicates whether the daily minutes spent in MVPA has increased since the previous module	PA Intensity Levels	DS
10	Daily MVPA Trend	Trend ($\beta 1$) of time spent in MVPA	PA Intensity Levels	RWLR
11	Positive MVPA Behaviour Changes	Number of days in which positive MVPA behaviour changes are detected	PA Intensity Levels	WBIQR
12	Negative MVPA Behaviour Changes	Number of days in which negative MVPA behaviour changes are detected	PA Intensity Levels	WBIQR
13	Positive Behaviours, MVPA Increase	Difference between the minutes spent in MVPA during positive MVPA behaviour changes and the mean minutes spent in MVPA	PA Intensity Levels	WBIQR
14	Negative Behaviours, MVPA Loss	Difference between the minutes spent in MVPA during negative MVPA behaviour changes and the mean minutes spent in MVPA	PA Intensity Levels	WBIQR
15	Longest Positive Habit	Positive step behaviour change that has been maintained for the highest number of days	Healthy and Sedentary Behaviours	MSWBIQR
16	Opportunity to Increase Steps	Day of maximum difference in step number between the student and the cohort	Recommended Physical Activity	DS
17	Follow-up	Number of days the adolescent did more steps than recommended (Attribute ID 16)	Recommended Physical Activity	DS

Figure 2. Map of the attributes to the iEngage topics and the algorithms used to calculate them.

3.3 Personalised Feedback Generation

We used the learner model attributes to generate timely personalised feedback to support the students' learning, motivation, and healthy PA behaviour changes. The feedback is generated by a combination of feedback types which are triggered by specific attribute values. Figure 3 show the list of feedback types and a sample of the rules used to trigger the feedback.

3.4 Illustration

To illustrate the personalised feedback each student receives, we sampled data from a real participant in a previous iEngage intervention (Yacef, Caillaud, & Galy, 2018) to generate the feedback received on a particular day. The values of each learner model attribute were calculated by our algorithms using the sampled data to trigger the feedback rules and to build the personalised feedback. This is presented to the participant as a narrative in an educational, constructive and adolescent-friendly way (Figure 4).

ID	Feedback types
A	Congratulate and encourage steps
B	Find positive changes to reinforce steps
C	Congratulate for steps goal
D	Find positive steps trends and reinforce them
E	Find positive step behaviours to reinforce them
F	Congratulate for achieving a positive habit
G	Promote following opportunity suggestions
H	Suggest opportunities for more PA during the day
I	Find negative behaviours to fix
J	Congratulate and encourage MVPA
K	Find positive changes to reinforce MVPA
L	Congratulate for MVPA goal
M	Find positive MVPA trends and reinforce them
N	Find positive behaviours of MVPA to reinforce them
O	Find negative behaviours to fix
P	State steps non-achievement
Q	State MVPA non-achievement

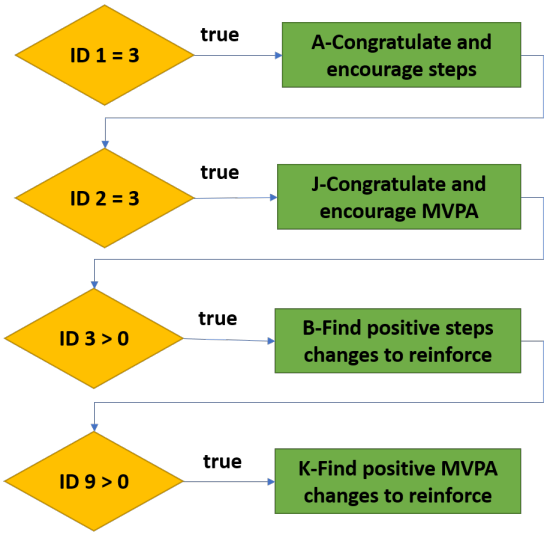


Figure 3. List of feedback types (left table) and sample of feedback trigger rules according to the combination of attributes values (right image).

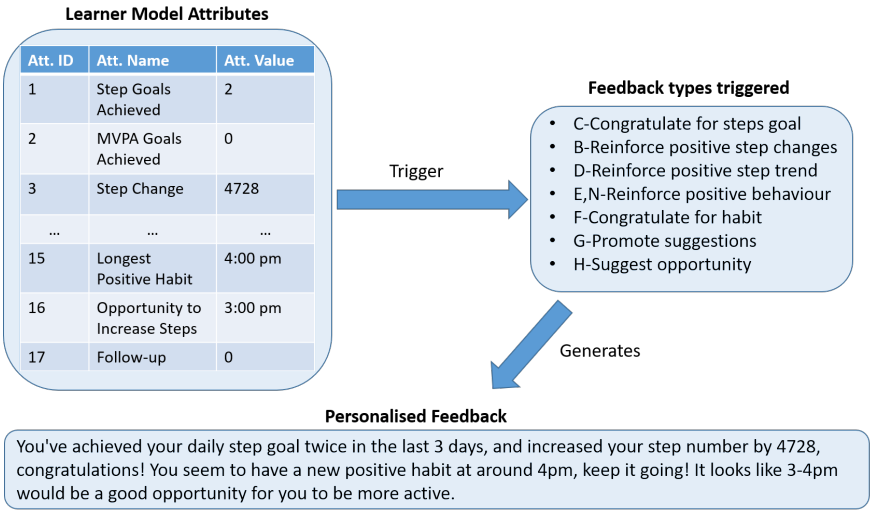


Figure 4. Illustration of the resulting personalised feedback using data of a real participant for a particular day.

4. Validation

We recruited four health sciences, exercise physiology, or physical education experts with experience working with adolescents. We asked them to evaluate via an online survey whether the automated personalised feedback was valuable (across 3 dimensions: helpful, important, and timely) for supporting adolescents who learn how to modify their PA behaviour. This online survey presented real PA data from 3 adolescents from a previous iEngage education session (Yacef, Caillaud, & Galy, 2018). The experts were asked to evaluate and comment on the feedback automatically generated by our algorithms on 3 days. This survey was approved by The University of Sydney Human Research Ethics Committee (2021/815). Each expert evaluated 44 automated feedbacks. The resultant 176 answers (4 experts * 44 feedbacks) positively evaluated the automated personalised feedbacks as helpful (mean=88%, s.d.= 11%), important (mean=85%, s.d.= 7%), and timely (mean=85%, s.d.= 8%). Experts were also asked to provide free text comments about the automated feedbacks. They found the automated feedback was good support for the students' learning process; however, they highlighted the importance of using an adolescent-friendly language in a live scenario that would need to be tailored

depending on the health intervention population demographics. They also suggested students should be incorporated into the feedback loop to increase their motivation and propose more relevant actions. The study shows that valuable personalised formative feedback can be automatically generated using a combination of knowledge-specific attributes mapped with the health education application, and PA behaviour-specific attributes extracted from the activity tracker.

5. Conclusion

We addressed the problem of adding the PA accelerometer time series data into a learner model for a health education system by providing algorithms that (1) detect PA behaviour changes relevant for knowledge acquisition by students, (2) build relevant learner model attributes, and (3) provide timely personalised feedback. Window-based algorithms provide an effective way to monitor individual's PA behaviour changes over time through relevant learning-related PA behaviour attributes mapped to the learning contents. Although the actual feedback was not implemented live in our system, we show what type of feedback can be generated using these algorithms and real data from a digital health education programme that teaches health knowledge and skills to promote PA behaviour change in adolescents. Health education experts appraised that the personalised feedbacks generated by our algorithms were useful, important, and timely, indicating it is valuable for health promotion and PA behaviour change.

Acknowledgements

Claudio Diaz acknowledges Universidad Adolfo Ibáñez and CONICYT 'Becas Chile' Doctoral Fellowship program Grant No. 72200111.

References

Dawadi, P. N., Cook, D. J., & Schmitter-Edgecombe, M. (2016). Modeling patterns of activities using activity curves. Pervasive and Mobile Computing, 28, 51-68.

Dias Pereira dos Santos, A., Yacef, K., & Martinez-Maldonado, R. (2017). Let's dance: How to build a user model for dance students using wearable technology. In Proceedings of the 25th conference on user modeling, adaptation and personalization (p. 183–191). New York, NY, USA: Association for Computing Machinery.

Dobbins, C., & Rawassizadeh, R. (2015). Clustering of physical activities for Quantified Self and mhealth applications. 2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (October), 1423–1428.

Diaz, C., Galy, O., Caillaud, C., & Yacef, K. (2020). A clustering approach for modelling and analysing changes in physical activity behaviours from accelerometers. IEEE Access, 8, 224123-224134.

Echeverria, V., Martinez-Maldonado, R., Power, T., Hayes, C., & Shum, S. B. (2018). Where is the nurse? towards automatically visualising meaningful team movement in healthcare education. In C. Penstein Ros´e et al. (Eds.), Artificial intelligence in education (pp. 74–78). Cham: Springer International Publishing.

Lee, P. H., Yu, Y.-Y., McDowell, I., Leung, G. M., & Lam, T. (2013). A cluster analysis of patterns of objectively measured physical activity in hong kong. Public Health Nutrition, 16(8), 1436–1444.

Martinez-Maldonado, R., Yacef, K., Dias Pereira Dos Santos, A., Buckingham Shum, S., Echeverria, V., Santos, O. C., & Pechenizkiy, M. (2017). Towards proximity tracking and sensemaking for supporting teamwork and learning. In 2017 ieee 17th international conference on advanced learning technologies (icalt) (p. 89-91).

Organization, W. H., et al. (2020). Who guidelines on physical activity and sedentary behaviour: at a glance.

Santos, O. C. (2015, 01). Education still needs artificial intelligence to support personalized motor skill learning: Aikido as a case study. In (Vol. 1432, p. 72-81).

Santos, O. C. (2016). Training the body: The potential of aied to support personalized motor skills learning. International Journal of Artificial Intelligence in Education, 26(2), 730-755.

Sprint, G., Cook, D. J., & Schmitter-Edgecombe, M. (2016). Unsupervised detection and analysis of changes in everyday physical activity data. Journal of Biomedical Informatics, 63, 54 - 65.

Yacef, K., Caillaud, C., & Galy, O. (2018). Supporting learning activities with wearable devices to develop life-long skills in a health education app. In Artificial intelligence in education (pp. 394–398). Cham: Springer International Publishing.

Kang J, et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Research on Classroom Interaction Behavior in a Mobile Tablet-based Unit Teaching Setting

Jian KANG, Ruixue LIU*
School of Educational Technology, Northwest Normal University, China
*isnow0211@163.com

Abstract: In February 2022, the Ministry of Education issued the *Key Points of the Ministry of Education's Work in 2022*, pointing out the importance of implementing educational digitalization strategy. To implement the requirements of China's policy and newly released discipline curriculum standards, this study developed a mobile tablet-based unit teaching setting and explored its effects on classroom interaction behavior. The research sample comprised 45 middle school students from western China. Combined with the Improved Flanders-type Analysis System, two Chinese course videos were analyzed before and after implementation. Results show that the mobile tablet-based unit teaching setting had positive effects on the quality of teacher-student interaction, student-student interaction, student-technology interaction, and the classroom interaction quality. The conclusions provide practical teaching guidance and references for various subjects' unit teaching implementation under the context of the educational digitalization transformation.

Keywords: Mobile tablet, Unit teaching, Classroom interaction behavior

1. Introduction

Rose (1999) argued that interaction often has the duality of privilege and neglect. In the teaching environment, classroom interaction behavior refers to the transmission of teaching information between several elements of classroom teaching (Sun, 2021). It can be found that previous studies provide various definitions of interaction generally as well as classroom interaction specifically. In this study, we define classroom interactive behavior as various transactions between teachers, students, teaching media, and resources. Example lists are teachers imparting knowledge, students' use of information technology, and students' active speech.

Many researchers have examined classroom interaction in different environments. For example, Balaji (2010) pointed out that adding an online discussion module to a traditional classroom can encourage students to participate more actively in classroom interaction. Another study found that the electronic whiteboard environment promotes interaction between students and improves teaching efficiency (Qi, 2011). Moreover, Fang et al. (2012) proposed an interactive analysis system that improved Flanders Interaction Analysis System (iFIAS) based on a digital classroom teaching environment. Though these previous studies provide important insights in terms of classroom interaction in various environments, there is still a lack of research on exploring class interaction behavior with the iFIAS method of learning behavior, particularly in the context of mobile tablet-based settings. Therefore, this study developed a mobile tablet-based unit teaching setting, and aimed to use the methods of iFIAS to answer the following questions:

1. Does the mobile tablet-based unit teaching setting elicits better classroom interaction behavior?

2. What kind of classroom interaction behavior changes significantly in a mobile tablet-based unit teaching setting?

2. Development of the Mobile Tablet-based Unit Setting

In this part, we present detailed information about the developed mobile tablet-based unit environment, to explore classroom interaction behavior effects.

2.1 Mobile Tablet-based Unit Teaching Setting Overview

Based on learning theories such as constructivist learning theory and zone of proximal development. T his study selected middle school Chinese as a subject and developed the mobile tablet-based Chinese p rogressive unit teaching setting to promote classroom interaction behavior and improve the classroom teaching effect. This setting uses online learning tools and resources and integrates both of them into t he learning activities, which are organized and conducted through nine steps, to be conducive to the op timization of classroom interaction behavior, as shown in Figure 1.

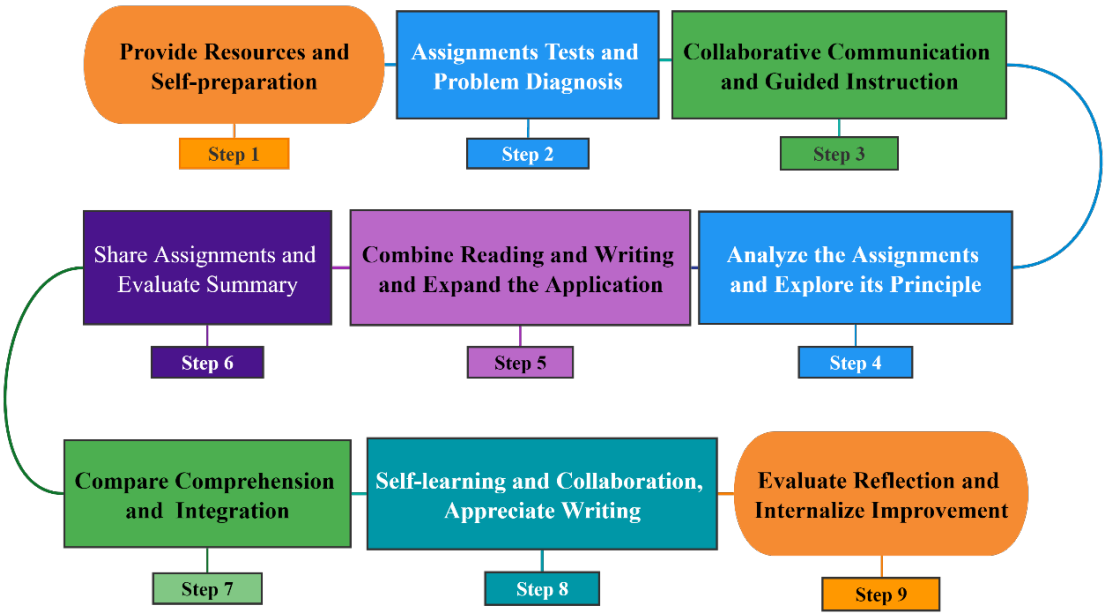


Figure 1. The Mobile Tablet-based Unit Teaching Setting

Among them, Step 1 is designed to help students complete the preview by using study task sheets and online resources. Step 2 is used to understand the students' basic situation for learning new knowledge from the text, such as basic word recognition, author background introduction, etc. Step 3 helps students gain high understanding levels of the new knowledge. Step 4 analyzes the works and explores the rules based on consolidating and deepening learning, aiming to achieve a level of understanding of new knowledge. Step 5 mainly focuses on practical application and evaluation, to detect the creative level of students' mastery of methods and laws. Step 6 guides students to continuously revise and improve the learning results. To promote students to form strategic knowledge and metacognitive knowledge in the process of practice, reflection, and summarization. Step 7 is the cultivation of the ability to comprehend unit knowledge, which aims to improve the student's overall cognitive structure of the unit. Steps 8 and 9 are comprehensive applications and evaluations of unit knowledge in real situations.

To further verify the effect of classroom interaction behavior under the mobile tablet-based subject unit teaching setting, this study conducted an empirical study.

3. Methodology

3.1 Participants

The study sample included 45 middle school students (25 boys and 20 girls) and a female teacher who taught middle school Chinese courses in Wuzhong City. The students had not been exposed to a mobile tablet-based subject unit teaching environment before, nor had they been exposed to the course content, by the study requirements.

3.2 Classroom Environment

The school completed the construction of the information technology teaching environment in June 2018. The classrooms are equipped with interactive electronic whiteboards, and mobile networks, and each student is equipped with a mobile tablet. The classroom mainly uses teaching assistant software such as "Class Optimization Master." This software can provide teaching functions such as random grouping, random questions to students, group grading, real-time display of student content, distribution and sharing of various resources, etc. The teaching environment is shown in Figure 2.



Figure 2. The Mobile Tablet-based teaching environment

3.3 Course Content

The texts in the sixth unit of the first volume of the eighth grade are all classics from ancient countries. This unit aims to cultivate students' ability to independently read classical Chinese using annotations and reference books, practice reading, and accumulate knowledge of classical Chinese. The texts in the sixth unit of the second volume of the eighth grade are all traditional masterpieces. The theme is the philosophical thoughts and emotions of the characters. This unit also aims to cultivate students' ability to read classical Chinese, practice reading, and accumulate knowledge of classical Chinese. Therefore, the themes of the two units are similar and the requirements are consistent. The two lessons collected in this study are the curriculum content of the sixth unit of the eighth-grade Chinese volume, "Born in distress, died in peace," and the sixth unit of the second volume, "Zhuangzi and Huizi in the trenches." The content of the courses is roughly similar and meets the experimental requirements.

3.4 Procedure

The experimental process is as follows: A 40-minute classroom teaching of "Born in distress, died in peace" was video recorded in a traditional teaching setting. The new model was adopted to teach "Zhuangzi and Huizi in the trenches." and another video of about 40 minutes was recorded, the two videos were then analyzed.

3.5 Instrument

To grasp the classroom interaction behavior and its changes, this study adopts the iFIAS. specifically, it includes four categories: Teacher language, Student language, Stillness, Technology, and 14 behaviors.

3.6 Data Analysis

After finishing the video recording, we cut out footage that was not related to the classroom b efore and after and kept two classroom videos (36 minutes each) as research samples. Classro om interaction behavior was encoded. The proportion of 14 behaviors after two-dimensional matrix statistical analysis according to the iFIAS coding system is shown in Table 1.

Table1. *Statistical comparison of interactive behavior coding two lessons*

Category		Coding	Specific behavior	*Born in distress, died in peace* Coding	*Zhuangzi and Huizi in the trenches* Coding
Teacher language	Indirect influence	1	Teachers accept feelings	0.70%	0%
		2	Teachers praise or encourage	0.56%	1.94%
		3	Teachers adopt students' views	1.11%	2.22%
		4	Teacher questions 4.1 Ask open questions 4.2Ask closed questions	5.56%	3.05%
		5	Teacher teaching	27.4%	3.47%
	Direct influence	6	Teacher instruction	7.37%	11.37%
		7	Teachers criticize or defend teachers' authority	0%	0%
Student language		8	Students' passive response	19.05%	4.58%
		9	Students speak actively 9.1 Students' active response 9.2 Learn to ask questions actively	0%	17.34%
		10	Discussion between students and peers	0%	32.18%
Stillness		11	It doesn't help the confusion of teaching	0.97%	1.39%
		12	Conducive to the silence of teaching	21.42%	1.53%
Technology		13	Teacher manipulation technology	15.86%	6.24%
		14	Student manipulation technology	0%	14.7%

4. Results

The iFIAS analysis tool was used to analyze the statistical proportions in the cases. For example, the first seven behaviors were all teacher language. The proportion calculation formula is shown in the next column. Statistical analysis was carried out on fifteen items including the silent proportion of teaching, the statistical items, and calculation formulas shown in Table 2.

Table 2.*Coding statistics of the proportion of interactive behaviors in two classes*

Statistics	Calculation formula	*Born in distress, died in peace* Coding	*Zhuangzi and Huizi in the trenches Coding*
Teacher language ratio	$\sum_{i=1}^{7}$ tally(i)/Total	42.7%	22.05%
Student language ratio	$\sum_{i=8}^{10}$ tally(i)/Total	19.05%	54.09%
The proportion of silence that is beneficial to teaching	tally(12)/Total	21.42%	1.53%
Technology application ratio	tally(12)/Total	15.86%	20.94%

The ratio of indirect influence to direct influence on students in teacher language	$\sum_{i=2}^{4}$ tally(i)/ $\sum_{i=5}^{7}$ tally(i)	20.8%	48.6%
The ratio of positive reinforcement to negative reinforcement to students in teacher language	$\sum_{i=2}^{3}$ tally(i)/ $\sum_{i=6}^{7}$ tally(i)	22.64%	36.59%
The proportion of questions asked in the teacher's language	$tally\ (4)\ / \sum_{i=1}^{7}$ tally(i)	13.03%	13.84%
The proportion of students actively speaking in the student language	$tally\ (9)\ / \sum_{i=2}^{10}$ tally(i)	0%	22.77%
The proportion of teachers manipulating technology in the application of information technology	$tally\ (13)\ / \sum_{i=13}^{1014}$ tally(i)	100%	29.8%
The proportion of students who manipulate technology in the application of information technology	$tally\ (14)\ / \sum_{i=13}^{1014}$ tally(i)	0%	70.2%
Teachers ask open-ended questions as a percentage of teachers' questions	tally(4.1) /tally(4)	0%	72.73%
Teachers ask closed questions as a percentage of teachers' questions	tally(4.2) /tally(4)	100%	27.27%
The proportion of students who actively responded to students who actively spoke	tally9.1) /tally(9)	/	57.6%
The proportion of students taking the initiative to ask questions to the students taking the initiative to speak	tally(9.2) /tally(9)	/	42.4%
The proportion of students who responded actively to the students who responded	$\frac{\text{tally(9.1)}}{\text{tally(8)}}$ + tally(9.1)	0%	68.57%

As shown in Table 2, the proportion of teacher language and student language in the classroom changed significantly after the practice, which shows that the classroom interaction behavior under the mobile tablet-based Chinese unit teaching setting underwent significant changes. There were significant increases in the proportion of students' language (from 19.05% to 54.09%), the proportion of students manipulating technology in classroom teaching, which rose to 70.2%, the proportion of indirect influence and direct influence on students of teacher language (from 20.8% to 48.6%), the proportion of students taking the initiative to speak (from 0% to 57.6%), and so on.

5. Discussion and conclusions

This study developed a mobile tablet-based unit teaching setting, in the context of middle school Chinese subjects. Using the iFIAS method for analyzing two videos (students in the same class in traditional mode and new mode, two lessons with the same teacher), we conclude the following findings.

5.1 Teacher-student Interaction, Student-student Interaction Changes

Under the mobile tablet-based Chinese subject unit teaching setting, the proportion of students' language increased significantly (from 19.05% to54.09%). The data show that in the mobile tablet-based subject unit teaching setting, classroom interaction behavior changed significantly, teacher-student interaction and student-student interaction behavior significantly improved, and the classroom changed from being teacher-led to both student-centered and teacher-led.

5.2 Interaction with the Media Improved

In traditional setting, the proportion of students operating technology was rare. After implementing the mobile tablet-based subject unit teaching setting for one semester, the proportion of students who used technology in classroom teaching rose to 70.2%. The proportion of students using technology has also increased significantly (from 0 to14.7%). indicating that students' use of information technology also improved under the mobile tablet-based subject unit teaching setting.

5.3 Improvement in the Quality of Classroom Interaction

The proportions of indirect and direct influences on students of teacher language increased significantly (from 20.8% to 48.6%), and the ratio of positive reinforcement and negative reinforcement in teacher language to students also increased noticeably (from 22.64% to 36.59%), the proportion of teachers asking open-ended questions increased significantly (from 0% to 72.73%), and so on. It can be concluded that the mobile tablet-based subject unit teaching setting is more conducive to allowing the lecturer to control the balance of classroom interaction, mobilize the enthusiasm of students to answer questions, and thus promote a dynamic balance of classroom interaction and improve classroom teaching efficiency, implementing the new curriculum requirements.

6. Limitations and future research

This study developed a mobile tablet-based unit setting under the context of a middle school Chinese subject and explored its impact on classroom interaction behavior. Combined with the improved iFIAS method. The results show that the mobile tablet-based unit setting had a positive impact on the quality of classroom interaction. The research conclusions provide practical teaching guidance and reference for the implementation of education digitization strategy in specific subject unit teaching.

However, the current study had certain limitations. First, the teaching content before and after the experiment was different. Second, quantitative analysis of teachers' subject teaching ability and students' learning level has not been carried out. Further research should select similar course content and add a qualitative method to explore the educational effect of this new setting.

Acknowledgment

This research was supported by the Gansu Province Educational Technology Innovation Project: Research on the interactive behavior of teachers and students in the delivery classroom under the background of "Internet + teaching support" No. 2021CXZX-338. 2022 Young Teachers Research Ability Improvement Program in Northwest Normal University (Project No. NWNU-SKQN2022-18).

References

Balaji, Diganta, C. (2010).Student Interactions in Online Discussion Forum: Empirical Research from media Richness Theory' Perspective. *Journal of Interactive Online Learning*,(9),5-22.

Fang, H., Gao, C., Chen, J.(2012). Improved Flanders Interactive Analysis System, and Its Application. *China Electronic Education*, (10),109-113.

Qi, L. (2011). Using the electronic whiteboard to effectively promote the interaction between students and students in the classroom. *Educational Informatization in China*, (14),41-44.

Rose, E. (1999). Deconstructing Interactivity In Educational computing, *EdcuationalTechnology*,39(1),43-49.

Sun, X. (2021). Research on the characteristics of teacher-student interaction ineffective Chinese classroom teaching in primary school. *Liaoning Normal University Master's Thesis*.

Zhao, F. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

A Technique for Tracking the Reading Rate to Provide Students' Learning Feedback

Fuzheng ZHAO[a] & Chengjiu YIN[b*]
[a]*Graduate School of System Informatics, Kobe University, Japan*
[b]*Information Science and Technology Center, Kobe University, Japan*
*yin@lion.kobe-u.ac.jp

Abstract: Reading rate is the number of words a user reads per unit of time, also known as reading speed. As online learning has been continuously growing, research on tracking learning behavior is undertaken, and the reading rate was beginning to be applied to measure the learning process. However, current calculating methods are based on the traditional learning environment of a paper-based course, and further research is needed to determine whether they can be used to track students' learning in an online learning environment and integrate with their online reading habits. To this end, this study proposed a new method based on online learning used to calculate reading rate as a measure to track students' learning status, after analyzing students' reading behavior in the e-book learning system. This method not only can identify students with learning difficulties through reading rate outliers, but also provide urgent feedback to students on their reading status.

Keywords: Reading rate, learning tracking, urgent feedback

1. Introduction

Accurately and timely understand students' learning status, discover abnormal learning habits and intervene, which puts forward higher requirements for evaluating the learning process. Many studies have proved that reading comprehension is correlated with learning outcomes (Crawford, Tindal, & Stieber, 2001), and the reading rate is one of the important indicators to measure reading comprehension. However, frequently used tests like classroom oral reading rate in the classroom (Deno & Mirkin, 1977) cannot detect students' learning status in real-time (Skinner et al., 2009).

With the continuous application of online learning environments and learning methods, the popularity of digital teaching materials has further accelerated the development of online reading (Zhao, Hwang, & Yin, 2021). Besides, studies have shown that students who use online reading achieve better reading comprehension than paper reading (Huang, 2014). Especially, online reading is capable to provide a reliable reading tool and learning environment that gives timely feedback. But, Vonderwell (2003) reported that students' greatest disappointment with online learning was the lack of communication and a sense of community, and Rovai (2002) proved that when students encounter difficulties, this sense of community is rewarded if teachers detect and help them.

To this end, this study will use the characteristics of the online learning environment to design a reading rate calculation method, which is used to detect students' reading rate behaviors in real-time, help teachers understand students' reading status, and find abnormal reading behaviors.

2. Literature Review

2.1 Tracking Learning by Reading Rate

Reading rate refers to the number of words read per unit time, also known as reading speed (Carver, 1992), which has been used as an independent variable to determine students' learning skills. While, various reading rate is correlated with reading errors and knowledge comprehension (Breznitz & Berman, 2003). The conventional view is that humanity's reading speed varies with the difficulty of the material and the purpose of the reader (Brysbaert, 2019). Besides, reading rate is often quantified by

reading speed, which is the number of words read correctly per minute or the length of time it takes a reader to complete an article (Hudson, Lane, & Pullen, 2005). Notably, the basic measure of reading rate is done by reading at regular intervals. Specifically, during the test, students have to complete the reading materials arranged by the teacher, to judge the number of words they read correctly and the number of errors in a certain period.

2.2 Relationship between Reading Rate and Learning Comprehension

The main function of reading is comprehension. Carver (1992) considered the reading rate to reflect the ability of word recognition and reading comprehension. Within this context, the reading rate is used as a diagnostic tool for dyslexia. Since slow reading leads to diminished comprehension, many interventions focus on improving reading speed (Bernacki, Byrnes, & Cromley, 2012). On the other hand, it is important to keep abreast of students' reading rates. Although a measure of reading speed is highly correlated with reading comprehension, it does not contain a direct measure of comprehension. Crawford, Tindal, and Stieber (2001) used reading rate to predict student performance on a national test consisting of reading and mathematics, and through correlation analysis, only validated a good way to monitor student progress by measuring reading rate.

2.3 Reading Rate Calculation

Current reading rate calculation methods include oral and non-verbal reading rates. The oral reading rate is divided into correct word rate, question-answer correct rate, and cloze correct rate, depending on the calculation method. Specifically, correct word rate is measured as the number of words read correctly per minute (Skinner et al., 2009). In the measurement, students are asked to read aloud a given passage within 1 minute, and teachers rate the accuracy of reading each word. Thus, the correct word rate is a speed measure that includes accurate speaking and reading speed (Good & Kaminski, 2002). In addition, to measure comprehension accuracy during reading, students answer pre-designed comprehension questions after reading a text passage aloud and measure the percentage of text per minute reading comprehension, which is referred to as the question response rate (Daly, Chafouleas, & Skinner, 2005). It is calculated by dividing the correct question rate by the number of seconds it takes to read the text and then multiplying by 60 seconds. Finally, the correct completion question rate is measured by asking students to select the correct answer from the options provided after reading aloud (Ridge & Skinner, 2011). Although the completion test is still widely recommended in instructional methods texts, it is also dependent on writing skills and can be time-consuming and frustrating. By contrast, a measure of non-verbal reading rate is a measure of reading speed, i.e. the number of words a student reads in one minute. This method has been used as a technique to track student reading and indirectly assess student reading comprehension (Liang & Huang, 2014).

3. Reading Rate Calculation Based on Online Learning Behavior

3.1 Online Reading Behavior

Currently, online reading is done using a learning management system, such as Moodle, using a dedicated e-book reading system, or a hardware device such as a reading kinder. Unlike paper-based reading, learning strategies or styles vary due to changes in a learning environment and reading tools, such as reading back during reading (Yin et al., 2019), skipping during the reading review (Yin et al., 2015), and repetitive manipulation of reading tools (Zhao, Hwang, & Yin, 2021). At the same time, these behaviors in the online learning environment are related to academic performance, learning style, and learning pressure. For this reason, this study will utilize reading behaviors in the online reading environment, such as page-turning, view manipulation, highlighting, underlining, and making memos, as one of the main bases for reading calculation methods. This choice mainly takes into account the e-book reading system, which can provide a wealth of reading tools such as turning pages, changing views, and annotating. In addition, the e-book reading system can record students' reading behavior in

the form of a log for easy data analysis. From the convenience of data collection, this study uses the e-book system developed by Yin et al.(2019) to collect data on students' reading behavior.

3.2 Calculation Method

The main idea of the reading rate calculation method is to select the features that best represent the occurrence of reading behavior, after which the core elements of reading are extracted from these reading behavior features, the core elements include time, word count, and page count. The specific reading rate calculation method is shown in Figure 1. The reading rate calculation method proposed in this study focuses on calculating the speed of reading material and reading content: calculating the speed of reading material means calculating the reading speed of students when using the page-turning tool; calculating the speed of reading content is obtained by calculating the reading speed of students when using underlined/highlighted annotated text.

The formulae were categorized based on the three learning strategies: read all the words in the material (RAW), read some of the words in the material (RSW), and did not read any words (RNW). These three learning methods are judged by comparing the number of words read by the student with the total number of words per page in the reading material. If the difference is equal to zero, then the student has read all of the words on that page; if it is less than zero, then the student has read some of the words on that page; if the difference is equal to the total number of words on no page of the reading material, then the student has not read any words.

RAW calculation. When the student has read all the words, the reading rate is equal to the total number of words on each page of the reading material divided by the time spent.

RNW calculation. When no text was read, the reading rate was zero.

RSW calculation. When the student reads some of the text, calculating the reading rate further involves determining whether the student is using underlining/highlighting when reading. In the case of underlined/highlighted reading, the reading rate for a single reading operation is first calculated as the total number of words marked in a single reading divided by the time spent. The average reading rate is then calculated as the sum of the single reading rates divided by the total number of reading operations. In the case of no underlined/highlighted reading, the reading rate is equal to the total number of untagged words divided by the total time spent reading minus the difference in time spent reading without underlining/highlighting. Finally, the size of the reading rate in these two cases is determined. If either reading rate is greater than 50%, it is considered to be the dominant reading rate.

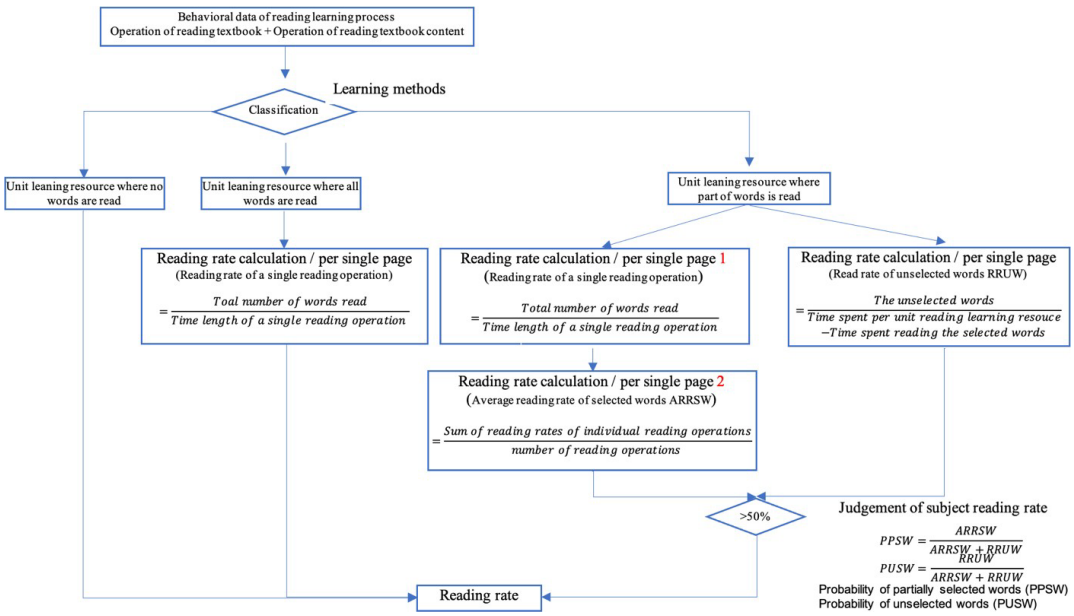


Figure 1. Reading Rate Calculation Flowchart

3.3 Data Collection and Processing

The reading rate calculation method proposed by this study requires the students' operation behavior number, operation behavior time, number of words read, and number of pages of reading materials. First, the data collection is done by the e-book system, and the data is stored in the database in the form of log data, as shown in Table 1. "ID" is the student's number, and "OperationName" represents the student's operation behavior: including specific operation behaviors, such as adding underscores; "id: 23" indicates the number of the operation behavior; "Makertext" indicates the number of words involved in the operation behavior, such as students using underscores to read a total of two words "more attention". "Operation data" represents the time when the student performed the operation behavior. "E-book" is a reading material. "Device" is the device used by the student. "Pages" indicate the total number of pages of the reading material.

Table 1. *Log Data of Reading Behavior Collected from E-book System*

ID	Operation name	Operation date	E-book	Device	Pages
1	ADD Highlight id:23 page:5 markertext: more attention	2021-05-19 11:00:48.329	23	pc	49
1	ADD Underline id:24 page:5 markertext: address some problems	2021-05-19 11:00:55.061	23	pc	49
1	ADD Memo id:46 page:5 markertext: make reflections	2021-05-19 11:01:44.326	23	pc	49

Data processing and analysis use Python, including Xlrd, Xlwt, Openpyxl, and Pandas modules. First, use the Pandas module to remove invalid and missing data. Second, use Xlrd, Xlwt, and Openpyxl to split the data of "OperationName", the split data types include the specific name of the reading operation and its number, the page number where the reading operation occurred, and the number of words read, and the total number of words. Finally, the Datetime module calculates the duration of the read operation behavior. The final results are shown in Table 2. Among them, "BehaviorN" is the name of the reading operation, and "PageN" is the page number where the reading operation is located. "MarkingT" is the number of words read, and "TextS" is the sum of the number of words read. "Time" is the start time of the read operation. "Timedelta" is the duration of the read operation. "WordN" is the sum of all the word counts of the page number on which the read operation occurs.

Table 2. *Processed Behavioral Data for Calculating Reading Rate*

ID	BehaviorN	PageN	MarkingT	TextS	Time	Timedelta	WordN
1	AddHL	1	Digital Learning	2	11:13:19.565	0:04:53.050	72
1	DeleteHL	1	Num	0	11:13:33.919	0:00:00.994	0
1	AddUN	1	Education	1	11:13:45.628	0:00:33.725	514
1	Next	2	Num	0	11:25:58.384	0:00:12.801	104
1	AddHL	2	It was found …	20	11:30:52.492	0:00:06.665	233

3.4 Reading Rate Indicator

Based on the relationship between reading rate and learning behavior, Huang et al. (2014) summarized the reading rate and status related to reading behavior, as shown in Table 3. Depending on the reading rate, there are seven reading states: "Slowing", "Memorizing", "Learning", "Raiding", "Skimming", "Scanning", and "Flipping". Each reading state has a corresponding reading behavior. For example, when the reading rate is less than 50 words per minute, it can be considered that the student is in a "slowing" reading state, and the student's learning behavior at this time is mainly one of the following or a concentrated combination. For example, "Excessively slow", "Inefficient reading", "Disfluent", "Labored", "Inexpressive", and "Unenthusiastic rendering".

Table 3. *The Relationship between Reading Rate, Status, and Reading Behavior*

Reading status	Reading rate	Reading behavior
Slowing	1-50	Excessively slow, Inefficient reading, Disfluent, Labored, Inexpressive, Unenthusiastic rendering
Memorizing	50-100	Sustained attention, In-depth reading
Learning	100-200	Oral reading, Concentrated reading annotation
Raiding	200-400	Silent reading
Skimming	400-700	Keyword spotting, One-time reading
Scanning	700-1000	Reading selectively, Browsing and scanning, Non-linear reading
Flipping	>1000	Flip page, Glance and glimpse

4. Reading Rate and Learning Status Tracking

The students' behavior prediction is mainly based on the data collected by the e-book system and the reading rate calculation method developed in this study. Based on the existing e-book system, the Spring MVC framework and Java were used to realize the predictive analysis based on the reading rate, as shown in Figure 2. Teachers can click the reading status submenu in the main menu bar, enter the course name after entering the page, and click search to obtain the seven reading rate status distributions of the students in the course. Each learning status is sorted in descending order of reading rate.

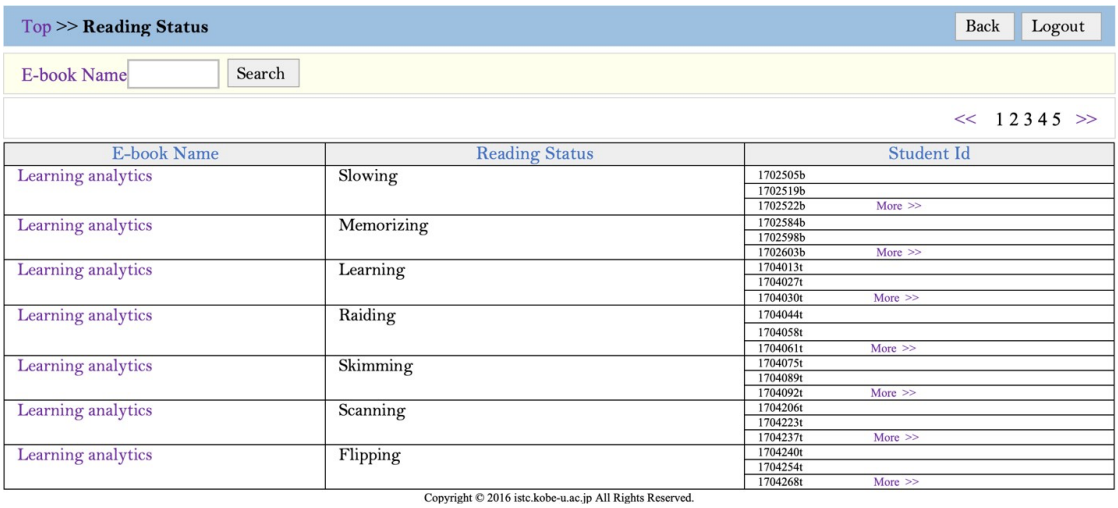


Figure 2. Reading Rate Feedback

5. Discussion and Conclusion

Finding the best way to track students' learning status has been getting a lot of attention. As an important evaluation method, the measurement method and evaluation index of reading rate is constantly developing. With the rapid development of online education, the learning environment and methods have also changed, shifting to mobility and fragmentation. Online reading has been described as an important learning method. In such an environment, there are two problems.

Firstly, whether the traditional methods of calculating correct reading words and reading speed can be applied to online learning platforms. This method has not been well applied to online education. On the one hand, this traditional method reveals many problems, such as being time-consuming and not allowing for real-time tracking. On the other hand, learning behavior and reading habits are also different a way of learning that differs from reading on paper in traditional courses. A second problem arises: how to make effective use of behavioral models of online learning while overcoming the drawbacks of traditional methods, has become an urgent issue to be addressed. After analyzing the strengths and weaknesses of existing methods for calculating reading rates, this study starts with two aspects of reading material manipulation and reading content: firstly, the reading speed method, which

is one of the traditional methods, is used in calculating reading material manipulation, which is the number of words read divided by the number of words read. Calculated by reading time.

Secondly, in terms of calculating reading content, the number of words read as annotations was divided by the number of words read as annotations by calculating students' reading behavior with annotations, i.e. calculating students' reading speed during the underlining and highlighting reading operations. time spent on reading. In this way, a method for calculating reading rates based on online reading was obtained for this study. It is worth noting that by introducing a correlation indicator between reading rate and learning behavior, the reading rate can be translated into seven reading behaviors of students that teachers can easily understand. This can help teachers to identify students' reading status on time, track their learning status, and identify which students are struggling with reading.

Acknowledgments

Part of this research work was supported by the Grants-in-Aid for Scientific Research Nos. 21H00905 from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan.

References

Bernacki, M. L., Byrnes, J. P., & Cromley, J. G. (2012). The effects of achievement goals and self-regulated learning behaviors on reading comprehension in technology-enhanced learning environments. *Contemporary Educational Psychology*, *37*(2), 148-161.

Breznitz, Z., & Berman, L. (2003). The underlying factors of word reading rate. *Educational Psychology Review*, *15*(3), 247-265.

Brysbaert, M. (2019). How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of memory and language*, *109*, 104047.

Carver, R. P. (1992). Reading rate: Theory, research, and practical implications. *Journal of Reading*, *36*(2), 84-95.

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment*, *7*(4), 303-323.

Daly, E. J., Chafouleas, S., & Skinner, C. H. (2005). Designing and evaluating measurably effective interventions for reading problems. *New York: Guilford Publications*.

Deno, S. L., & Mirkin, P. K. (1977). Data-based program modification: A manual. Reston, VA: Council for Exceptional Children

Good, R. H., & Kaminski, R. A. (Eds.). (2002). Dynamic Indicators of Basic Early Literacy Skills (6th Ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Huang, H. C. (2014). Online versus paper-based instruction: Comparing two strategy training modules for improving reading comprehension. *RELC journal*, *45*(2), 165-180.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how?. *The Reading Teacher*, *58*(8), 702-714.

Liang, T. H., & Huang, Y. M. (2014). An investigation of reading rate patterns and retrieval outcomes of elementary school students with e-books. *Journal of Educational Technology & Society*, *17*(1), 218-230.

Ridge, A. D., & Skinner, C. H. (2011). Using the TELLS prereading procedure to enhance comprehension levels and rates in secondary students. *Psychology in the Schools*, *48*(1), 46-58.

Rovai, A. P. (2002). Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks. *The Internet and Higher Education*, *5*(4), 319-332.

Skinner, C. H., Williams, J. L., Morrow, J. A., Hale, A. D., Neddenriep, C. E., & Hawkins, R. O. (2009). The validity of reading comprehension rate: Reading speed, comprehension, and comprehension rates. *Psychology in the Schools*, *46*(10), 1036-1047.

Vonderwell, S. (2003). An examination of asynchronous communication experiences and perspectives of students in an online course: A case study. *The Internet and higher education*, *6*(1), 77-90.

Yin, C., Okubo, F., Shimada, A., Oi, M., Hirokawa, S., & Ogata, H. (2015, November). Identifying and analyzing the learning behaviors of students using e-books. In *Proceedings of the 23rd international conference on computers in education* (pp. 118-120). Hangzhou, China: Asia-Pacific Society for Computers in Education.

Yin, C., Ren, Z., Polyzou, A., & Wang, Y. (2019, July). Learning Behavioral Pattern Analysis Based on Digital Textbook Reading Logs. In International Conference on Human-Computer Interaction (pp. 471-480). Springer, Cham.

Zhao, F., Hwang, G. J., & Yin, C. (2021). A Result Confirmation-based Learning Behavior Analysis Framework for Exploring the Hidden Reasons behind Patterns and Strategies. *Educational Technology & Society*, *24*(1), 138-151.

Deep, A. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

An Evaluation of the Online Course on Climate Change and Emotion

Anurag DEEP*, Aditi PATHAK & Nandini CHATTERJEE
UNESCO MGIEP, 35, Firozshah Road, New Delhi - 110001 India
*a.deep@unesco.org

Abstract: Education possesses the potential to enable people to understand the causes and consequences of climate change, make informed decisions and take appropriate actions to address it. Behavioral research shows that decision-making is a cognitive-emotional process. Therefore, any intervention to enable appropriate actions for climate change must include the knowledge and skills necessary to enable sustainable living. Based on such an integrated model, an online interactive course was designed for middle school learners in the age group of 12-15 years. Here, we present the design of the course and a pilot study conducted (N=52) to investigate the efficacy of the course. Efficacy was assessed through a self-reported pre-post Sustainable Development Self-Efficacy (SDSE) questionnaire. A paired sample t-test between the post-and pre-test SDSE questionnaire scores revealed a significantly increased score ($p < 0.05$). Specifically, increased scores were seen on the subscales of "Influence of environment" ($p < 0.05$) and "Economic welfare" ($p < 0.05$). This study provides preliminary evidence of an online course to build knowledge of climate change and mindsets to affect pro-environmental action to mitigate its impact.

Keywords: Climate Change, Social and Emotional Learning (SEL), Sustainable Development Self-Efficacy (SDSE) questionnaire

1. Introduction

Climate change is now recognized as a grave challenge to the present and future of humanity. A serious consequence is rise in the earth's temperature by 1.5 degrees. The earth's temperature has varied dramatically during its 4.5-billion-year history, but the current rapid warming is not natural. These changes will grow and become more severe and inescapable with time. To mitigate the challenges ahead, humanity must learn to live sustainably (Osbaldiston & Schott, 2012). Accumulating evidence points to the fact that the effects of climate change can only be addressed and mitigated by altering human behavior (Mogensen & Schnack, 2010). In fact, emerging research shows that learning about climate change "*is not only a cognitive endeavor but also involves cultivating skills to manage emotions evoked by the seriousness and complexity of these problems."* (Ojala, 2012). The Action for Climate Empowerment (ACE) recognizes that including climate change in curricula is crucial in solving the complex challenges presented by climate change. There is an urgent need to not only develop courses on climate change for all ages but also ensure that any climate change-related curriculum should address cognitive and emotional learning domains.

An internet search revealed that while many courses that provide information about climate change exist, there were hardly any courses that combined knowledge of climate with emotional skills. To address this gap, an online course on climate change that combined knowledge with social and emotional learning (SEL) was developed. SEL can be defined as the process of developing the competencies, abilities, and attitudes necessary to recognize and control emotions, develop caring and concern for others, form positive relationships, make responsible decisions, and deal with challenging situations (Payton et al., 2000; Greenberg et al., 2003; Weissberg et al., 2015). Embedding SEL competencies structured around the **EMC^2** framework (Singh & Duraiappah, 2019): **E**mpathy, the ability to understand how someone is feeling. When students become empathetic towards nature and the environment, they no longer see global issues such as climate change, refugee crisis, discrimination and racism as externalities but are ready to act and engage in pro-social behavior. **M**indfulness is the ability to be present in the moment, acknowledging our thoughts without judgment. Mindfulness tools help

learners understand the knowledge and information about climate change with strategies to recognize the emotional reaction to this change and ways to cope and manage one's emotions through mindfulness exercises. **C**ompassion happens when our empathy makes us want to help. A compassionate learner is more likely to take climate action. **C**ritical Inquiry is the ability to think clearly, rationally, identify, analyze, and find ways to solve problems. This skill helps learners critically analyze their 'ideas'.

2. About the course

The course entitled **Climate change: Understand, Reflect, Empathize, and Act** was designed for middle school learners and focused on two aspects (a) information on key climate change-related concepts such as carbon footprint, the science of climate change, causes, impacts, adaptation, and mitigation (2) the role of self-awareness, emotional regulation, critical inquiry, perspective-taking, and compassion to address the impact of climate change. The course is hosted on the digital learning platform Framerspace (www.framerspace.com/course/climate-change), available free of cost to all learners above the age of 13 years. The duration of the course is 20 hours and is offered in English. The course could be integrated into the regular teaching-learning practice and embed concepts from math, science, and social science. The specific learning outcomes of the course are

1) Understanding varying and often conflicting points of view around the issue of climate change.
2) Noticing and becoming aware of one's emotional response to global issues such as climate change.
3) Critically evaluating and inquiring about various aspects related to climate change.
4) Practicing mindfulness activities and learning to be mindful of one's emotional and cognitive state throughout the course and beyond.
5) Understanding the knowledge and information about climate change with strategies to recognize the emotional reaction to climate change and ways to cope and manage one's emotions through mindfulness exercises.

The online course uses the Libre pedagogical framework (Rautela, & Singh, 2019) that encourages the use of multiple pedagogies to cultivate learning, namely 1) Reflection; 2) Discussion; 3) Inquiry; 4) Storytelling; 5) Simulations and Games. The use of reflective practice encourages learners to reflect on issues related to climate change or introspect and self-evaluate their values, beliefs and attitude related to the topic. Discussion activities enable learners to learn different perspectives, express their opinions, and engage with multiple points of view. In the course, tools of inquiry are used to help learners investigate and explore the difference between fake and real news and information related to climate change. Storytelling explains the concept of trade-offs and synergy in the context of low carbon emissions alternatives, which not only helps in simplifying complex topics but also makes the module engaging. Simulations and games are used to investigate the economic impact of pro-environmental decisions by policymakers.

3. Evaluation of the Course

A single group pre-post research design was implemented to evaluate the impact of the course on learning, usability, and its efficacy in building competencies to take appropriate decisions for climate change. Figure 1 below presents the study design.

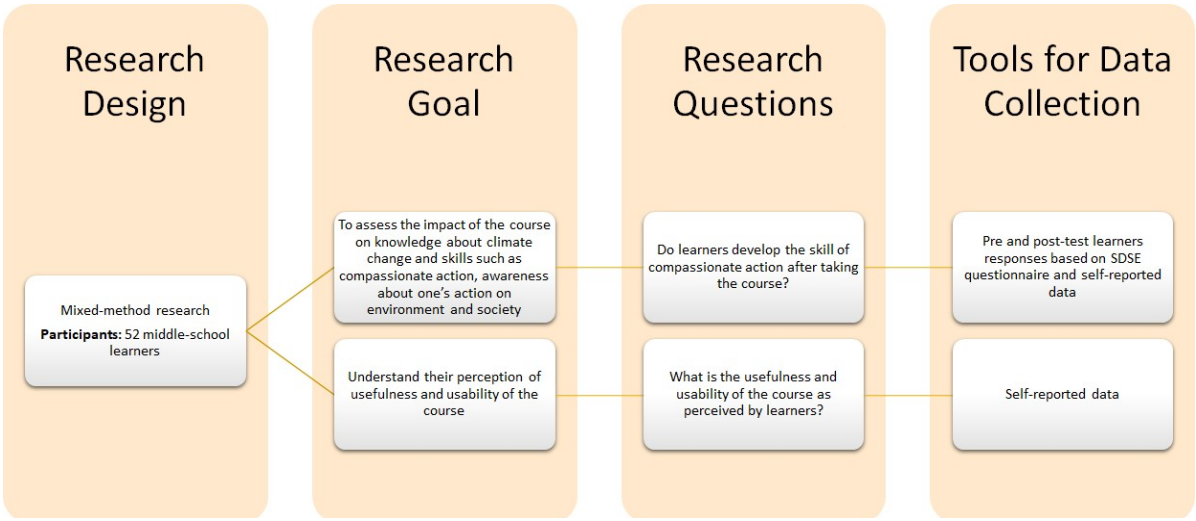


Figure 1. Study design for evaluation of the course

3.1 Participants, Study Procedure, Data sources and Analysis Technique

A total of 52 participants (N=52) were recruited from three urban schools in India from grades 7 and 8. Before the start of the course, a workshop session was conducted for the teachers to introduce them to the course, steps involved in the research study followed by the workshop session for the learners. For the first research goal the Sustainable Development Self-Efficacy (SDSE) questionnaire (Hanss, & Böhm, 2010) was used. The 20 SDSE questions were divided into 3 groups: **Influence on the environment** (reflects beliefs about the influence of one's actions and of humans in general on the environment), **Economic welfare** (represent beliefs about one's capability to contribute to a socially fair distribution of resources and economic welfare) and **Influence on others** (represents beliefs about whether one's actions to foster sustainable development motivate others to do the same). Example items included: Influence on the environment ("I believe my actions have an influence on global warming and climate change."), Economic welfare ("With my everyday consumption and buying behavior I can contribute to a socially fair distribution of resources around the world."), Influence on others ("My actions to reduce the effects of global warming and climate change in my community will encourage others to do the same."). Learners responded to the SDSE questionnaire before and after interacting with the course by choosing the Likert scale responses, which had options from "strongly disagree" to "strongly agree".

For the second research goal, self-reported questionnaire was used to evaluate usability and usefulness of the course. These questions evaluated learners' perceptions of the contribution of the course to learning - "Level of skill/knowledge at start/end of course". There were other questions regarding feedback on the course organization, learning objectives, workload, and participation. Besides this, open-ended questions such as, "What aspects of this course were most useful or valuable for you?" and "What can be improved/added in the course to be more beneficial to you?" were included.

3.2 Results

3.2.1 Quantitative Analysis

A paired t-test between the post-and pre-test SDSE questionnaire scores was conducted to assess the learning gain after interacting with the course material. Overall, there was a statistically significant difference between the post-and pre-tests ($p < 0.05$) (Figure 2). Statistically significant differences were found in questions that evaluated learners' understanding of the influence on the environment ($p < 0.05$) and in the construct of economic welfare ($p < 0.05$), which focused on a socially fair distribution of resources.

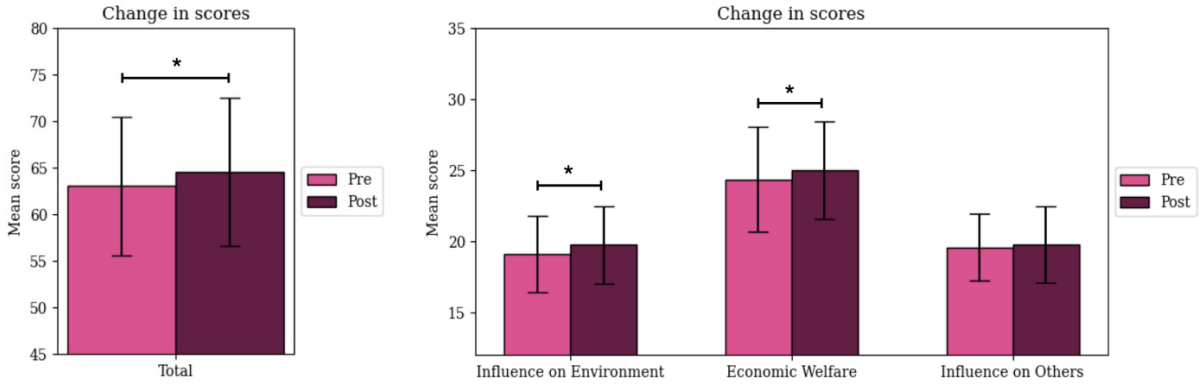


Figure 2. SDSE Questionnaire Mean Score Variation Across Pre-and post-test: Total, Influence on Environment, Economic Welfare and Influence on Others

Data analysis of the self-reported questionnaire on usability and usefulness of the course revealed a statistically significant difference ($p<0.01$) in the perception of the level of knowledge before and after the course.

3.2.2 Thematic Analysis

Data analysis of the responses to feedback on the course organization, learning objectives, workload and course participation revealed that 91% of the learners agreed that the course content was well organized, well planned and had clear learning objectives. Moreover, 88% of the learners agreed that the course workload was appropriate and allowed them to participate fully.

Table 1: *Themes and Respective Sample Excerpts*

Theme	**Instance of response from participant's artefacts**
What aspects of this course were most useful or valuable for you?	
Pro-environmental Behaviour	"Actions are taken by the people in the world to stop climate change, students can also have their contribution for stopping climate change." "The aspects were that it encouraged me a lot to do something about climate change."
Course Content	"The videos and articles were very useful to me because they helped me gain a full understanding of the topics." "Aspects of videos and chart or poster making activities more useful."
What can be improved/added in the course to be more beneficial to you?	
Live Sessions	"Live interaction with students for solving climate change problem"
Adding resources and activities	"The course was very effective in teaching, but if there were more articles in the course that would benefit me more."
Availability in another language	"Make it in different languages so that people around the world can understand it more clearly"

4. Discussion

The purpose of this study was to evaluate the efficacy of a course on climate change. Analysis of the result shows that overall, there was a significant difference between the pre-and post-tests which had SDSE questions. This could be attributed to the reflection activities done by learners across modules of the course. Since participants had to reflect on varying and often conflicting points of view around the issue of climate change, we conjecture that the course helped learners take compassionate action and understand that one's actions and of humans in general influence the environment. Additionally, a statistically significant difference in the questions on economic welfare, which evaluates learners' beliefs about one's capability to contribute to a socially fair distribution of resources, was also found. This finding may be attributed to the use of inquiry in the course that helped learners critically evaluate and inquire about various aspects of climate change.

Thematic analysis of open-ended responses indicated that this course helped learners to identify their influence on the environment and motivated them to act for pro-environmental behaviors. An excerpt from the participant's response "Actions are taken by people around the world to stop climate change, students can also have their contribution for stopping climate change" led to the triangulation of the findings from the SDSE quantitative analysis, where we found overall significant difference along with the significant difference in the theme of influence on the environment and influence on others. Furthermore, learners recognize that interaction with the course helped them to understand the knowledge and information about climate change. Also, learners perceive that the course is highly usable and useful for the learners. It is recommended that longitudinal research based on a quasi-experimental research design must be conducted to assess and evaluate the long-term impacts of such an intervention.

References

Greenberg, M. T., Weissberg, R. P., O'Brien, M. U., Zins, J. E., Fredericks, L., Resnik, H., & Elias, M. J. (2003). Enhancing school-based prevention and youth development through coordinated social, emotional, and academic learning. *American psychologist*, *58*(6-7), 466.

Hanss, D., & Böhm, G. (2010). Can I make a difference? The role of general and domain-specific self-efficacy in sustainable consumption decisions. *Explaining sustainable consumption. Findings from cross-sectional and intervention approaches*.

Mogensen, F., & Schnack, K. (2010). The action competence approach and the 'new' discourses of education for sustainable development, competence and quality criteria. *Environmental education research*, *16*(1), 59-74.

Ojala, M. (2012). Regulating Worry, Promoting Hope: How Do Children, Adolescents, and Young Adults Cope with Climate Change?. *International Journal of Environmental and Science Education*, *7*(4), 537-561.

Osbaldiston, R., & Schott, J. P. (2012). Environmental sustainability and behavioral science: Meta-analysis of pro environmental behavior experiments. *Environment and behavior*, *44*(2), 257-299.

Payton, J. W., Wardlaw, D. M., Graczyk, P. A., Bloodworth, M. R., Tompsett, C. J., & Weissberg, R. P. (2000). Social and emotional learning: A framework for promoting mental health and reducing risk behavior in children and youth. *Journal of school health*, *70*(5), 179-185.

Rautela, R., & Singh, N. C. (2019). Libre—Nourish the Brain So the Future Can Flourish. *Childhood Education*, *95*(5), 34-43.

Singh and Duraiappah. (2019). EMC2 – a social and emotional learning framework to build human flourishing. Position paper, UNESCO MGIEP.

Weissberg, R. P., Durlak, J. A., Domitrovich, C. E., & Gullotta, T. P. (2015). Social and emotional learning: Past, present, and future.

Omair, H. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Augmented Reality Applications in the Classroom: Teachers' Experience

Hajar AL OMAIR[a*], Antonette MENDOZA[b], Marie BODÈN[c], Nilufar BAGHAEI[d]
[a] *College of Computer Sciences and Information Technology, King Faisal University, Saudi Arabia*
[a,c,d] *School of IT and Electric Engineering*, *The University of Queensland, Australia*
[b] *Faculty of Engineering and IT, The University of Melbourne, Australia*
*h.alomair@uq.edu.au

Abstract: Augmented Reality is overlaying three-dimensional images or videos on a surrounding physical environment. Many studies have proven that Augmented Reality (AR) can enhance users' learning experience and promote their engagement and excitement. There are also studies showing that teachers have a positive attitude towards using AR as an educational tool in the classroom. However, there is a lack of AR applications implemented in schools and there is limited research examining K-12 teachers' use of AR applications. Hence, this study followed a qualitative approach by conducting in-depth interviews to understand the experiences of six K-12 teachers in Australian schools that had used and implemented AR in their classrooms. This research aimed at uncovering the underlying advantages and challenges teachers and their students have faced when working with AR in the classroom, and the teachers' intentions for future use. Our findings show that the teachers believed AR is a valuable learning experience, and they would use it in the classroom if it served their needs. Teachers also believed AR provided new ways for children to express learning and enhanced school and community engagement. They listed curriculum, time constraints and support, followed by students' ages, space, equipment, and technical issues as challenges for AR use. The most emphasized challenges were curriculum fit and mapping AR applications with curriculum, as well as lack of educational AR applications and applications that are co-designed with educators and students. We believe our findings will help designers, researchers, and schools to better understand teachers' requirements which can be used to inform the design of more efficient and effective AR applications that will serve the needs of its users and minimize the disruption on the classroom environment.

Keywords: Augmented Reality, AR, Implementation, K-12 Education, Advantages, Challenges, Teachers, Teacher's Experiences, Future Attitudes, Student Experiences

1. Introduction

Advancements in technology are continuously bringing opportunities for new ways of supporting learning. Evolutions in technology have allowed virtual illustrations to be in 3D which can promote excitement and engagement (Irshad and Rambli 2016). One such technology is Augmented Reality (AR) where computer-generated 3D images, videos and sometimes audio are presented on a surrounding physical environment. This emerging technology is still trying to find its place in education despite significant evidence indicating its potential as a valuable teaching and learning tool. According to the review held by Chen et al. (2017), AR's main advantages in education are: better learning performance, increased learning motivation, deeper engagement of students, perceived enjoyment, and positive attitudes. Yet, little is known about why current AR applications have not been adopted and retained successfully in the 'Kindergarten to 12th grade' educational environment. Although the concept of AR initiated in the 1990s, there is still a lack of AR applications implemented in the educational sector (Radu 2014). According to Lee (2012), this could be due to their perceived incompatibility with the traditional learning methods, the costs incurred for developing and maintaining AR, and general resistance to new technology. Akçayır and Akçayır (2017) pointed out two major challenges facing educational AR applications that involve usability and technical issues.

Existing literature have focused on the short-term effect of AR on students' learning as well as the emotional states of students when interacting with AR in schools (Ibáñez and Delgado-Kloos 2018), while few focus on the longer-term effects of using the technology, after the temporary "wow" moment

of any novel technology. Also, several studies have explored teachers' intentions for adopting AR in the classroom and they have shown that teachers had positive acceptance rates for using AR in the classroom (Alkhattabi 2017; Banerjee and Walunj 2019; Ozdamli and Karagozlu 2018). However, more research is needed on how teachers have implemented AR and their experiences in their classrooms. Hence, there is an absence of studies that investigate the experiences of the users of educational AR applications in the longer-term of using it, especially in the context of K-12 education.

This study aimed to interview teachers currently using AR in their classroom and understand what is working, what is not and what they think about using AR in the future. While not neglecting the importance of students input for understanding these concerns in the future, teachers' opinions are required as they play a key role in the adoption of these technologies. In addition, teachers from diverse backgrounds and specialties may have different views on their experiences with AR. Understanding the existing experiences of K-12 school teachers who have initiated the use of AR in their classrooms, will aid in understanding the needs of the users and what challenges they face during implementation.

2. Literature Review

It is common to hear the terms Augmented Reality (AR) and Virtual Reality (VR) together, which is not wrong as they are both under the genetic term Mixed Reality, however, they have different degrees of reality. VR is when someone is immersed in an entirely virtual environment where there is no connection to the real world. Whereas AR is like a middle point that connects what is completely real (e.g. a physical table that you can touch) with completely virtual objects (e.g. a 3-D image of a synthetic lamp) and aligns them with each other interactively (Azuma 1997).

In recent years, there has been an upsurge of published papers regarding AR in education. A dramatic increase started in 2011 and escalated in 2013 to 2017 (Akçayır and Akçayır 2017). It is believed that the spread of mobile devices and AR development tools, even in developing countries, increased the use and study of AR (Akçayır and Akçayır 2017; Garzón, Pavón, and Baldiris 2019).

AR enriches individuals' understanding of the real world through virtual aspects that cannot be easily detected by their senses (Azuma 1997). While using AR in education, it can help students gain deeper understandings of the human anatomy (Fuchsova and Korenova 2019), learn the English vocabulary in an attractive and interactive manner (Lee et al. 2017) or even introduce animals in early childhood education (Nainggolan et al. 2018). AR can serve students with different learning abilities to *understand spatial data* and *improve reading comprehension* (Billinghurst and Duenser 2012). Also, AR aids the future direction of *individual learning* (Pence 2010).

Many reviews have collected the advantages of AR in education, and similarly they found that the top two are "Learning Gains" and "Motivation" (Akçayır and Akçayır 2017; Altinpulluk 2019; Bacca et al. 2014; Chen et al. 2017; Garzón et al. 2019; Radu 2014). *Learning gains* is where students show higher learning performance and achievement. This was proven through studies that adopted a pre-test - post-test approach on every experimental (with AR) and control (traditional teaching) groups of students. While *motivation* is where students feel more motivated to learn compared to traditional learning methods. Garzón et al. (2019), believed that this motivation is driven by *sensory engagement*, where students use more than one of their senses to learn. Since AR is a representation of 3D elements, it helps in *visualizing abstract concepts* that are difficult for students to grasp with traditional teaching methods which eventually may help in greater *memory retention* (Garzón et al. 2019; Radu 2014). Another advantage of AR in learning is that it may promote students' *self-learning* (Papanastasiou et al. 2019; Pence 2010), or even providing means for *collaboration* between students in cases like in-formal learning where students team up to solve an outdoor mission (Chang, Chung, and Huang 2016; Chiang, Yang, and Hwang 2014).

Despite the great evidence that prove AR's potential as an effective learning tool, there are some challenges that had been reported. Usability and technical drawbacks are the most reported issues, as AR is still an emerging technology. While AR can be in many types, the technical issues may differ. A *Marker-based AR* tracks a marker which may be interrupted if the marker is out of camera's sight and cause students' confusion (Dalim et al. 2016). Also, if the marker is disturbed by light and pointing angles, this may cause identification failure (Chang and Hwang 2018). Using a *Marker-less AR* may solve those problems; however, this type is still evolving as the tracking work is much harder in this type (Bacca et al. 2014). For *Location-based* AR, GPS signals would sometimes be inaccurate in locating a specific learning target (Chiang et al. 2014; Dunleavy and Dede 2014).

In regards of interaction design, students' *cognitive overload* was a common reported issue, due to difficult tasks, app-related *technical issues,* and *teacher's inadequate training* (Papanastasiou et al. 2019). Studies have already proposed to solve this through simplistic and enhanced design (Dunleavy and Dede 2014), however, more usability studies are needed that consider the state-of-art characteristics with an educational environment (Akçayır and Akçayır 2017). This is important to understand user preferences and establish suitable design principles and evaluation criteria for AR use in education to deliver AR to its full potential (Billinghurst and Duenser 2012). Additionally, curriculum was considered as a challenge that may limit teachers' adoption and use of AR applications. Tzima et al. (2019) surveyed teachers and curriculum was found to be the major restricting factor. Other teachers pointed out that they lack the knowledge and training of how to map curriculum contents to AR applications (Oliveira da Silva et al. 2019).

School's Support was mentioned as well in the literature, and this can be considered as a general challenge with the adoption of new technology. In order for teachers to adopt new technology in teaching, school administrators need to effectively offer implementation support (Sugar, Crawley, and Fine 2004; Tondeur et al. 2017). In schools where devices are not available, hard to get due to financial reasons and school policies restrict students to bring their own devices are challenges for teachers to implement AR (Alalwan et al. 2020). Moreover, applying AR in a lesson may take more *time* than a usual lesson. This may be due to the need of training for teachers and students, the complexity of the application or because of the novelty of AR (Cheng and Tsai 2013). Oliveira da Silva et al. (2019) surveyed teachers and found that time was a factor that prevented teachers from implementing AR more often as more planning and class time is needed. Lack of time was also an AR challenge mentioned by teachers (Alalwan et al., 2020).

3. Methodology

The aim of this research was to examine K-12 teacher's experience with AR in the classroom, we did a qualitative study, and conducted semi-structured interviews. The method was chosen as it offers deeper insights on the entire experience rather than a specific part or aspect of an experience as with quantitative methods (Moustakas 1994).

Due to the closure of schools during the pandemic of COVID-19, alternative ways of contacting teachers were through social media platforms (such as Twitter, Facebook, Instagram), by attending an online STEM conference, dedicated online AR websites, through school homepages and participants in this study nominating colleagues that are familiar with using AR. There were no incentives offered to the participants. Due to the difficulty in recruiting teachers that are familiar with the use of AR in the classroom and the low response rate in the circumstances of COVID-19, this study has a limited set of 6 teachers where one is an educational technologies advisor as well. While Creswell et al. (2007) inferred that recruiting 5-25 participants is recommended for phenomenal studies to explore individual experiences with a certain phenomenon, this study is in the recommended range. All participating teachers in this study teach in Australian schools. Where two teach in primary school, two teach secondary school students and two teach in high school. Further details about their backgrounds are in the results section.

Data was collected after an ethics approval was obtained. Prior to the interview, the participants were initially asked to read the plain language statement that had information about the research and purpose of the study. Then they were asked to fill in an online pre-questionnaire through Qualtrics that requested their consent to gather data and audio-record the interview. The responses were stored by the Qualtrics platform that was secured through an official username and password. The interviews were held through an online meeting room via *Zoom*. The interviews lasted for about 50 minutes on average. After the interview, the audio-recordings were transferred from the recorder to the researcher's laptop and then transcribed and stored safely.

All six participants completed the pre-questionnaire and only five had the time to proceed with the interview. The pre-questionnaire was initiated by collecting demographics about the teachers, followed by teacher's familiarity with the use of technology in the classroom and particularly AR. For example, their age, teaching period, what subjects they teach, which grade levels…etc. The final section of the pre-questionnaire was a Likert scale about teacher's experience with AR in the classroom. This scale was inspired from a study held by Santos et al. (2014) that was aimed for hand-held AR

applications. Noting that on initial planning of testing, the aim was to focus on Handheld AR, however due to the difficulty of finding teachers that only use this type, the scope was expanded. Results of the usability scale were not analysed quantitatively because teachers' AR application methods were different.

Interview questions were based on the following four themes: *1)* Motivations for AR Use; *2)* Planning Strategies; *3)* Positive and Negative Usability Aspects; 4*)* Future Use Attitudes. After data was collected, the data was analysed thematically following the six stages proposed by Braun and Clarke (2006). Thematic analysis is suitable for the research in-place because it can be a "realist method, which reports experiences, meanings and the reality of participants" (Braun and Clarke 2006), uncovering teacher's real experiences and attitudes with the implementation of AR in the classroom.

4. Findings

The findings are organized based on the themes extracted from the data collected. Beginning with teachers' biographical backgrounds, followed by their current use of AR and finally their attitudes towards future use. The participants were teachers in Australian schools, three males and three females with diverse age ranges (Table 1). As for their educational background, three had a bachelor's degree, one had a master's degree and two had a master's degree with an in-progress PhD degree.

Table 1. *Age Distribution of Teachers*

Age	20-29	30-39	40-49	50-59
Count	1	1	3	1

According to teaching periods, one had been a teacher for less than 5 years, three teachers had 7 years of experience on average, whereas two had been teachers for about 23 years. In terms of subjects, majority teach Digital Technologies or STEAM (Science, Technology, Engineering, Arts and Math) (5 teachers), Science and Humanities & Social Sciences (3 teachers each), Arts (2 teachers), Mathematics, Future Literacies, Religion & Values and ESL (English as a second language) (1 teacher each). As most of the teachers teach digital technologies or STEAM, they all frequently use technology in their lessons and say they have school's support to use it as well. The responses from the six teachers have been annotated here in the findings as T1-T6. All teachers had used AR in their classes; however, their use was limited. Five teachers applied AR occasionally and some very rarely. While one teacher used AR frequently as in almost every week. They had used different AR applications, 5 used *"Merge",* one had used *"Aurasma"* (Hp Reveal), *"Artivive"*, *"SolarSystems"*, *"Metaverse"*, *"Adobe Aero"* and *"CoSpaces*".

It is important to note that learning with AR can be by creating AR content and viewing AR content. A couple of teachers found it relevant to share their experience with VR as well. There were different methods of how the participated teachers implemented AR or VR with their students:

- Individually in class, each student created their own AR content using a desktop computer and then their work was viewed out of class in a school art exhibition by iPads (T4).
- Individually in class, each student had their own Google Cardboard and inserted their own mobile phones to view VR content (T3).
- Individually at home as assignments for students, each student created a 3D model of their "Minecraft" world and then uploaded their model in "Adobe Aero" to be able to overly their virtual Minecraft world on their real world (T2).
- Individually at home as a virtual art exhibition, students could walk through their artwork (This method was adopted when schools were closed during the COVID-19 pandemic) (T4).
- Collaboratively in class, where two students took turns to use a headset to view AR or VR Google Expeditions content (T1).
- Collaboratively in class, using an iPad to view AR content from a Merge cube (T1, T2).

4.1 Motivational Factors

There were several reasons why participants considered adopting AR in their lessons. This study divided these motivational factors into three main categories; (1) Enhancing Students' Learning, (2) Technology Capabilities, (3) Extending Teacher's Education (see *Table 2*).

Table 2. *Teachers' Motivational Factors for Using AR*

Enhancing Students' Learning	*Technology Capabilities*	*Extending Teacher's Education*
Learning to code	Bringing things to life	Conducting a PhD study on AR
Exposure to new technology	Immersive	Learning new ways of teaching
Creating AR content	Novel	
New ICT skills		
Spatial ability		

Four teachers were motivated to use AR because it enhances students' learning. Two teachers chose it to teach students about coding and gain new skills, the teacher explains: "*...we need to teach students how coding works. There's a little bit of coding in AR, not very much, it is more ICT skills."* (T1). They also expressed that they wanted to teach students how to create 3D content. T1 also believed that primary students should get exposed to AR as a technology, T1 states: *"...the creation really should be taught in years 8,9,10...but I think if children are already going to be using it, they should learn how it works, so they can have a better understanding of it".*

The second motivational factor was the special technological capabilities that AR has to offer. Two teachers had used AR because it could transform static work into an interactive live piece of work. Teachers also saw that it helped students understand spatial ability. T2 had used AR to transform students Minecraft Worlds into a 3D image that is overlaid on their school desks, the teacher illustrated how this was a motive: *"if I didn't it was just going to stay there, when I transformed it into 3D, I'm able to print that picture in the school newsletter, and students could give each student to do a selfie with their own world that they built. I can send it to students, I can be more proud of the work that they've done. Furthermore, we do it for them to see the world in three dimensions in a bigger way. So, it's quite interactive for them."*

In addition, findings showed that AR may increase students' attention and maintain focus in a subject that is usually not that engaging, due to its novelty and immersive nature. Two teachers had commented on these factors: *"So, the immersive nature was the main reason that I wanted to use it because I thought it would attract students and it would keep them focused on the task. They would actually be having fun I thought that while they're having fun, it may be easier for them to learn."* (T3, ESL class), *"So, I was kind of motivated to just make my subject a bit more engaging, because a lot of the time, because it's not one of the mainstream subjects, kids tend to take it and think they can slack off, they're on computers, you know, I'll do whatever I want kind of thing. And I want to just get them a bit more engaged in the technology aspect of the subject."* (T4, Art Class)

The third factor is extending teacher's education where teachers state that they learnt more about AR and its capabilities in the classroom. T2 and T3 are conducting scientific research about AR to complete a PhD degree. Where others seek new ways to teach in the classroom and pass the experience to other colleagues to try. T4 was recommended by another teacher to use AR in her subject: *"'Teacher X' actually suggested that I should do something with augmented reality and at first, I was quite resistant because it seemed like a big learning curve, and when I looked into the applications of it in an art context, I realized it could really, really pay off for the students just a really cool thing"* (T4)

4.2 Teachers' Positive Experiences

The participants expressed positive experiences that they and their students faced when AR was implemented in their lessons. In this study, the positive experiences were divided into two main themes: (1) Engagement, and (2) Interactivity.

All teachers strongly agreed that using AR in the classroom is engaging. T1 says that students sometimes needed to be *"tricked"* to learn something new. Another teacher found students appreciating learning in a new method such as creating AR content, "*It's beyond, you know, getting them just to*

watch a video, they're really engaged in the process" (T4). Two teachers also mentioned that engagement is not just for students of the subject, but engagement of the school as well with their projects. T2 could take selfies of the students with their Minecraft worlds and posted it in the school newsletter. The teacher also introduced "Aurasma" in a STEM fest for the whole school community to be exposed to AR. While the other teacher's great experience was in a school art exhibition where students of all ages and staff engaged with the artworks: *"…and it was just, you know, really excellent at the end, watching all of the kids that weren't even in that class, just engaging with the artworks and just really engaging with the art exhibition, which was, it's actually quite rare."* (T4). This teacher also had another engagement experience during COVID-19, where a student's whole family could visit a virtual AR art gallery of the students' artworks. The teachers also agreed that using AR in the classroom is fun and exciting. One teacher said students expressed AR as "magic," *"I found kids saying this a lot with augmented reality, like it's like magic, you know, because they can see their world through the lens of the camera as well and their world is being changed"* (T4). It wasn't just exciting for the students but for the teachers as well, "*The positive part was hearing the students saying, wow, this is amazing! ... That's what got me excited as a teacher to say that they're actually focused on the tasks, and they are enjoying their classes"* (T3).

In addition, learning by AR can be more interactive than a traditional taught lesson. A teacher commented that if complicated concepts can be explained through AR, *"it is a lot more fun way than listening to a teacher."* (T1). It is not just fun but can be a more effective way of learning certain concepts as well, "*Love seeing their work in AR out in the real world - helps put learning into context for them"* (T5). Five teachers strongly agreed that when using AR, it is possible to describe concepts that are difficult to explain without it. Teachers explained how students can manipulate AR objects and feel more in control of their learning, *"...with Merge Cube, we were looking at the flower, how a bee pollinates a flower, and we rotated it around in our hands, and can look at different steps, we're looking at the plant cell cytoplasm, and from mitochondrion, all those things, it's quite interactive. So, when you just have a potential draw diagram on the board, or even if you show a power point slide show, it's quite different for them. It involves for interactivity and self-exploration; the student can go over with in their own time later on."* (T2). Another teacher adds that AR could give an opportunity for students to express their learning when sometimes they were not good at expressing their learning in a traditional classroom, "*So, that's what drives me to students sometimes they're not very good at showing their knowledge in traditional ways and an opportunity to show knowledge in non-traditional ways,"* (T1).

4.3 Teachers' Challenging Experiences

Even though five teachers in the pre-questionnaire stated that the AR applications were easy to use in class, they expressed their experiences in the interviews more differently. The following derived themes were the challenging experiences that teachers had faced when they implemented AR in their lessons.

Curriculum: The most frequent challenges mentioned by the teachers were related to curriculum. Some refer to it as the most difficult aspect to consider when thinking about using AR in class. One participating teacher said that it is hard to link between the technology and the learning area, "*The hardest thing is to relate the use of the technology into another, we call them KLA's or Key Learning Areas"* (T1). Teachers explained that AR content must fit very well in the curriculum, or it will be difficult to justify its use, "…*certain activities won't be at a simple level, it may not directly map with the curriculum that I'm teaching to the students at that stage, so I'll have to map that out properly as well."* (T2). While teachers believed that content fit was the most important, they acknowledged that using AR would make students engage better with the content even if it didn't map perfectly to the curriculum. It was easy for the STEAM teacher to use AR as the focus was the skill of using the technology. Other teachers found it challenging to find AR applications with content that is related to the curriculum or even for general educational purposes*, "…I was not able to find a lot of appropriate games for educational purposes or .. related to the curriculum .. related to general lack of values or sustainability,..."* (T3).

Time Constraint: Four teachers agreed that using AR in the classroom is time consuming. Before using AR in the classroom, the teachers needed to understand how AR worked and sometimes they just didn't have the time, two teachers expressed this: "*it's quite a learning curve to begin with to implement augmented reality and a lot of times it you know, teachers will have to do that in their own time.*" (T4). Teachers found that students must learn the skill of using the technology before being able

to comprehend any content from it, which requires more class time, *"... if you're trying to fit augmented reality into, say a HASS subject and then with all the skills, you have to teach the kids to be able to function the augmented reality, it takes away a lot of time from the actual content that you're required to learn."* (T4). In addition, as with using something new like AR in class, more time was needed to solve unexpected problems and issues with technology.

Support, Space and Equipment: The findings showed that support was an important factor that affected teachers' use of AR in the classroom. Different forms of support were school support, teachers' support, and companies' support. All teachers agreed that they are more likely to use AR if they have school support. A teacher commented that schools need to be willing to try new technologies, "*…you need to have a school that's really forward thinking or really wants to approach new technologies…. if the schools were able to take into account and support teachers, by giving them like, extra time or extra support, um, that would really help as well"* (T4). Technical support from school was also named as important for teachers to use AR. Teachers had experienced issues such as logins, file sizes and licenses. One teacher said, "*If they're going to have any negatives, it will probably be on the technical side where getting the app installed that would be one thing, would the app be free or not, available to sign in with their school credentials or the school email like some that might have single sign on with Google or single sign on with Microsoft."* (T2). Other than school's support, support from colleagues was another aspect that had been articulated by the teachers, *"'Teacher X' was willing to help, so I think having the support of like, the other teachers really kind of helped it work, I think without that it never would have got off the ground."* (T4). Access to space and equipment were also mentioned as an important factor for teachers to use AR. Some teachers said that a regular classroom with desks and chairs worked fine with AR, however, younger students needed more space. One teacher stated: "*They were able to actually control themselves while they would play because they had some space in between desks and that Blackboard or the whiteboards, they were good. But with the younger ones, the year sevens and eights, they were more mobile while they were actually playing the game."* (T3).
The findings of teachers' current experiences with AR in the classroom are summarized in *Figure 1.*

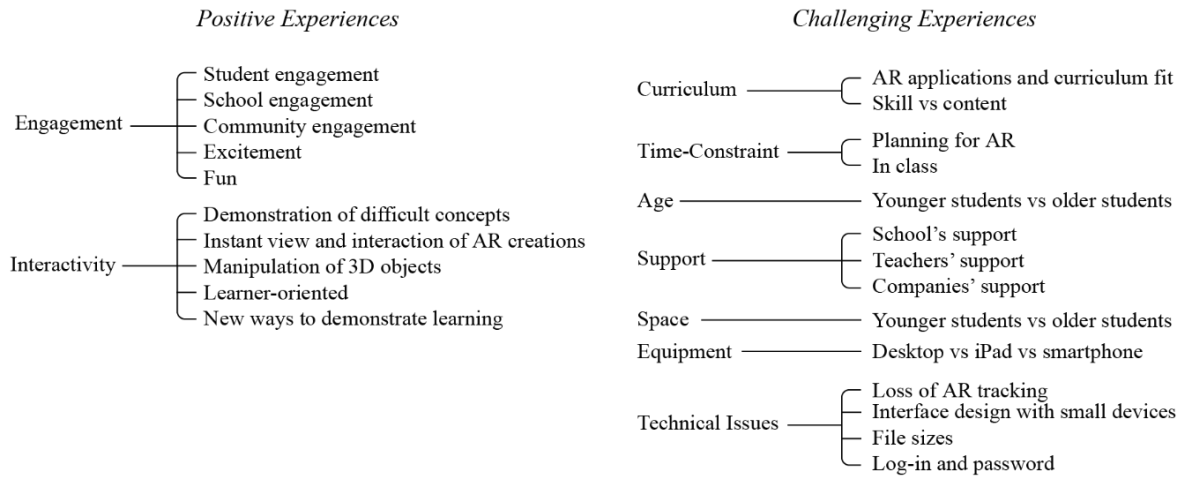


Figure 1. Summary of Teachers' Experiences with AR

4.4 Teacher's Future Use of AR

Teachers were asked about their attitudes toward future use of AR in the classroom. Teachers' replies were positive, and they would recommend other colleagues to use AR in their lessons, "*Yes, I would and I'm actually going to demonstrate merge cube to my other colleagues so they could use it in their science lessons, because that's when it's going to really make an impact"* (T2). Though, one teacher added that teachers believe it's difficult. A teacher said that she would recommend teachers, however caution them about the time and effort that needs to be put in, *"I would because of how engaging it is for the students but I would also caution them as to how much time it does take to set up and we had a lot of great payoff for it…It definitely has a good payoff, but it requires an extra effort."* (T4).

Teacher's confidence and mind-set were also factors related to teachers' use of AR, *"The biggest problem primary teachers have is when technology doesn't work, they don't feel confident, and they are losing that confidence."* (T1). While the other teacher explained, if a teacher implemented methods of independent learning, the adoption of AR would fit seamlessly, *"I believe, that would go back to their teaching practices because in a traditional setting when they place themselves at the center and they just ask the students to listen to themselves and follow their instructions all the time, where there is a lack of independent learning skills, I believe that's when it could be problematic is when they could have that kind of fear. But in a classroom where a teacher sets their routines well from the beginning and if students are learning independently, then I believe the transition would be much smoother."* (T2).

Frequency of AR use varied between teachers from once a week to as often as there was a use for it or the equipment was available, *"…as long as I find more links more relevant content to use it, I would make effort to implement it."* (T2), *"I think any innovation in a classroom should be tried as long as you can find a use for it"* (T1). Teachers had many recommendations for improving educational AR applications. First, developing more AR applications for educational purposes, where AR content is related to curriculum, "*The game designs that I think are essential. So, the games that are going to be designed need to actually meet the needs of the curriculum"* (T3). Second, developing AR applications that are co-designed with educators, *"I would like more applications made by educators or made for education, that were more like easily accessible"* (T4). Third, demonstrating and training how an AR application can directly map with the curriculum may encourage more educators to use it, *"Oh definitely if it was PDS about particular augmented reality and how they could work, how they could fit into curriculum in an easy way that would definitely help a lot of educators approach it, I think."* (T4). Fourth, companies should pilot there products as soon as possible to get more people involved in using the application, *"with new technologies like AR and VR, the companies that design it really need to try and pilot it, get it out to classrooms and educational facilities and universities and just promote its use and try getting a whole lot of people using something new rather than wasting so much time developing instructions how to use it."* (T1). Fifth, educational institutions should release products as soon as possible and not worry so much about the rules of using it, as this limits creativity, *"…education is too concerned about this will affect a child. The main problem is we have to write all these rules about how we can use something before we use it, and the rules stop us from using it in a creative way. It's not always about being safe, sometimes the education departments want to make sure they cover every base before they release a product and sometimes, they just need to release it."* (T1). Sixth, one teacher believed that school culture can restrict the use of AR, she compares between a "teach to test" culture vs an "open approach learning" culture, *"…teachers tend to fall back on the common more rote learning models to really get specific facts across. I think it's because of the teach to test culture that a lot of schools have where we just want them to know the answer to the test like, so they test well, and test scores are all it's important, I think, perhaps in somewhere more open like a Montessori school or school kind of like ballpark where they are approaching learning in a different way that it could possibly flourish a lot more."* (T4).

5. Discussion and Conclusions

In this study we have explored K-12 teachers' opinions on their current experiences with AR in the classroom. The aim was to find out how AR is currently being adopted and what advantages and challenges are present with the implementation. Through a qualitative study, with six participants we found teachers believed that AR is a valuable learning experience, and they will continue to use it as long as there is a purpose for it.

In this paper we found that majority of the advantages that the teachers mentioned comply with existing literature. Further, we found that AR provided new ways to express learning and AR provided means for school and community engagement such as a school art exhibition where students of all ages and staff engaged with the artworks.

While the teachers believed that more AR and 3D content creation will be used widely in the future, our findings show that teachers current use of AR and 3D content was limited due to challenges, that involve curriculum preparation that align with using such technologies; the lack of time to learn and use them to build curriculum; and the lack of support provided to teachers to trial such new ways to design curriculum. Further, students' age and the lack of space, equipment, and technical issues

restricted use of the technology. In particular, the most emphasized challenges were curriculum fit and how to map the AR application with curriculum. Suggestions to solve the challenge of curriculum fit could be articulated from different viewpoints. For example, (1) developing more educational AR applications that fit the curriculum, (2) developing AR applications that are co-designed with educators and students, (3) Training and demonstrating how AR maps with curriculum. These suggestions from the teachers comply with the suggestions found in previous literature, however, this indicates the importance of these aspects and the absence of strong solutions in place.

Teachers' future use highly depends on the purposefulness design of AR applications. Teachers' recommendations fall under this umbrella and suggest: (1) More AR applications developed for educational purposes, (2) Co-designed educational AR applications, (3) Training and demonstrating mapping to curriculum and (4) Considering an open approach learning culture to harness students self-learning. These recommendations are complimentary to previous studies. While, other recommendations were targeted to companies and educational institutions respectively, (5) Conducting pilot studies rapidly to involve as much users as possible with the application and (6) Less restricting rules of how to use a technology as rules usually limit creativity.

The results of this study can help schools understand what challenges teachers have faced with AR and provide more support throughout the implementation. Nevertheless, it showed how vital it is for companies, designers, and developers to work close with teachers and students to develop efficient educational AR applications that line seamlessly in the classroom environment considering the challenges that they may face. This study was the first step towards a larger study where we will co-design an AR application with teachers and implement and test it with the students. This will help us understand how to serve the curriculum and find subjects or concepts that are usually difficult for teachers to deliver and examine whether the AR technology can resolve that difficulty.

Acknowledgements

This study was funded under a scholarship from the Saudi Arabian Government represented by King Faisal University. We would like to thank all the teachers for their valuable input to this study.

References

Akçayır, Murat, and Gökçe Akçayır. 2017. "Advantages and Challenges Associated with Augmented Reality for Education: A Systematic Review of the Literature." *Educational Research Review* 20:1–11. doi: 10.1016/j.edurev.2016.11.002.

Alalwan, Nasser, Lim Cheng, Hosam Al-Samarraie, Reem Yousef, Ahmed Ibrahim Alzahrani, and Samer Muthana Sarsam. 2020. "Challenges and Prospects of Virtual Reality and Augmented Reality Utilization among Primary School Teachers: A Developing Country Perspective." *Studies in Educational Evaluation* 66:100876. doi: 10.1016/j.stueduc.2020.100876.

Alkhattabi, Mona. 2017. "Augmented Reality as E-Learning Tool in Primary Schools' Education: Barriers to Teachers' Adoption." *International Journal of Emerging Technologies in Learning (IJET)* 12(2):91–100.

Altinpulluk, Hakan. 2019. "Determining the Trends of Using Augmented Reality in Education between 2006-2016." *Education and Information Technologies* 24(2):1089–1114. doi: 10.1007/s10639-018-9806-3.

Azuma, Ronald T. 1997. "A Survey of Augmented Reality." *Presence: Teleoperators & Virtual Environments* 6(4):355. doi: 10.1162/pres.1997.6.4.355.

Bacca, Jorge, Silvia Baldiris, Ramon Fabregat, Sabine Graf, and Kinshuk. 2014. "Augmented Reality Trends in Education: A Systematic Review of Research and Applications." *Journal of Educational Technology & Society* 17(4):133–49.

Banerjee, Gargi, and Siddhesh Walunj. 2019. "Exploring In-Service Teachers' Acceptance of Augmented Reality." Pp. 186–92 in *2019 IEEE Tenth International Conference on Technology for Education (T4E)*.

Billinghurst, Mark, and Andreas Duenser. 2012. "Augmented Reality in the Classroom." *Computer* 45(7):56–63. doi: 10.1109/MC.2012.111.

Braun, Virginia, and Victoria Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3(2):77–101. doi: 10.1191/1478088706qp063oa.

Chang, Rong-Chi, Liang-Yi Chung, and Yong-Ming Huang. 2016. "Developing an Interactive Augmented Reality System as a Complement to Plant Education and Comparing Its Effectiveness with Video Learning." *Interactive Learning Environments* 24(6):1245–64. doi: 10.1080/10494820.2014.982131.

Chang, Shao-Chen, and Gwo-Jen Hwang. 2018. "Impacts of an Augmented Reality-Based Flipped Learning Guiding Approach on Students' Scientific Project Performance and Perceptions." *Computers & Education* 125:226–39. doi: 10.1016/j.compedu.2018.06.007.

Chen, Peng, Xiaolin Liu, Wei Cheng, and Ronghuai Huang. 2017. "A Review of Using Augmented Reality in Education from 2011 to 2016." Pp. 13–18 in *Innovations in Smart Learning*, *Lecture Notes in Educational Technology*, edited by E. Popescu, Kinshuk, M. K. Khribi, R. Huang, M. Jemni, N.-S. Chen, and D. G. Sampson. Springer Singapore.

Cheng, Kun-Hung, and Chin-Chung Tsai. 2013. "Affordances of Augmented Reality in Science Learning: Suggestions for Future Research." *Journal of Science Education and Technology* 22(4):449–62.

Chiang, Tosti H. C., Stephen J. H. Yang, and Gwo-Jen Hwang. 2014. "An Augmented Reality-Based Mobile Learning System to Improve Students' Learning Achievements and Motivations in Natural Science Inquiry Activities." *Journal of Educational Technology & Society* 17(4):352–65.

Creswell, John W., William E. Hanson, Vicki L. Clark Plano, and Alejandro Morales. 2007. "Qualitative Research Designs: Selection and Implementation." *The Counseling Psychologist* 35(2):236–64. doi: 10.1177/0011000006287390.

Dalim, C. S. C., A. Dey, T. Piumsomboon, M. Billinghurst, and S. Sunar. 2016. "TeachAR: An Interactive Augmented Reality Tool for Teaching Basic English to Non-Native Children." Pp. 82–86 in *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*.

Dunleavy, Matt, and Chris Dede. 2014. "Augmented Reality Teaching and Learning." Pp. 735–45 in *Handbook of Research on Educational Communications and Technology*, edited by J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop. New York, NY: Springer New York.

Fuchsova, Maria, and Lilla Korenova. 2019. "Visualisation in Basic Science and Engineering Education of Future Primary School Teachers in Human Biology Education Using Augmented Reality." *European Journal of Contemporary Education* 8(1):92–102. doi: 10.13187/ejced.2019.1.92.

Garzón, Juan, Juan Pavón, and Silvia Baldiris. 2019. "Systematic Review and Meta-Analysis of Augmented Reality in Educational Settings." *Virtual Reality* 23(4):447–59. doi: 10.1007/s10055-019-00379-9.

Ibáñez, María-Blanca, and Carlos Delgado-Kloos. 2018. "Augmented Reality for STEM Learning: A Systematic Review." *Computers & Education* 123:109–23. doi: 10.1016/j.compedu.2018.05.002.

Lee, L., C. Chau, C. Chau, and C. Ng. 2017. "Using Augmented Reality to Teach Kindergarten Students English Vocabulary." Pp. 53–57 in *2017 International Symposium on Educational Technology (ISET)*.

Moustakas, Clark. 1994. *Phenomenological Research Methods*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: SAGE Publications, Inc.

Nainggolan, E. R., H. H. Asymar, A. R. A. Nalendra, Anton, F. Sulaeman, Sidik, U. Radiyah, and Susafa'ati. 2018. "The Implementation of Augmented Reality as Learning Media in Introducing Animals for Early Childhood Education." Pp. 1–6 in *2018 6th International Conference on Cyber and IT Service Management (CITSM)*.

Oliveira da Silva, Manoela Milena, Rafael Alves Roberto, Iulian Radu, Patricia Smith Cavalcante, and Veronica Teichrieb. 2019. "Why Don't We See More of Augmented Reality in Schools?" Pp. 138–43 in *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*.

Ozdamli, Fezile, and Damla Karagozlu. 2018. "Preschool Teachers' Opinions on the Use of Augmented Reality Application in Preschool Science Education." *Mišljenja Nastavnika Predškolskog Odgoja o Upotrebi Aplikacije Proširene Stvarnosti u Predškolskom Obrazovanju u Području Prirodnih Znanosti.* 20(1):43–74. doi: 10.15516/cje.v20i1.2626.

Papanastasiou, George, Athanasios Drigas, Charalabos Skianis, Miltiadis Lytras, and Effrosyni Papanastasiou. 2019. "Virtual and Augmented Reality Effects on K-12, Higher and Tertiary Education Students' Twenty-First Century Skills." *Virtual Reality* 23(4):425–36. doi: 10.1007/s10055-018-0363-2.

Pence, Harry E. 2010. "Smartphones, Smart Objects, and Augmented Reality." *The Reference Librarian* 52(1–2):136–45. doi: 10.1080/02763877.2011.528281.

Radu, Iulian. 2014. "Augmented Reality in Education: A Meta-Review and Cross-Media Analysis." *Personal and Ubiquitous Computing* 18(6):1533–43. doi: 10.1007/s00779-013-0747-y.

Santos, Marc Ericson C., Takafumi Taketomi, Christian Sandor, Jarkko Polvi, Goshiro Yamamoto, and Hirokazu Kato. 2014. "A Usability Scale for Handheld Augmented Reality." Pp. 167–76 in *Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology*, *VRST '14*. New York, NY, USA: Association for Computing Machinery.

Sugar, William, Frank Crawley, and Bethann Fine. 2004. "Examining Teachers' Decisions To Adopt New Technology." *Journal of Educational Technology & Society* 7(4):201–13.

Tondeur, Jo, Johan van Braak, Peggy A. Ertmer, and Anne Ottenbreit-Leftwich. 2017. "Understanding the Relationship between Teachers' Pedagogical Beliefs and Technology Use in Education: A Systematic Review of Qualitative Evidence." *Educational Technology Research and Development* 65(3):555–75. doi: 10.1007/s11423-016-9481-2.

Tzima, Stavroula, Georgios Styliaras, and Athanasios Bassounas. 2019. "Augmented Reality Applications in Education: Teachers' Point of View." *Education Sciences* 9.

Li, J. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Design and Effect of AR Sign Language Teaching Software Based on Natural Interaction

Jiangxu LI[a], Peng ZHANG[b], Junting CHANG[c] & Su CAI[a, d*]
[a]*VR/AR+Education Lab, Faculty of Education, Beijing Normal University, China*
[b]*Graduate school of education, Peking University, China*
[c]*University College London, United of Kingdom*
[d]*Beijing Advanced Innovation Center for Future Education, Beijing Normal University, China*
*caisu@bnu.edu.cn

Abstract: Sign language is an important way for deaf people to communicate and learn knowledge. In the process of the promotion and standardization of national common sign language, sign language teaching is becoming more and more important. However, the current sign language teaching is mainly based on multimedia courseware, sign language videos, and sign language action decomposition diagrams, which lack dynamic and all-round display. Therefore, this study focuses on the teaching of alphabetic sign language for young children. Based on platforms and devices such as Unity 3D and Leap Motion, two functions of this sign language teaching software were designed and developed, namely learning function and practice function, and feedback from students, parents and teachers were collected. It enables deaf children and other sign language learners to learn, practice and communicate sign language through Android devices. Finally, the software implements learning function and practice function. Among them, the sign language learning function is intuitive, three-dimensional and interactive, while the practice function has the characteristics of natural interaction and real-time feedback. The average judgment accuracy of the practice function is 90%, calculated through the use effect of eight children, which is relatively high. The interview results show that students, parents and teachers have good feedback on it. They said that although the software needed to be improved, the three-dimensional display of its learning function was very interesting, and the practice function could help students correct their mistakes.

Keywords: Sign language teaching, Natural interaction, Augmented Reality

1. Introduction

According to the World Federation of the Deaf, there are over 70 million deaf people in the world (Murray, 2018). Therefore, sign language is an important means of communication and learning knowledge for deaf people. However, the school does not have enough sign language teachers to provide one-on-one instruction, and deaf children's sign language learning efficiency is low because of the limitation of traditional special education methods (Rastgoo, Kiani, & Escalera, 2021). In addition, parents of deaf children are also faced with the challenge of learning sign language and need suitable resources for learning sign language. Currently available tools to assist sign language learning are mostly based on online courses and mobile devices (Martins, Rodrigues, Rocha, Francisco, & Morgado, 2015). The former are mostly frontal displays and recordings, lacking all-round display. While the latter can facilitate users to learn sign language anytime and anywhere, but they are mostly static pictures of sign language, lacking dynamic and comprehensive display, which is undoubtedly a challenge for deaf children whose comprehension is not fully developed. The technical features of Natural interaction and Augmented Reality (AR) can compensate for the deficiency to a certain extent.

Natural interaction is more intelligent and natural than traditional human-computer interaction that requires the use of external hardware devices (e.g., mouse, keyboard, etc.) (Valli, 2008). AR is a technology that applies virtual information to the real world to achieve a sensory experience beyond reality (Azuma, 1996). The combination of the two enables students to conduct natural three-dimensional sign language learning with the machine, resulting in a more realistic learning experience.

Even through Natural interaction and AR have been initially applied in sign language learning, there are still some issues to be improved. User-friendliness as well as accuracy of Natural interaction in sign language learning still need to be improved. The interaction applications are mostly based on wearable devices and visual capture. The former such like gloves and other wearable somatosensory devices make users feel bulky and inflexible (Jiang, Kang, Song, Lo, & Shull, 2022), while the latter such like Kinect can recognize body movements, but the accuracy of gesture recognition is not high (Nguyen, Falk, & Ebling, 2014). The application of AR technology in sign language teaching lacks natural interactivity. Existing AR teaching software is limited to recognizing the static model of the card, lacking the transformation from hand movements to texts.

Based on the above analysis, we focus on two main objectives, developing an AR sign language teaching software and implementing an empirical study of the AR sign language teaching software. Design and develop an AR sign language teaching software. The AR sign language teaching software has two main functions of sign language learning and sign language practice, combining the advantages of AR and Natural interaction. The sign language learning function is intuitive and three-dimensional, while the sign language practice function is with natural interaction and real-time feedback. Implement an empirical study of the AR sign language teaching software. After the development of AR sign language teaching software, the accuracy of the practice function was tested, and users were interviewed. The data were collected to analyze the advantages and disadvantages of the teaching software.

2. Related Work

2.1 Sign Language Teaching

The World Health Organization reports that 466 million people in the world need rehabilitation to address their "disabling" hearing loss, including 34 million children (WHO, 2021). And Smith's research shows that about 1 in 1,000 children in developed countries is born with a bilateral hearing loss of at least 40 decibels. More serious is deafness. The World Federation of the Deaf reports that there are about 70 million deaf people in the world. Among them, there are more than 20 million deaf people in China, including nearly 800 thousand children under the age of seven. Therefore, attention needs to be paid to issues such as treatment, education, and employment of hearing loss or deaf groups (Smith, Bale, & White, 2005).

In education, the main channel for learning sign language is still in the form of class teaching in special schools. Sign language teachers teach according to the outline requirements of sign language textbooks. This teaching method is difficult for deaf people, especially beginners, to interact with the teacher during the learning process, and the learning efficiency is very low (Rastgoo, Kiani, & Escalera, 2021).

Therefore, with the development of information technology, various technologies are gradually applied in the teaching of deaf children. For example, using information technology to learn through storytelling (Alsumait & Al-Musawi, 2013; Flórez-Aristizábal et al., 2019), developing serious games for gamified sign language learning (Economou et al., 2020) and developing the sign language teaching and translation software or hardware (Alrabiah, AlMuneef, AlMarri, AlShammari, & Alsunaid, 2020). These applications basically use gesture recognition technology for sign language action matching and natural interaction for sign language dialogue (Rastgoo, Kiani, & Escalera, 2021). What's more, the addition of AR enhances the sense of experience and immersion.

2.2 Augmented Reality in Sign Language

AR has the characteristics of virtual and real combination, real-time interaction and three-dimensional registration (Azuma, 1996), which is a bridge connecting the real world and the virtual world (Cai, Wang, Ynag, & Liu, 2016). Based on AR, users can see the real world and virtual objects integrated in the real world and interact with them naturally. At present, AR has been widely used in scientific practice teaching, and has gradually entered the field of special education (Kose & Guner-Yildiz, 2021; Yenioglu, Ergulec, & Yenioglu, 2021).

As shown in Table 1., the current researchers mainly use three kinds of devices such as data gloves, Kinect, and Leap Motion combined with AR to assist in sign language learning. Some studies have shown that this system can effectively improve the sign language learning of deaf students (Deb, Suraksha, & Bhattacharya, 2018).

Table 1. *Applications of natural interaction technologies in learning sign language*

Device	Applications	Conclusions
Data gloves (Amin, Rizvi, & Hossain, 2022)	Chinese Sign Language Thesaurus	Data gloves are required for interaction. High device requirements and inconvenient interaction.
Kinect (Baldeon, Oñate, & Caiza, 2022)	Self-directed learning and practice	Strong interaction support for students, but more limited in the use of equipment (e.g., the distance limited, etc.)
Leap Motion (Chuan, Regina, & Guardino, 2014)	Sign language practice	Higher accuracy, relatively low cost of equipment, but high system requirements for computers.

This study adopts Leap Motion, a lightweight gesture recognition device based on visual capture, which can capture hand movements more accurately and does not rely on a wearable device. There have been many studies applying Leap Motion to special education, training children with autism spectrum disorder or other developmental disabilities (Cai, Zhu, Wu, Liu, & Hu, 2018; Hu, Lee, Tsai, Yang, & Cai, 2020). The development in Leap Motion needs to be more integrated with practical teaching and Chinese sign language. Currently, Leap Motion is developed with high accuracy in sign language (Vaitkevičius et al., 2019). The development of both static (Rakesh, Kovács, Mokayed, Saini, & Pal, 2021) and dynamic (Hisham & Hamouda, 2021) sign language mostly focuses on feature combinations or 30 letters, lacking of the integration with pictorial Chinese sign language. This study starts from the actual needs of Chinese sign language teaching.

Therefore, this study is based on Leap Motion and AR to realize the natural interaction of sign language movements, develop learning and practice functions, and collect feedback from students, parents and teachers.

3. Design and Development

We apply 3D modeling, Natural interaction and AR as the core technology and Android application development technology. The AR sign language teaching software can demonstrate sign language models on Android mobile devices and precisely realize Leap Motion preset functions on PC respectively.

3.1 Application Technological Foundation

3.1.1 3D Modeling

Considering the demand of hand model establishment, we implement Maya as the modeling tool and applied techniques such as bone binding, skinning as well as animation to realize 3D modeling. The modeling progress can be divided into procedures as the establishment of hand models, the adjustment of the joint position and angle parameters, and the animation. To make the model more vivid, we replace the static model with partly dynamic model, which enhances the interactivity, fidelity, accuracy and interestingness of the application.

3.1.2 Natural Interaction

Leap Motion is utilized to accurately capture sign language and realize natural interaction. The sign language learning scenario is established and developed on the Unity 3D platform.

Gesture capture based on Leap Motion: To customize the interactive gestures in real time and define the sign language, we apply Leap Motion controller to obtain the raw data of gestures in the camera's field of view, including the number and the names of fingers and the 3D co-ordinates and the vectors of the palm and joints.

Gesture definition based on Unity 3D: Using the 2018.2.15f1 version of Unity 3D development tool that supports Leap Motion, we compile the script to realize the definition of gestures and the interaction of the context and sign language.

3.1.3 Augmented Reality

This study is based on Unity 3D platform, combined with Vuforia AR Toolkit, overlaying virtual information in real scenes to achieve AR effect.

Unity 3D technological route: Based on Unity 3D platform, the application is developed in Windows system and operates in Android system. In Unity 3D platform, we build scenes according to the requirements, add components, write corresponding scripts and keep debugging and running, and finally realize different functions required for various scenes.

Vuforia technological route: We design the target images of corresponding sign language course, add the 3D static and dynamic model of sign language, and determine the best display position by adjusting the parameters such as model position, angle, size, etc.

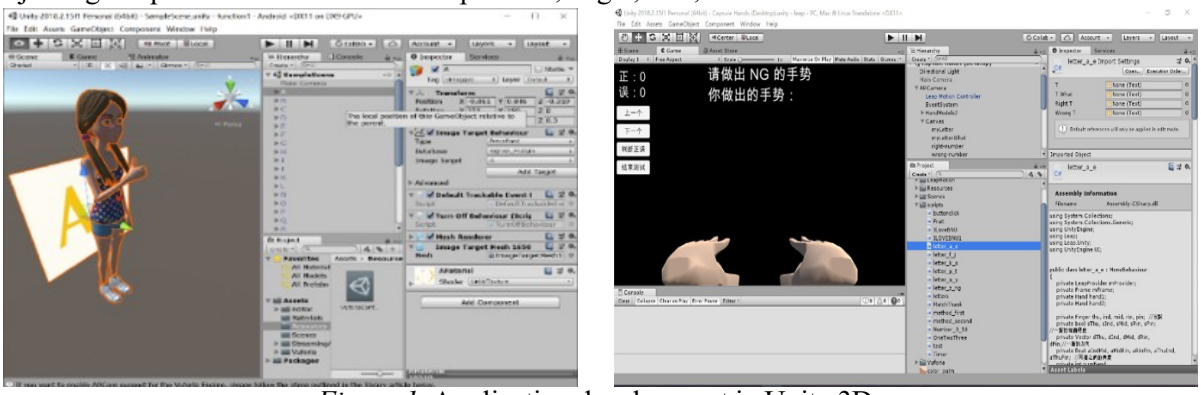


Figure 1. Application development in Unity 3D

3.1.4 Android Application Development

The hardware device of the teaching software is selected as a lower cost Android phone or tablet, so we add Android SDK to realize the design and development in Android platform and finally realize the AR effect in Android devices.

3.2 Application Function

In the learning function part, 30 sign language models and corresponding target images were designed and developed. The learning function of the APP is running on Android system. The target images are recognized by the phone or tablet camera and the 3D static or dynamic models are presented in the form of AR. Users can zoom in and out of the models through the screen and observe the models in 360 degrees for vocabulary learning and review.



Figure 2. Screenshot of learning function

As shown in Figure 3, in the practice function part, the sign language teaching software captures sign language movements in real scenes through a computer camera and matches them with virtual 3D hand model through Leap Motion device, which judges the gesture of the virtual hand and returns the result of correctness or incorrectness. Learners are enabled to practice various sign languages in a customized way according to their learning progress. The practice data is collected so that teachers and students can be aware of the practice status in real time.



Figure 3. Screenshot of practice function

4. Experiment and Results

4.1 Experiment Procedure

To ensure the accuracy of the sign language teaching software and collect participants' using experience, 8 children aged 8-10 (4 deaf children and 4 normal children) with similar sign language learning levels would use the sign language teaching software to learn and practice 30 alphabetic sign language actions.

The experimental procedure is mainly divided into three parts: learning, practicing and interviewing (Figure 4.).

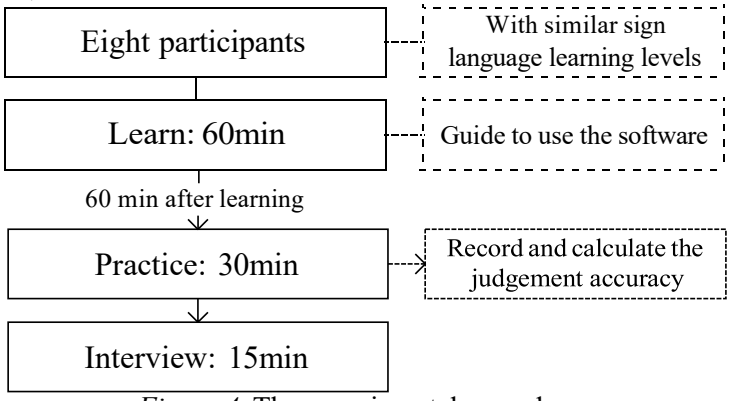


Figure 4. The experimental procedure

Before the experiment, parents of participants were asked to read and sign an informed consent form. In the formal experiment, a sign language teacher from deaf school and a researcher guided the whole experiment. Participants first used the learning function of this software to learn the 30 alphabetic sign language actions. They could observe the sign language action from multiple angles by rotating and zooming, and then, imitate the sign language action. One hour after learning, they used the practice function of this software to review and consolidate sign language actions. When a letter appeared on the screen, the participants made the corresponding sign language action within the range that the camera could capture. Then, the software judged it right or wrong. At this time, the researcher would record whether the software judged correctly, so as to calculate the judgement accuracy of this software. After the practice, participants and their parents would be interviewed about their experience and views of the sign language teaching software.

Calculate the judgment accuracy of the software's practice function through the accuracy formula (1).

$$ACC = (CC+IC)/(CC+IC+CI+II) \quad (1)$$

Among them, CC/IC refer to the number of times that correct/incorrect sign language is correctly recognized, while CI/II refer to the number of times that correct/incorrect sign language is incorrectly recognized. So ACC refers to the proportion of software correctly judged

4.2 Experiment Results

4.2.1 Accuracy

The 30 letters are divided into 5 groups in turn. The average accuracy of the software during the review process of the eight subjects was calculated, as shown in Table 2. The software is 90.0% accurate in judging a certain letter. Among them, the judgment accuracy of the second and fifth groups is low, indicating that the judging scripts of the two still need to be improved.

Table 2. *The accuracy of the sign language teaching software*

Group	Letters	ACC
1	A/B/C/D/E/F	93.3%
2	G/H/I/J/K/L	85.9%
3	M/N/O/P/Q/R	90.6%
4	S/T/U/V/W/X	93.8%
5	Y/Z/ZH/CH/SH/NG	86.6%
	Average	90.0%

4.2.2 Interview Results

Finally, interviews were conducted with the participants who were able to communicate, all parents and the sign language teacher about their feelings and feedback during using or viewing process.

Participants said that the way to scan the cards to see the 3D model is very interesting, and the practice function is also very convenient. I can practice many times and get corrections. Compared with the form of video and pictures, AR software can better meet the visual needs of students (Al-Megren & Almutairi, 2019). But participants also raised some questions: "The sensitivity of the 3D model character is too high during the drag and rotation process, and it takes a lot of time to adjust the angle."

Parents said, "Although, it is worth affirming that the effect of teachers' instructing should be better, with this AR method, children can continue to learn sign language without teachers around." Some parents also suggested that this is also a way for them to learn sign language. Indeed, for families with deaf children, family members also need to study sign language. And compared with deaf children who can study sign language at school, other family members can only study sign language through online classes, sign language books or videos in their spare time (Frank, 2017). And this more visual sign language teaching software undoubtedly provides a new way for them to study sign language. Parents are also concerned about the children's health problems brought about by long-term use of electronic devices for learning.

The sign language teacher believes that the concept of this software is feasible to a certain extent, but if it is to be truly applied in sign language classrooms, more vocabulary and functions are needed, such as learning various related sign languages in different life scenarios. Sign language teachers believe that the concept of this software is feasible to a certain extent, but to really apply it to the sign language classroom, more vocabulary and functions are needed, such as learning various related sign language situations in different lives. Studies have shown that deaf children have difficulty learning sign language out of context (Almutairi & Al-Megren, 2017). And sign language teachers also said that sign language teaching is generally carried out based on a certain story situation or a dialogue scene. In addition, the design of the practice function is very helpful, it would be better if the test method can be more flexible. Regarding the accuracy of the software, the teacher said, "among the 30 alphabetic sign languages in this experiment, individual sign language movements are very similar". So it is difficult for participants to remember and distinguish, and the software may also make mistakes in judgment.

For example, "H" and "I", "M" and "N", “Z” and “ZH” are only slightly different. This is a good explanation for the lower accuracy of the second and fifth groups. That is, there are sign language movements with similar characteristics in the group, but the script does not distinguish them well. In addition, the sign language teacher also said that if the software is used reasonably, it can reduce the burden of one-on-one guidance for the sign language teacher, guiding students to learn sign language and supervising students to practice sign language through the software.

5. Conclusion and Future Work

The purpose of this study is to design and develop alphabetic sign language learning software for deaf children. Then, calculate the accuracy of the software and collect participants' feedback after participants use the software.

The software can assist students in sign language learning without teacher guidance, and also help parents in basic learning of sign language. The learning function of the software enables students to conduct all-round observational learning. Students can act as both "speakers" and "listeners" to observe the characteristics of the front and back of sign language movements. The practice function of the software allows students to get feedback on their practice, so that students can carry out targeted practice according to their weak sign language vocabulary, improving the effect and efficiency of practice.

However, some challenges have also been raised. Accuracy and interface of the software need to be further improved. In the future, it is necessary to develop more sign language vocabulary in combination with life scenarios to realize sign language situational learning. This experiment recruited a small number of participants, and no quasi-experimental research or case study was conducted to explore the learning effect of participants, only the accuracy and experience of use were investigated.

Moreover, sign language expression is not only dependent on the movements of the hands, but also related to the expresser’s expression at that time, the speed of the movement, and the context at that time (Shaffer, 2018). We need to note that technology can reduce the burden on teachers but not replace teachers, especially in the field of special education, which requires more humane care.

Acknowledgements

This research was supported by National Natural Science Foundation of China (61977007).

References

Al-Megren, S., & Almutairi, A. (2019). User Requirement Analysis of A Mobile Augmented Reality Application to Support Literacy Development Amongst Children with Hearing Imparirments. *Journal of Information and Communication Technology-Malaysia, 18*(2), 207-231.

Almutairi, A., & Al-Megren, S. (2017). Augmented Reality for the Literacy Development of Deaf Children: A Preliminary Investigation. Paper presented at the *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, Baltimore, Maryland, USA.

Alrabiah, M., AlMuneef, H., AlMarri, S., AlShammari, E., & Alsunaid, F. (2020). A Robotic Hand for Arabic Sign Language Teaching and Translation. Paper presented at the *Intelligent Systems and Applications*, Cham.

Alsumait, A., & Al-Musawi, Z. S. (2013). Creative and innovative e-learning using interactive storytelling. *International Journal of Pervasive Computing and Communications, 9*(3), 209-226.

Amin, M. S., Rizvi, S. T. H., & Hossain, M. M. (2022). A Comparative Review on Applications of Different Sensors for Sign Language Recognition. *8*(4), 98.

Azuma, R. T. (1996). A Survey of Augmented Reality. *Presence Teleoperators Virtual Environments, 6*(4), 355–385.

Baldeon, K., Oñate, W., & Caiza, G. (2022). Augmented Reality for Learning Sign Language Using Kinect Tool. Paper presented at the *Developments and Advances in Defense and Security*, Singapore.

Cai, S., Wang, P., Ynag, Y., & Liu, E. (2016). Review on Augmented Reality in Education. *Journal of Distance Education (in Chinese), 34*(05), 27-40.

Cai, S., Zhu, G., Wu, Y.-T., Liu, E., & Hu, X. (2018). A case study of gesture-based games in enhancing the fine motor skills and recognition of children with autism. *Interactive Learning Environments, 26*(8), 1039-1052.

Chuan, C. H., Regina, E., & Guardino, C. (2014, 3-6 Dec. 2014). American Sign Language Recognition Using Leap Motion Sensor. Paper presented at the *2014 13th International Conference on Machine Learning and Applications*, Detroit, MI.

Deb, S., Suraksha, & Bhattacharya, P. (2018). Augmented Sign Language Modeling(ASLM) with interaction design on smartphone - an assistive learning and communication tool for inclusive classroom. *Procedia Computer Science, 125*, 492-500.

Economou, D., Russi, M. G., Doumanis, I., Mentzelopoulos, M., Bouki, V., & Ferguson, J. (2020). Using Serious Games for Learning British Sign Language Combining Video, Enhanced Interactivity, and VR Technology. *Journal of Universal Computer Science, 26*(8), 996-1016.

Flórez-Aristizábal, L., Cano, S., Collazos, C. A., Benavides, F., Moreira, F., & Fardoun, H. M. (2019). Digital transformation to support literacy teaching to deaf Children: From storytelling to digital interactive storytelling. *Telematics and Informatics, 38*, 87-99.

Frank, A. (2017). Deaf Families' Unique Experiences and Obstacles. *Journal of Social Work in Disability & Rehabilitation, 16*(3-4), 216-229.

Hisham, B., & Hamouda, A. (2021). Arabic sign language recognition using Ada-Boosting based on a leap motion controller. *International Journal of Information Technology, 13*(3), 1221-1234.

Hu, X., Lee, G. T., Tsai, Y.-T., Yang, Y., & Cai, S. (2020). Comparing Computer-Assisted and Teacher-Implemented Visual Matching Instruction for Children with ASD and/or Other DD. *Journal of Autism and Developmental Disorders, 50*(7), 2540-2555.

Jiang, S., Kang, P., Song, X., Lo, B. P. L., & Shull, P. B. (2022). Emerging Wearable Interfaces and Algorithms for Hand Gesture Recognition: A Survey. *IEEE Reviews in Biomedical Engineering, 15*, 85-102.

Kose, H., & Guner-Yildiz, N. (2021). Augmented reality (AR) as a learning material in special needs education. *Education and Information Technologies, 26*(2), 1921-1936.

Martins, P., Rodrigues, H., Rocha, T., Francisco, M., & Morgado, L. (2015). Accessible Options for Deaf People in e-Learning Platforms: Technology Solutions for Sign Language Translation. *Procedia Computer Science, 67*, 263-272.

Murray, J. (2018). World Federation of the deaf. Retrieved from http://wfdeaf.org/our-work/

Nguyen, P. L., Falk, V., & Ebling, S. (2014). Building an Application for Learning the Finger Alphabet of Swiss German Sign Language through Use of the Kinect, Cham.

Rakesh, S., Kovács, G., Mokayed, H., Saini, R., & Pal, U. (2021, 14-15 June 2021). Static Palm Sign Gesture Recognition with Leap Motion and Genetic Algorithm. Paper presented at the *2021 Swedish Artificial Intelligence Society Workshop (SAIS)*.

Rastgoo, R., Kiani, K., & Escalera, S. (2021). Sign Language Recognition: A Deep Survey. *Expert Systems with Applications, 164*, 113794.

Shaffer, I. R. (2018). Exploring the Performance of Facial Expression Recognition Technologies on Deaf Adults and Their Children. Paper presented at the *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, Galway, Ireland.

Smith, R. J. H., Bale, J. F., & White, K. R. (2005). Sensorineural hearing loss in children. *Lancet, 365*(9462), 879-890.

Vaitkevičius, A., Taroza, M., Blažauskas, T., Damaševičius, R., Maskeliūnas, R., & Woźniak, M. (2019). Recognition of American Sign Language Gestures in a Virtual Reality Using Leap Motion. *9*(3), 445.

Valli, A. (2008). The design of natural interaction. *Multimedia Tools and Applications, 38*(3), 295-305.

WHO. (2021). Deafness and hearing loss. Retrieved from https://www.who.int/en/news-room/fact-sheets/detail/deafness-and-hearing-loss

Yenioglu, B. Y., Ergulec, F., & Yenioglu, S. (2021). Augmented reality for learning in special education: a systematic literature review. *Interactive Learning Environments*, 1-17.

Hwang, W. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Designing a Recommender System for Mobile Applications Focusing on Relative Importance Weights of Learner-related Variables

Woorin HWANG[a], Hyo-Jeong SO[a*], Chiyoung SONG[b], & Hyeji JANG[a]
[a]*Ewha Womans University, Republic of Korea,* [b]*MOTOV, Republic of Korea*
*hyojeongso@ewha.ac.kr

Abstract: To embrace why and how people learn and how to combine learner characteristics for recommending foreign language learning mobile applications (apps), this research presents a recommender system based on the relative importance weights of learner-related variables. In developing the system, 100 adult learners used 4-6 foreign language learning apps, resulting in 557 user-satisfaction data to calculate the relative importance of 14 learner-related variables in four categories: (a) demographic information, (b) motivational orientation for language learning (instrumental/integrative), (c) learning style, and (d) learning experience. The result showed that the model considering the relative importance weights of learner-related variables outperforms the dummy model in predicting users' satisfaction with the apps.

Keywords: Recommender System, Mobile Applications, Feature Importance, Learner Variables

1. Introduction

In the flood of information and the expansion of informal learning spaces, the learners face up with the difficulty in selecting appropriate learning resources since they cannot judge what kind of learning information is meaningful to them. Also, as most of the app recommendations have been developed in the commercial area, it is uneasy for them to adopt the recommendations directly for educational or learning purposes. While these commercial algorithms are grounded on general factors such as demographic information, installed apps, age groups, user device information, and preference (Cheng et al., 2016), Garcia-Martinez and Hamou-Lhadj (2013) emphasized that educational recommender systems should be distinguished from those developed in the business area in terms of goal, context, the influence of pedagogical factors, and classification of users. Essalmi et al.(2015) also mentioned that, in education, personalized recommendation needs to focus on the combination of learners' characteristics coupling with learning objects. With this backdrop, this research proposes a method to design an educational recommender system of apps for foreign language learning. First, the newly-developed system reflects the fundamental factors for learning ― why and how people learn. Moreover, to harmonize learner-related variables considering that not all variables are equally important, it is based on the relative importance of learner-related variables in model estimation.

2. Method

2.1 Participants

The participants were recruited from online bulletin boards and included 100 adult learners (33 males and 67 females) in Korea, aged 19 to 39. At the time of the study, 53 participants were students, followed by 32 working as office or professional workers, and 9 unemployed. In terms of their educational background, they include 38 undergraduates, 41 graduates, and 21 with master's or higher degrees.

2.2 Constructing a Recommender System

2.2.1 Data Collection

First, the participants self-reported their characteristics in four categories: (i) demographic information (age, gender, education level, job-related factors), (ii) language learning experience, (iii) learning style by Felder-Solomon's Index of Learning Style (ILS), and (iv) motivational orientation for language learning (instrumental/integrative). Second, four to six apps among 18 foreign language learning apps were randomly assigned to each participant. The 18 were selected from App Store (iOS) and Google as they rated three points or higher (out of five). As the result, 557 satisfaction data (1= "not satisfied at all" to 5= "very satisfied") were collected. All procedure was conducted after the approval by the IRB committee of the researchers' institution.

2.2.2 Relative Importance and Algorithm

RandomForestRegressor (Scikit-Learn 1.0.2., 2022) was used in the data analysis process to calculate the relationship between user satisfaction with 18 apps and 14 sub-variables. Table 1 presents the relative importance weights of each variable according to the rank of importance.

Table 1. *Description of Learner-related Variables and Relative Importance Weights*

Category	Code ID	Description	Importance Weight(Rank)
Demographic Information	AgeR	*Age group* (1: 19-24, 2:25-29, 3:30-34, 4: 35-39)	0.0657(6)
	Gen	*Gender*: Whether the learner is male (0) or female (1)	0.0569(7)
	Schooling	*Education*: The learner's highest education completed (1: high school level or under, 2: college/university level, 3: graduate school or above)	0.0514(8)
	N-yn	*Job*: Whether the learner has a job (1) or not (0)	0.0247(13)
	S-yn	*Job*: Whether the learner is a student (1) or not (0)	0.0263(12)
Learning Style	LS1	*Information Processing*: active (1) or reflective (-1)	0.0658(5)
	LS2	*Information Perception*: intuitive (1) or sensory (-1)	0.0399(9)
	LS3	*Information Reception*: visual (1) or verbal (-1)	0.0245(14)
	LS4	*Information Understanding Progression*: global (1) or sequential (-1)	0.0337(10)
Language Learning Motivation	MIT	*Motivation in total:* The extent to which the learner is motivated for learning a language	0.1674(2)
	MI1	*Integrative Motivation:* The extent to which the learner is oriented toward integrative motivation (e.g., understanding the target culture)	0.1520(3)
	MI2	*Instrumental Motivation*: The extent to which the learner is oriented toward instrumental motivation (e.g., reaching practical goals such as getting a job)	0.1808(1)
Language Learning Experience	LE	*Whether the learner has ever used any mobile apps* for language learning (1) or not (0)	0.0283(11)
	LLPW	*The frequency of language learning per year*	0.0826(4)
Sum of Importance Weights			1.0

The top five variables with high importance weights are all three variables in language learning motivation (MI2=.1808, MIT=.1674, MI1=.1520), usage frequency in the learning experience (LLPW=.0826), and perception in learning style (LS1=.0658).

2.2.3 System Interface and Evaluation

The proposed system can be described in Figure 1. In calculating the similarity between learners, the weighted cosine similarity was employed by reflecting the importance weights of the variables. The interpolation in the learners' satisfaction scores that they did not rate was made with Inverse Distance Weighting(IDW) — giving greater weight to the value of the former learner with the closest similarity.

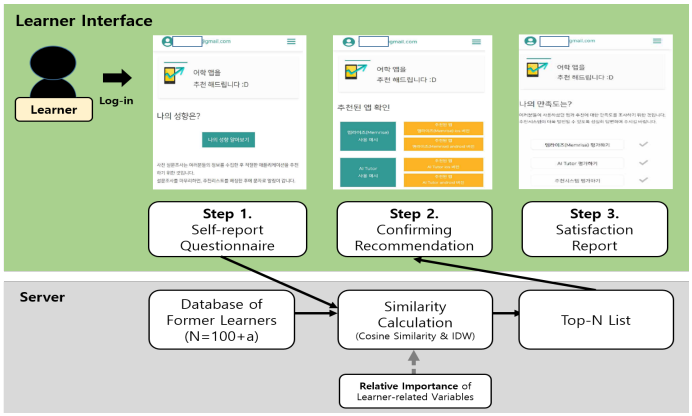


Figure 1. Recommender System Architecture

The performance of the developed system was compared with the dummy model which postulated recommending every 18 apps to everyone. As the result, the former (Precision=.647, Recall=.403, F1-score=.494) outperformed the latter (Precision=.564, Recall=.388, F1-score=.459).

3. Conclusion and Future Work

This research presents how the relative importance weights among learner-related variables can affect a recommender system of language learning apps. Overall, we confirmed that learner characteristics in the affective domain, such as language learning motivation, learning style (Information Processing), and app usage frequency carry higher importance weights than the general demographic information such as gender and job. This led to the result that the proposed system reflecting the relative importance of the variables was better in predicting learner satisfaction than the random recommendation. Some limitations of this study should also be noted. First, since the importance weights were calculated with a limited number of variables, we need to expand the model with more diverse variables to develop more meaningful recommender systems. Second, this study used self-reported questionnaires to determine learning style and motivation, requiring learners of extended time and effort. Thus, future research on educational recommender systems should take into account utilizing proxy indicators or automatic detectors to capture learning motivations and styles. Third, the complex interaction among learner variables can be detected by adopting more advanced ways for analysis, such as Support Vector Machine, Gradient Boosting, or Deep Neural Network Model.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT)(No. 2020R1F1A1073469).

References

Cheng, H. T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., ... & Shah, H. (2016, September). Wide & deep learning for recommender systems. *Proceedings of the 1st workshop on deep learning for recommender systems* (pp. 7-10).

Essalmi, F., Ayed, L. J. B., Jemni, M., & Graf, S. (2015). Generalized metrics for the analysis of E-learning personalization strategies. *Computers in Human Behavior, 48*, 310-322.

Garcia-Martinez, & S., Hamou-Lhadj, A. (2013). Educational recommender systems: A pedagogical-focused perspective. In *Multimedia services in intelligent environments* (pp. 113-124). Heidelberg: Springer

Raste, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Learning Through Patient Stories: Fostering Sensemaking in A Technology Enhanced Learning Environment for Clinical Diagnosis

Sunita RASTE [a*], Sahana MURTHY [a]
[a]*Indian Institute of Technology Bombay, India.*
sunita_20@iitb.ac.in

Abstract: Clinical diagnosis of genetic disorder requires sensemaking based on the symptoms of a patient. Making a diagnosis needs information gathering from the patient, interpreting the information, conducting tests and advice on further treatment. Furthermore, the experience of diagnosis requires developing a personal connect with the patient, which allows immersion in the case being investigated. This has been known to promote learners' attention so that they can better notice relevant information, and minute or unusual details. It can help them recall and relate to their own previous experiences. *Karyotype* learning environment provides learners with an authentic situation of solving video-based cases of patients suffering from genetic disorders . Students perform the clinical diagnosis by building and revising scientific explanations for the given cases while going through a series of learning activities. In this paper, we report a study conducted with 26 bioscience students to understand whether working with karyotype improves students' sensemaking and how the design of karyotype supports students' learning experience.

Keywords: Technology enhanced learning environment, sensemaking, clinical diagnosis, case-based reasoning, empathy

1. Introduction

Clinical diagnosis of genetic disorders caused due to abnormalities such as change in the number or structure of entire chromosomes, or a specific part of the chromosome carrying a particular gene. This requires sensemaking based on the symptoms of a patient, to make decisions regarding the treatment. Students need sensemaking and reasoning skills to build a scientific explanation by analysing the symptoms and connecting it with chromosomal abnormalities. However, students often learn genetics as separate, unrelated concepts, and face difficulty in understanding how they are related (Kılıç and Sağlam, 2014). Due to the lack of experience and not being immersed in the phenomena being investigated, novice learners, when introduced to new concepts, perceive them as an activity of mere memorization of facts and procedures (CTGV,1990). Though sensemaking and reasoning are essential to the way scientists construct knowledge, the majority of science undergraduate students lack these skills (Odden and Russ, 2018). This is partly due to the undergraduate science laboratories being fact-laden, non-inquiry based, with activities that do not support the development of sensemaking and reasoning skills explicitly (Daempfle,2006). Apart from these cognitive aspects, clinical diagnosis also involves affective aspects such as forming a personal connect with the cases. This is important to be immersed and involved in the process of diagnosis. Being empathetic towards patients and their stories helps in relating the given situation to prior knowledge and experiences, to build explanations regarding the new situation encountered. Empathy helps the learner to be more attentive and considerate towards minute details that might be significant for the diagnosis.

Karyotype is a web-based technology enhanced learning environment (TELE) where students are given the task of diagnosing the genetic disorders of patients while assuming the role of a geneticist. It uses a case-based reasoning approach where the students are given a set of symptoms and asked to explain them with the help of given information to suggest a diagnosis. *Karyotype* provides hands-on learning experience of the diagnosis process where students build and revise their explanations while solving cases of genetic disorders through a series of learning activities (Raste, Deep and Murthy, 2021).

In this paper, we describe the findings from a research study conducted with 26 bioscience majors' students interacting with *Karyotype.* The focus of this research study was to understand:

1. Does working with *Karyotype* improve students' sensemaking?
2. How does the design of *Karyotype* support students' learning experience?

2. Theoretical basis and design of *Karyotype*

Sensemaking about a phenomenon involves collecting observations, analysing the data and constructing interpretations. In this study, we use the definition of sensemaking as "s*ensemaking* is a dynamic process of building or revising an explanation in order to "figure something out"—to ascertain the mechanism underlying a phenomenon in order to resolve a gap or inconsistency in one's understanding." (Odden &Russ, 2018). Sensemaking and reasoning are closely associated in order to strengthen the interpretations. Learning to make sense of phenomena in the classroom can help the learners by preparing them to make sense of unfamiliar, complex, or seemingly inconsistent scientific phenomena that they may encounter in real life .

Case-Based Reasoning (CBR) approach allows students to learn by doing problem solving and other activities that pique their interest and apply what they've learned in a way that provides immediate feedback (Kolodner et al, 2003). CBR has been used in medicine for diagnosis as the methodology of CBR systems closely resembles the thought processes of a physician. In CBR, the remembered cases are used as a means of efficient problem-solving. Here elicitation becomes a task of gathering case histories and implementation is reduced to identifying significant features that describe a case (Watson and Marrir, 1994). CBR is typically a cyclical process comprising the four "REs". A new problem is matched against cases in the case base and one or more similar cases are retrieved. A solution suggested by the matching cases is then reused and tested for success. However, the previous diagnoses cannot be applied to new cases as-is, they need to be adapted to fit the new case/situation. Unless the retrieved case is a close match the solution will probably have to be revised producing a new case that can be retained. (Kolodner, 1992; Watson & Marrir, 1994). The cases presented in the form of an authentic and engaging narration can help learners understand and empathize with the patients. Empathy is an emotional experience between an observer and a subject in which the observer, based on visual and auditory cues, identifies and transiently experiences the subject's emotional state (Hirsch, 2007).

Karyotype is designed for the target learners of biology majors students. The learning environment consists of a series of learning activities that helps the learners to get a walkthrough of the process followed for clinical diagnosis of chromosomal abnormalities. The system makes use of interactivity to enhance learners' engagement with the content and provides scaffolds for the learners to help them make progress in the learning activities. A patient narrative is presented in the form of a video case to establish a personal connect with the content. Each case contains in-video questions where learners are being pointed towards details related to the symptoms. This puts an emphasis on relevant observations that help in making a correct diagnosis.

3. Study design

3.1 Participants and context of study

The participants of this study were 26 bioscience majors' students from Kerala, Karnataka and Maharashtra, India. The context of this study was clinical diagnosis of genetic disorders caused due to chromosomal abnormalities. The problems are given in the form cases presented as patients' narrative. Students need to understand the context of the problem, make relevant observations regarding the case, perform basic tests for diagnosis and come up with an explanation that provides reasons and justification for a possible diagnosis associated with the given case. All the participants of this study had basic prior knowledge of the domain as they had undergone an introductory course in genetics during their first/second year of undergraduate studies.

3.2 *Procedure*

This study was conducted as a part of a workshop for bio-science students. A pre-post study design was followed as we intend to check the effectiveness of pedagogy and design elements used in *Karyotype* to support sensemaking. The workshop was conducted online using Google Meet as the video conferencing platform, in the presence of the instructor in a supervised setting. *Karyotype* was accessed as a Google Sites using the standard web browser for capturing data of the learning activities related to the task of clinical diagnosis. The study had five steps and followed a pre-post design. A registration form along with informed consent was floated among bioscience undergraduate and postgraduate students. The registration form included the personal and academic details of the participants. There were 2 close-ended questions to check students' prior knowledge about the topic, and a confidence questionnaire. Participants who have undergone an introductory course on genetics during their bachelors or masters were recruited for this study on the basis of the registrations received. Participants were asked to solve a minimum of 5 cases. 86.5 % participants solved >5 cases (lowest-2, highest-9) spending an approximate of 20 minutes on each case during their interaction with *Karyotype.* For each case solved, participants created a final diagnosis report. After the intervention, participants were given one common case in the post-test where all the scaffolds were removed. A post workshop interview of 20-25 minutes was conducted with each participant.

3.3 Data collection and analysis

Mixed methods approach was used for the data collection. Quantitative data was collected through the diagnosis reports prepared as artifacts by students during the intervention and post-test. Qualitative data was collected through semi-structured interviews with students.

3.3.1 *Quantitative data*

The diagnosis reports included notes and observations made for the respective cases. The pre-test consisted of two questions which asked the learners to analyse the given problem and interpret it in the context of a chromosomal disorder. The purpose of these pre-test questions was to understand how learners make sense of the given problem and the role of providing an authentic context. Student artifacts (diagnosis reports) for the pre-test, intervention and post-test were evaluated using a 4-point rubric scale (0-missing, 1-needs work, 2-adequate and 3-target. Three criteria corresponding to the disciplinary practices involved in the process of clinical diagnosis were used to check if students are able to:

1. Make explicit and relevant clinical observations
2. Connect the observable clinical symptoms with the chromosomal condition
3. Explain how chromosomal changes lead to clinical symptoms

These criteria are consistent with the disciplinary practices of making relevant observations (R1), making connections across concepts (R2) and giving a scientific explanation respectively (R3).
R1 checks if learners have pointed out relevant symptoms and noted specific observations, while R2 checks if a clear connection has been stated between the observed symptoms and the chromosomal condition identified. R3 checks for reasoning that explains and justifies the diagnosis by pointing out the cause-effect relationship between the observed symptom and the chromosomal condition. To ensure the consistency of the rubric used for evaluation of students' responses, we carried out an inter-rater reliability test with 2 independent raters. The raters were given information about *Karyotype*, the problem context and the learning activities. They were provided with the rubric that explained the target level descriptor corresponding to each score on the 4-point scale of 0 (missing) to 3 (Target). The raters were also provided with an exemplar sample response for their reference. Each rater was first asked to individually rate 6 sample reports on the 3 criteria mentioned in the rubric (Ref. Table 1). The researcher had separate meetings with the raters to clarify their doubts regarding the rubric and TELE, if any. After one round of individual rating, the two raters discussed their scores with each other to resolve the discrepancies and reach a consensus. Both raters submitted their revised scores based on the discussion, which were then used to calculate the percentage agreement and Cohen's kappa. A percentage

agreement of 83.33 % was observed between the two raters, giving a kappa value of 0.66 indicating substantial agreement.

3.3.2 *Qualitative data*

Semi-structured interviews were conducted with 22 students based on their availability and willingness to participate in the interviews. The conversations were recorded and transcribed to obtain the verbatim. The intended purpose of these interviews was to get insights regarding students' experience of interacting with *Karyotype*. Students were probed about the procedure they followed to solve the given cases, what features of the TELE helped them in the process of diagnosis and how? Students were also asked about their prior knowledge of the topic, their prior classroom learning experience and how it is similar or different from the learning activities in *Karyotype*. Each interview lasted for about 15-20 minutes. These interview conversations were coded and analyzed using a thematic analysis approach. The unit of analysis was one sentence/series of sentences related to a broad theme of questions asked by the researcher. The broad themes included questions related to the prior knowledge, features of TELE, procedure followed for diagnosis and how did they build a plausible explanation about the case.

4. Findings

4.1 Students' performance on sensemaking criteria

We compared students' pre and post test scores to see the difference in their performance on each criterion. The paired t-test result shows significant difference for all 3 criteria suggesting that the intervention did help the students in sensemaking. Table 2 summarizes the results from the paired t test between the pre-test and post-test scores, represented graphically in figure 2 to indicate students' learning gains.

Table 2*: Students' learning gain*

Criteria	Pre-test	Post-test	Paired t test	Interpretation
			T value	
R1: Making relevant observations	0.27	1.27	5.70	Significant
R2: Making connections across concepts	0.35	1.04	3.49	Significant
R3: Giving a scientific explanation	0.00	0.54	4.24	Significant

4.2 Affordances of the TELE supporting sensemaking.

Thematic analysis of the interview transcripts gave us insights about how the features of *Karyotype* support students to engage in sensemaking while performing the clinical diagnosis. We present excerpts from the interviews where students were seen to engage in the cognitive and affective processes involved in clinical diagnosis. The excerpts indicate how the design features of *Karyotype* aided this process. As sensemaking requires prior knowledge, students' interaction with *Karyotype* builds on their existing knowledge through a set of activities that are different from the regular classroom activities and enhance the learning experience. Sharing her experience with *Karyotype S21* was quoted saying*: "Many of the syndromes which I have not studied in my graduation. We did a theoretical experiment about karyotype. I have studied during my MSc, regarding mutations and chromosomal abnormalities. In this activity I have engaged myself in doing the diagnosis, looking for what are the symptoms and what are the causes."*

As a way of thinking about science, sensemaking is a cognitive process, and a type of discourse (Odden and Russ, 2018). *Karyotype* learning activities are designed to support this cognitive process. Students were seen engaging in sensemaking while performing diagnosis. *S14:"Also, in cri-du-chat one portion of the chromosome was broken, the P arm. And given that it could affect the voice and speech that could have definitely been a reason for the symptoms that appear. The genes that could have been*

deleted or affected in other patients that directly connect to the symptoms that they have."

Students were seen relating the cases in *Karyotype* with their real-life experiences (if any) and empathizing with the people who were a part of the case narrative. An example of one such instance, *S10: "In the common diseases, there was this one Nora baby and it showed the cleft lip. I know one man like that. So, I just remember that guy. Yeah in my experience, what is happening, I used to go there because the surgery was going, he was explaining the experience. He was trying to understand why it was happening, at that time I also didn't know. Yesterday I came to know, ok it is because of this reason."*The process of clinical diagnosis involves not only the cognitive aspect but also the affective aspects as it concerns the lives of patients. Having an empathic approach helps the students to get involved in the diagnosis by being more careful and observant about the given cases.

During sensemaking, experts can spot meaningful patterns that may not be apparent to novices. Thus, students being novices require assistance in mapping their understanding of the concepts. *Karyotype* provides scaffolds for students to complete the diagnosis process. Students found these scaffolds useful and pointed out how it helped them navigate through the process. *S24: "The video gives you a brief idea about the patient's history. So it is better to understand and to judge the karyotype, because visually you can understand what is going on with the patient. The interactive questions, if it is wrong, you will get that there is some mistake or something is missing. So, you have to go back and read and listen carefully."*

5. Conclusion

Making links to the real world or lived experience, coordinating different representations, analyzing the rationality of solutions, and perceiving the problem as a reasonable one to address are all behaviors connected with sensemaking (Chen, Irvin and Syre, 2013). In the efforts to answer RQ1, we observed that students' sensemaking improved after interacting with *Karyotype*. We believe that design of the learning environment has led to this positive change. *Karyotype* is novel in terms of the use of instructional strategy and interactivity introduced to support learners' empathy. The technological affordances provide a more immersive experience of the diagnosis process. A personal connection established between the patients shown in the cases and the learners solving the cases is something that can not be achieved with the traditional pen and paper activity. Further, thematic analysis of student interviews show evidence regarding RQ2 i.e how the design of *Karyotype* supports students' learning experience. The design and pedagogical decision of using video-based cases including patient narrative provided veridical representation of the events which led to the improvement of students observation skills. The use of real-life cases to provide authentic learning scenarios, explores both cognitive and affective aspects of the diagnosis process. In our context, the visual representation of issues faced by patients proved to be more effective over the mere textual representation of facts. It got the learners more involved in the diagnosis process rather than reading and imagining a list of symptoms. Use of scaffolds have shown some significant change in students' abilities to perform tasks related to sensemaking. Design elements such as look-up table, reflection spots, Hints and feedback, prompts scaffolded students' sensemaking while navigating through the learning activities in *Karyotype*.

The focus of this current study was to highlight the design of *Karyotype* in supporting students' sensemaking during the process of clinical diagnosis. Thus, we used pre-post study design to evaluate their performance on certain sensemaking criteria before and after the interaction with *Karyotype*. Since there was no prior exposure to the TELE, the difference observed in the pre-post comparison can be attributed to learning activities in *Karyotype*. However, we acknowledge that the observed positive effect maintained post the withdrawal of scaffolds could be due to the post-test being conducted immediately after the intervention. Going further, to evaluate the effectiveness of *Karyotype* in improving sensemaking, a quasi-experimental study will be conducted along with a delayed post-test to account for the retention of acquired skills. *Karyotype* being aligned with the curriculum, it can be deployed in classrooms as a bridge between the theory classes and laboratory sessions. Instructors can also use it for assessment purposes.

References

The Cognition and Technology Group at Vanderbilt (CTGV). (1990). Anchored Instruction and Its Relationship to Situated Cognition. *Educational Researcher*, *19*(6), 2–10. https://doi.org/10.3102/0013189x019006002

Chen, Y., Irving, P. W., & Sayre, E. C. (2013). Epistemic game for answer making in learning about hydrostatics. *Physical Review Special Topics - Physics Education Research*, *9*(1).

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., ... & Soloway, E. (2018). A scaffolding design framework for software to support science inquiry. In *The Journal of the Learning Sciences* (pp. 337-386). Psychology Press.

Daempfle, P. (2006). The effects of instructional approaches on the improvement of reasoning in introductory college biology: A Quantitative Review of Research, *Journal of College Biology Teaching*, v32 n4 p22-31 Dec 2006.

Herreid, C. F. (Ed.). (2007). *Start with a story: The case study method of teaching college science*. NSTA press.

Hirsch E. M. (2007). The role of empathy in medicine: A medical student's perspective. *The virtual mentor : VM*, *9*(6), 423–427. https://doi.org/10.1001/virtualmentor.2007.9.6.medu1-0706

Jonassen, D. H., & Hernandez-Serrano, J. (2002). Case-based reasoning and instructional design: Using stories to support problem solving. *Educational Technology Research and Development*, *50*(2), 65–77.

Kılıç, D., & Sağlam, N. (2013). Students' understanding of genetics concepts: the effect of reasoning ability and learning approaches. *Journal of Biological Education*, *48*(2), 63–70.

Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, *6*(1), 3–34.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., Puntambekar, S., & Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting Learning by Design into practice. *Journal of the Learning Sciences*, *12*(4), 495–547.

Odden, T. O. B., & Russ, R. S. (2018). Defining sensemaking: Bringing clarity to a fragmented theoretical construct. *Science Education*, *103*(1), 187–205.

Raste, S., Murthy, S., & Deep, A. (2021). Karyotype: An interactive learning environment for reasoning and sensemaking in genetics through a case-based approach. Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education. 29(1), 392-397.

Watson, I., & Marir, F. (1994). Case-based reasoning: A review. *The Knowledge Engineering Review*, *9*(4), 327–354.

Soundararaj, G. et al. (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Unpacking Contextual Parameters Influencing the Quality of Personalized Adaptive Learning EdTech Applications

Gomathy SOUNDARARAJ[a*], Vishwas BADHE[a], Ishika ISHIKA[a], Meera PAWAR[a], Chandan DASGUPTA[a], Sahana MURTHY[a]
[a]*Indian Institute of Technology, Bombay, India*
*gomathy.soundararaj@gmail.com

Abstract: Personalized Adaptive Learning (PAL) EdTech products are specialized applications geared towards fostering self-regulated learning of individual students. The design of these applications incorporates specific features to afford such kinds of interaction. As such, the quality of the applications is judged by the presence or absence of such specific features. However, such an approach towards designing and evaluating PAL applications is unlikely to attend to contextual parameters on the ground that may influence students' perception of the quality of the application. Large scale user studies previously done using such applications typically focus on infrastructural capabilities needed to deploy the application on the ground or the learning outcomes for a large set of users because of using the application. Such studies overlook the need to understand why certain applications are well-received by students and hence may lead to significant learning gains while others are not viewed favorably. Moreover, such an approach to designing and evaluating PAL applications and lack of such user studies likely lead to the creation of contextually insensitive applications that do not appeal to the intended students. This study aims to unpack the contextual parameters that might affect a PAL application's quality when used in the classroom. We use a design-based implementation research (DBIR) approach since it enables researchers and practitioners to work closely with each other and understand the nuances of application implementation on the ground, inform generation of local theories, and work towards a sustainable change on the ground. We present six contextual parameters from Indian classrooms that are likely to influence the quality of 'PAL at school' EdTech applications. These findings advance our knowledge of both classroom learning and implementation of PAL applications, and inform the design of contextually sensitive applications.

Keywords: User studies, Personalized adaptive learning (PAL), Quality of EdTech, DBIR, context sensitive

1. Introduction

India is one of the biggest potential educational technology (EdTech) markets with a student population of 250 million. There is a growing interest from global EdTech companies to implement at scale in India and an upsurge of Indian EdTech startups catering to local needs. The EdTech space is powered by the need for quality products having a strong focal point on innovation, scalability, acceleration, value for money and a systematic change (Sikander & Rahman, 2021). Researchers have also highlighted the importance and benefits of designing educational technology that are context-sensitive (Mishra & Koehler, 2009; Zimmerman, Land, & Jung, 2016). In fact, research suggests that being sensitive to context enables students to demonstrate their strengths during the learning process (Sternberg, 2018). Thus, the presence or degree of context-sensitivity has bearing on EdTech applications' quality and knowledge of context-sensitivity in the Indian EdTech market is the need of the hour.

One of the ways of understanding what comprises an EdTech application's context is by forming insights from real classroom settings. While large scale user studies on Personalized Adaptive Learning (PAL) applications have been conducted in the past, these studies focused on measuring students' learning outcomes (Muralidharan, Singh, & Ganiman, 2019) and the infrastructural

capabilities needed to deploy EdTech applications at scale (CSF EdTech Lab 1 Report, 2019). Thus, existing user studies provide little insight into the problem of practice of what comprises an EdTech application's context and how we can inform a systemic change towards designing context-sensitive EdTech applications to fulfill local and global demand for quality applications. To address this gap, in this study, a team of researchers (authors of this paper) from a leading research institution Indian Institute of Technology Bombay - worked with Central Square Foundation, a non-governmental organization focusing on educational policy and strategy and Teach for India (TFI), a non-profit organization which runs a fellowship that brings India's brightest and most promising youths, from the nation's best universities and workplaces, to serve as full-time teachers to children from low-income communities in under-resourced schools. The problem of practice resonated with all these three stakeholders who have been working towards establishing a healthy EdTech ecosystem in India (Patel et al., 2021). This case study is part of a larger project informed by the design-based implementation research (DBIR) approach which facilitates the collaboration of research and practice in order to bring the effectiveness, sustainability, and scalability in educational interventions (Fishman et al., 2013). In this study we look at a subset of EdTech applications i.e., Personalized Adaptive Learning (PAL) deployed in Indian classrooms.

2. Background and Need

The concept of personalized learning can be traced back to Confucius's "teaching students according to their aptitude" and Socrates's elicitation teaching theory (Peng, Ma, & Spector, 2019). It started gaining importance again with the boom in student-centered approach in education. With the emergence of technology and rise of big data, the concept of adaptivity was integrated into personalized learning. Based on the four elements of individual characteristics, individual performance, personal development, and adaptive adjustment, Personalized Adaptive Learning (PAL) can be defined as a technology-empowered effective pedagogy which can adaptively adjust teaching strategies timely based on real-time monitoring of learners' differences and changes in individual characteristics, individual performance, and personal development (Peng, Ma & Spector, 2019). Thus, 'PAL at school' (or PAL, in short) EdTech applications are expected to provide personalized and adaptive instructions according to the individual learners' needs and pace (Badhe, Banerjee, & Dasgupta, 2021).

Based on prior research (Patel et. al, 2021), the EdTech Tulna index has been created as a public good that evaluates the quality of EdTech applications along 3 dimensions - Content quality, Pedagogical alignment and Technology & Design. Each dimension contains multiple evaluation criteria. Content quality measures the accuracy and content/ skill coverage through the criteria of content accuracy and clarity, alignment to national standards, and inclusivity in content representations. Pedagogical Alignment includes criteria of adaptivity, learner-centered pedagogy, enhancing learner experience, assessment of learning, and teacher support. It measures the extent of alignment of the pedagogical strategies with national educational policies, Learning Sciences theories and design principles to create a meaningful learning experience. Technology & Design measures the impact of technological affordances and the user interface design integration with the pedagogy and context and includes criteria like user interface design and affordances that facilitate learning.

3. Role of context in EdTech and inherent challenges

The learning context is defined as a depiction of the current situation of a learner related to a learning activity in addition to the attributes of the physical world model and as the information and content used to guide a specific learning activity (individual or collaborative) in a particular physical environment (Guabassi et al., 2018). The context supports and guides the learners in the learning process. Prekop and Burnet (2003) divided learning contexts into two dimensions - internal (which surrounds the user) and external (which surrounds the application).

According to Gwizdka (2000), the external dimensions consist of location, temperature, time, lighting levels and so on. To extend context-aware application into more cognitive domains, such as information retrieval, decision making, situation monitoring, product design, and so on, the cognitive domain consists of context such as users' goals, tasks, work context, business processes, personal events, communication, emotional and physical state (Gwizdka, 2000). According to Gumbheer et al.

(2022), the extrinsic context comprises Learning design (Subject, Learning Objectives, Pedagogical strategy, Learning Activities, and Learning Resources), Device Wearable & Handheld, and Hardware and Software Resources. Configuration and Physical Properties and Surrounding (nearby Resources, Learning Partners, Activities). Intrinsic context comprises contextual information that is related to the learner's profile including the knowledge level, concentration, motivation level, learning style, cognitive load, and learning behavior.

These extrinsic and intrinsic factors influence the context of use of EdTech products. The context of use is crucial, especially in developing nations like India which comprise diverse contexts, as it talks about how the EdTech applications' quality is perceived by the students. According to Datta (2019) in "one size does not fit all" article, the challenge in implementing EdTech applications in Bottom of the Pyramid (BoP) context is that it becomes a time consuming task since these applications are not designed with multiple contexts in mind. Apart from the practical challenges such as device and internet availability, the article suggests that there needs to be customization according to the demographic and context so it boosts learners' engagement and experience with the application.

Taking into account the existing categories of quality and the external and internal aspects of context mentioned above, the paper focuses on these specific dimensions of the learning context to inform and comment on the context-sensitiveness of PAL EdTech applications. The research question explored in the study is: What are the contextual parameters that influence the quality of PAL Ed-Tech applications?

3. Method

The user-study process started with taking prior consent from Grades 6-8 students, parents and TFI teachers to ensure the willingness of all the stakeholders. The teacher and student demographic survey along with the student diagnostic tests were taken. After collecting the survey data, each classroom was allocated one of two the PAL EdTech applications - PAL 1 and PAL 2. We proceeded with the two PAL applications for our user study in Grades 6-8 since they had national and international presence and had been found to be well-designed for classroom use. The Learning Unit's (LUs) that students needed to work on were chosen by researchers in consultation with the teachers based on the topics covered in the respective classrooms. The teachers were given access to the application and trained on the usage of the application and time to test the application before they implement it with their students. Training included familiarizing them with the application's features and introducing them to the survey tools to be answered by students after using the application. Researchers and the teachers jointly ensured the hardware and software availability before the user studies. Each grade-wise user study took 3-5 days to complete as the students were expected to spend a minimum of 5 hours with the application after school. The teachers and researchers observed the students use the PAL application and supported them wherever needed. The Learning Object Evaluation Survey-Students (LOESS) and Learning Object Evaluation Survey-Teachers (LOEST) was conducted on the last day of the user study (Kay, Robin, Knaack, & Liesel, 2009). The broad process followed in the user study is represented in Figure 1:

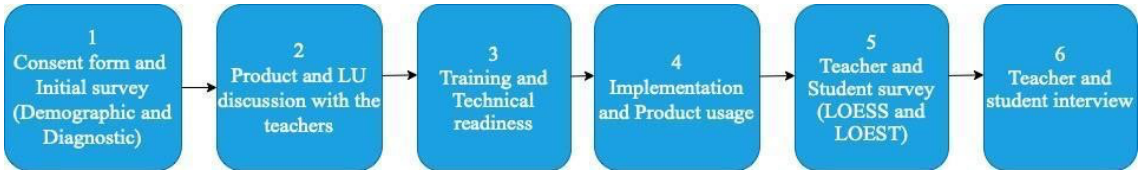


Figure 1: Depicts the broad process of user studies

3.1 Participants

The user study was conducted in four different government schools from Hyderabad and Mumbai with TFI fellows. Some reasons for choosing these classrooms are the homogeneity in the sample group, resource availability since researchers were based in these cities, students with good English language proficiency and TFI fellow volunteers.

Based on both demographic and diagnostics survey data, we followed purposive sampling (Patton, 1990) to ensure that the students who participated in the study had the following characteristics - good English language proficiency level, access to devices, technical skills and knowledge and no previous use of allocated applications. With reference to English language proficiency levels, for PAL 1, 72.97 % (27) of students were at paragraph level in reading, and 27.02% (10) at sentence level. For PAL 2, 84.21% (16) of students were at paragraph level in reading, and 15.78% (3) at sentence level. Some of the educational apps that students had used earlier were Read Along, Bolo, Byju's, and Class edge. The students who were already familiar with one of the PAL applications were assigned the other one to avoid bias due to prior experience with the application.

The compatibility test of the PAL EdTech applications showed that the applications were more than 95% compatible across smartphones and laptops and that the user experience was similar across these devices. For LU allocation, the list of topics covered in the classroom was matched with the units in the respective application according to the grade. The approximate time needed for one LU completion was identified by choosing a sample unit which helped in assigning the number of LUs. The approximate time required to complete one LU was 1 hour 50 minutes for PAL 1 and 1 hour 10 minutes for PAL 2.

Table 1 summarizes the information on grades, city, LUs allocation and number of participants (total number of students who used the application, total number of males and females, and number of students interviewed). Note that the number of students interviewed depended on their availability and hence the numbers varied across the locations.

Table 1: *The details of grade-wise LU allocation*

Grade	PAL Edtech Application	City	Learning Units				Student count (N)
			1	2	3	4	
6	PAL 2	Hyderabad	Integers	Fractions	Mensuration	–	8 (1-M,7-F) Interviewed-4-F
7	PAL 1	Hyderabad	Integers	Operations on Fractions	Lines and Angles	Area and Perimeter	20 (11-M, 9-F,) Interviewed-1-M
8	PAL 2	Hyderabad	Square and Square roots	Exponents and powers	Mensuration	–	11(11-F) Interviewed-4-F
	PAL 1	Mumbai	Real numbers	Exponents and roots	Algebraic expressions	Operations on Fractions	17(6-M, 11-F) Interview-10 (4-M,6-F)

3.2 Data collection and Analysis

The data was collected through surveys, teacher, and student interviews as well as process data from the applications. The demographic survey provided information on the context and background of the students. The diagnostic survey showed the students' reading and conceptual understanding level of mathematics. The teacher demographic survey provided information about usage of technology in the classroom and their perspectives on integration of technology. The LOESS and LOEST survey provided data on three key constructs: Learning, Quality and Engagement (Kay, Robin, Knaack, & Liesel, 2009). Through the student and teacher interviews, additional perceptions on the three constructs and general responses on the PAL application were collected. The PAL applications provided data through teacher dashboards and reports. While the surveys collected both quantitative and qualitative data, the process data was focused on the former and the interviews on the latter. The names of the students and teachers mentioned in this study have been anonymized.

We followed the thematic analysis approach (Braun & Clarke, 2006) to analyze the student and teacher interview data. During the analysis, we sought to identify parameters recommended by literature and evident in the data that likely defined the context of use of the PAL applications in our classroom settings.

4. Findings

This section outlines the contextual parameters derived from the evidence and inferences gathered from qualitative as well as quantitative data from students, teachers, and process data.

4.1 Contextual parameter 1: Correctness and clarity in assessment

Grade 8 students at government school from Mumbai were provided with a PAL 1 application with a set of learning units as given in Table 1. We observed that students faced difficulty in understanding the assessment even though the questions and their solutions were factually correct with accurate scientific explanations. The wording of the assessment questions conveyed the intended meaning and expectations from the students. Yet students sought clarification from peers and teachers on the questions, answers, and explanations. One of the teachers (Ananya) corroborated this observation when she mentioned that there was confusion with respect to the method and explanation provided for solving the questions.

> Ananya (PAL 1 Teacher): "*...not exactly a misconception but just confusion with respect to the methods on reaching a solution because the app used different sorts of methods and different questions"*

Students also faced difficulty in understanding the assessment questions which were not covered in the classroom. Sonali's account below illustrates this point.

> Sonali (student): *"I did not understand the question because I have not done this type before."*

Commenting on the quality of EdTech application: The assessment questions in the PAL 1 application were correct, clear, and unambiguous. However, in the user studies we observed that students were struggling to comprehend the questions. Students' prior knowledge influenced the clarity of the assessment questions. Further, PAL 1's design promoted a one-right approach to solving the given problems which likely led to confusion when it deviated from the teacher's approach.

4.2 Contextual parameter 2: Language comprehensibility

Language comprehensibility refers to the language used in the PAL application content. The language used should be comprehensible by the target group of students. This emerged as an important consideration for both PAL 1 and PAL 2 applications used by students of grade 8 in the government schools in Mumbai and Hyderabad with a set of learning units as given in Table 1. From the students' interview responses, we gathered that students were facing difficulty in understanding the language used in both the PAL applications. For instance, the teacher (Ananya) mentioned that in PAL 1, "*...not everyone could understand especially word problems."* Pranjali (student) corroborated this in the case of PAL 2 application when she highlighted that the speech pace and accent was problematic.

> Pranjali (student): "*They talk very fast, American English, so I have to repeat 3 times to understand."*

Furthermore, students (e.g., Shreya) mentioned that the writing style of PAL 2 was difficult to understand in the content videos.

> Shreya (student): "*English was very fast. handwriting is poor, like we are stuck sometimes like when it is 'l' written and it looks like 's' or 'I'..."*

Commenting on the quality of EdTech application: The above instances highlight the importance of the language used in PAL applications. Unfamiliar pace and accent of the verbal content hindered students' progress. Handwritten texts that are usually preferred due to their personalized and authentic feel were also problematic owing to the poor quality. These findings were also supported by the LOESS survey data, in which 13 students out of 19 have reported obscurity in handwriting and accent.

4.3 Contextual parameter 3: Motivational features

Motivational features such as virtual agents and rewards prompt the learners to engage with the content. We found that the presence of such features was highly motivating for the students. For instance, grade 8 students from the government school in Mumbai were using PAL 1 for a set of learning units as given in Table 1. Students experienced an increase in motivation due to the sparkies, coins, rewards, and points in PAL 1 application.

> *Drishti (student): "I like the sparkies that I get and also, I see the friends' sparkies too. When I feel that I have less sparkies than friends then I feel that I have to solve more questions to get more sparkies. I liked that whenever one topic was completed, the game was enabled. So, for games, I asked more questions."*

Similarly, another student mentioned about the motivating effect of the rewards, points, etc.

> *Vinit (student): "To get more coins, rewards, points, I want to do more questions and whatever I have gained that means more coins, more better rank. Those who have more coin rank are at the top."*

These observations are corroborated from LOESS data for PAL 1 and PAL 2 - 54.05% (20) and 47.36% (9), respectively, strongly agreed that the apps motivated them. The LOEST survey results for PAL 1 and PAL 2 show that three teachers confirmed that the features motivated the students while using the app.

Commenting on the quality of EdTech application: Both PAL 1 and PAL 2 applications included a comprehensive set of motivational features (e.g., Sparkies, leaderboards, and badges) that were well-integrated and prompted the learners to interact further with the content. These features had similarities with gamified elements and indicate the relevance of considering gamified elements when designing quality PALs.

4.4 Contextual parameter 4: Teacher support

Support should be provided to teachers to effectively integrate the LUs in their classroom teaching and to enhance the learning experience. More teacher support with deeper insights on the teacher dashboard about the current student's performance state is required. The support should allow teachers to make choices and customize to a limited extent. One of the teachers from the Hyderabad government school commented in this regard while observing students of grade 7 using PAL 1 application.

> Devashish (teacher): "*We need detailed feedback of what the student is doing presently. There is a teacher dashboard and we are getting all the detailed information for this example this type of conceptual question, they are not able to do it. For example, if I'm getting the information today, that at night they were using it, and this is the question they were struggling with next, I will be preparing the lesson plan like that. In the teacher dashboard, the following data is given: Percentage of progress, Attempts, Number of questions done, Accuracy, Time taken for each question and over all app usage, Number of right and wrong question"*

From LOEST data, two teachers reported that they wanted a more detailed teacher dashboard that gives real time information of the students' progress. They expressed that they should be able to know the practice hours, practice content and efficiency of the students.

Commenting on the quality of EdTech application: Adequate teacher support has been provided including customization and assigning tasks or relevant materials to students. However, in PAL 1, the real time feedback or support while students are using the application is needed to know the real-time status of students and to make the teaching learning process effective.

4.5 Contextual parameter 5: Task value

Task value refers to the perceived value of completing a task by the student. To encourage students to engage with a task and work on it persistently to achieve their learning goals the task should be

meaningful for the students (National Academies of Sciences, Engineering, and Medicine, 2018). We found that once students realized that after attempting a particular number of questions, they will get a game to play, they began guessing answers for questions. Sometimes their guesses were correct and got them through more questions in less time. When their guess was incorrect and the application provided corrective feedback, students ignored the feedback just to attempt more questions in less time. Thus, students were having low task value for completing the questions (task) properly. These insights were collected from students in both Mumbai and Hyderabad for both PAL 1 and PAL 2. The instances below illustrate this point. Students were initially valuing the process of attempting the questions diligently i.e., high-task value.

Rahul (student, PAL 1): "After trying the same question I was thinking what can I do to make it correct."

Swadha (student, PAL 2): "I did not understand the video the first time, so I listened to it again and again, three times."

When students realized that they could enable games after answering a set of questions, they started guessing answers and spent approximately two seconds on a problem. From the teacher dashboard available in PAL 1**,** it was observed that most of the students spent very little time on most of the questions. For example, one student spent 329 minutes in the app and answered 344 questions. The detailed data shows that the student took about 2-3 seconds per question. The interviews also revealed students' focus on solving more questions to play the games.

Vinit (student, PAL 1): "Games are enabled when I do a certain topic when I see and complete a topic. Some games that create fun for us and motivate me to solve more questions. What I like most in the app is the game feature and also examples given in it help me to understand more."

Drishti (student, PAL 1): "I liked that whenever one topic was completed, the game was enabled. So for games, I did more questions."

Commenting on the quality of EdTech application: The presence of games within PAL applications motivates students to engage with more questions. However, this engagement is beneficial only when students find the task meaningful and ascribe high value to completing it diligently. To get to the game, students resorted to a guess and check approach and the PAL application was unable to counter such an approach.

4.6 Contextual parameter 6: Social Learning

PAL applications are considered to be used by individual students. As such, PAL applications are designed for individual work that doesn't support collaborating with one another. However, a typical classroom setting provides opportunities for students to discuss and collaborate with each other. Evidence taken from interviews of grade 8 students from Mumbai and teachers from Hyderabad give insights on social learning while using the application. For instance, a teacher (Devashish) mentioned the following.

"*The first one is, for example, when the students are using in the class, I think, the main advantage is that there, the application has that inbuilt thing that they will find curiosity and all but here I am seeing that, one student is doing it. The second one is more exciting (by looking at it). Okay. He's also doing and he's also using it since the last half an hour. I will also do it, there is something and there is one thing that you see that the group mentality starts working, right, yeah... they start discussing the doubts, they start solving it. So, it's a, I think a collaborative effort starts to kick in when they're using this type of application in the school environment... I think under the school context, in the school environment, it's much more productive than what can be provided at homes, because at homes that I don't think there is such a conducive (collaborative) environment for them to study."*

A student also commented on the importance of social interaction with peers.

Khushboo (student, PAL 1): "At last I did not understand the question, then I saw my friend doing that question then I understood. "

Commenting on the quality of EdTech application: Learning from peers via discussion and collaboration are often not given a consideration for PAL applications. The design of such applications usually does not support social learning. Our findings show that when PAL applications are used in classrooms, the teacher and peers serve as critical resources for facilitating learning. However, this view of learning using PAL applications that gives importance to the social interactions is at tension with the common notion of PAL being used by individual students to enable it to adapt to the students' needs. Thus, ensuring a balance between social learning and adaptivity in PAL applications becomes important.

These six contextual parameters are related and in sync with the context dimensions specified by Gwizdka (2000) and Gumbheer et.al (2022) as follows - a) parameter 1 is linked to the extrinsic context of learning activities and subject, b) parameter 2 addresses the communication aspect of the learning context, c) parameter 3 is directly related to the intrinsic context of motivation level, d) parameter 4 is aligned to device aspect, more specifically to the software resources, e) parameter 5 is derived from cognitive domains of tasks and goals, and f) parameter 6 is linked to the surrounding, specifically to the learning partners.

5. Discussion and Conclusion

In this research, we set out to unpack the contextual parameters that might affect a PAL application's quality when used in the classroom. We use a DBIR approach that emphasizes transcending research/practice boundaries and integrates wisdom of practice into innovations in a way that informs ongoing design (Penuel & Potvin, 2021). With this approach in mind, the user-study is an outcome of the collaboration between researchers, evaluators, teachers and students where all the stakeholders brought multiple perspectives about the quality of PAL applications.

We found six contextual parameters that potentially affect the quality of PAL EdTech applications when used in real classroom settings. In the first contextual parameter the students and teachers highlighted the confusion due to the difference between the method of teaching in the classroom and the method followed to solve the questions in the applications. The novelty of the type of questions for users who did not know the concept before led to lack of clarity of the assessment questions. Second, we found that the comprehensibility of the content was affected by the accent used and quality of handwritten text. Foreign accents and unfamiliar handwriting slowed down students' progress. These two contextual parameters are tied to the situation of a student and information and content used in specific learning activities (Guabassi et al., 2018). Participants in this study had sufficient mathematics knowledge and reading comprehension level for engaging with the PAL application productively. However, their interaction with the application was affected by the lack of adequate consideration of the context of use in its design. Third, motivational features (e.g., Sparkies, leaderboards, and badges) played an important role in prompting the learners to interact further with the content. Thus, such gamified elements are important in facilitating effective use of the PAL applications in the school setting. However, when students are too much focused on interacting with these elements, there arises the risk of encouraging a guess-and-check approach where students pay minimal attention to the content (fifth parameter). Thus, there is a need to balance the focus on intrinsic context and cognitive domains in PAL applications (Gwizdka, 2000; Gumbheer et al., 2022). The contextual parameter of teacher support notes that the features present in the app don't offer sufficient data and information on student's learning and progress for the teachers to customize their teaching. Finally, we found that in the classroom context, students were learning from each other's mistakes, peer interactions, seeking teacher's support in understanding the questions and then attempting questions correctly. Importance of social interactions while using PAL applications is often overlooked and not supported via the application's design.

These contextual parameters highlight ways in which PAL EdTech applications can be designed to be sensitive to the local context and support teaching and learning effectively. Presence of TFI as an implementation stakeholder was important for arriving at these findings because the TFI fellows brought a unique perspective that comes from having worked with children from low-income communities in under-resourced schools in India. The DBIR approach helped facilitate the partnership with TFI fellows for this user study and impressed upon the significance of attending to context-

sensitivity while designing PAL applications. The contextual parameters are independent in nature, however there can be an indirect dependency which needs to be explored further.

One of the limitations of the study is that the six parameters listed above are not an exhaustive set in this study and more contextual parameters might evolve from further user studies conducted at scale with different methods, students, teachers, and PAL applications. The findings from this study are not generalizable due to the small sample size and lack of longitudinal data from students and teachers across a complete school year. Another limitation would be that the internet speed and technical errors that occurred during the user study may have affected the student experience of the application. These technical challenges may have compounded the issues already faced by the students due to the parameters listed above.

On the whole, the knowledge of the contextual parameters through the user studies can be leveraged in future research to design contextually sensitive EdTech applications. These findings also advance our knowledge of both classroom learning and implementation of PAL applications in classrooms. The contextual parameters identified from this study can inform the design of large-scale user studies which acknowledge the role of context, culture, and students' background in determining the quality of EdTech applications.

Acknowledgement

The study was conducted as a part of the EdTech Tulna project hosted by the Department of Educational Technology, Indian Institute of Technology, Bombay in partnership with Central Square Foundation (CSF). The authors acknowledge the support of Central Square Foundation for funding and partnering in the EdTech Tulna project. The authors are grateful to the Tulna team members for their comments and suggestions, and to the Educational Technology interdisciplinary program at IIT Bombay for overall support. Teach For India (TFI) facilitated the implementation of the study with the schools GHS NBT Nagar, GHS Mehbubia, GHS Durgabhavani in Hyderabad and Siddharth Nagar UPS in Mumbai. The teachers at the schools and the students were enthusiastic participants of the study.

References

Badhe, V., Banerjee, G., & Dasgupta, C. (2021). Design Guidelines for Scaffolding Self-Regulation in Personalized Adaptive Learning (PAL) systems: A Systematic Review. In Rodrigo, M. M. T. et al. (Eds.), Proceedings of the ICCE 2021 (pp. 372-380). Bangkok, Thailand: APSCE

CSF EdTech Lab 1 Report (2019). Retrieved on 1 June, 2022 from https://www.idinsight.org/wp-content/uploads/2021/09/EdTech-Lab-Report_November-2019.pdf

Datta, S. (2019, October). One size does not fit all. Impact, Inspirations & Insights for Social Innovation, 30. Retrieved August 7, 2022, from http://assistasia.org/impact_magazine/Ed30_OctDec2019.pdf

EdTech Tulna, IIT Bombay (2021). TR-ET-Tulna-PAL-2021-01. Retrieved on May 31, 2022 from https://edtechtulna.org/?jet_download=5956

EdTech Tulna Evaluation Report For PAL 1 and PAL 2 for grade 6, 7 and 8 (2021). Retrieved on 31 May, 2022 from link 1 and link 2.

Ferreyra, M. M., & Liang, P. J. (2012). Information asymmetry and equilibrium monitoring in education. Journal of Public Economics, 96 (1-2), 237-254.

Fishman, B.J., Penuel, W.R., Allen, A.-R., Cheng, B.H., Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. In B. J. Fishman, W. R. Penuel, A.-R. Allen, & B. H. Cheng (Eds.), Design-based implementation research. National Society for the Study of Education Yearbook, 112(2), 136–156.

Gumbheer, C. P., Khedo, K. K., & Bungaleea, A. (2022). Personalized and Adaptive Context-Aware Mobile Learning: Review, challenges and future directions. Education and Information Technologies, 1-27.

Guabassi, El. I., Bousalem, Z., Al Achhab, M., Jellouli, I., & Mohajir, B. E. E. (2018). Personalized adaptive content system for context-aware ubiquitous learning. Procedia Computer Science, 127, 444-453.

Gwizdka, J. (2000, April). What's in the context? In Computer Human Interaction (Vol. 2000, pp. 1-4).

Johnson, A. M., Jacovina, M. E., Russell, D. E., & Soto, C. M. (2016). Challenges and solutions when using technologies in the classroom. In S. A. Crossley & D. S. McNamara (Eds.) Adaptive educational technologies for literacy instruction (pp. 13-29). New York: Taylor & Francis.

National Academies of Sciences, Engineering, and Medicine. (2018). How people learn II: Learners, contexts, and cultures. National Academies Press.

Kay, Robin, Knaack, & Liesel. (2009). Assessing Learning, Quality and Engagement in Learning Objects: The Learning Object Evaluation Scale for Students (LOES-S). Educational Technology Research and Development. 57. 147-168. 10.1007/s11423-008-9094-5.

Kay, Robin, Knaack, & Liesel (2007). Teacher evaluation of learning objects in middle and secondary school classrooms. Retrieved November 1, 2007 from http://faculty.uoit.ca/kay/papers/LOES_Teacher_2007.doc.

Braun, V. & Clarke, V. (2006). Using thematic analysis in psychology, Qualitative Research in Psychology, 3:2, 77-101, DOI: 10.1191/1478088706qp063oa

Means, B., & Penuel, W. R. (2005). Research to support scaling up technology-based innovations. In C Dede, J. Honan, & L. Peters (Eds.), Scaling up success: Lessons from technology-based educational improvement (pp. 176–197). New York, NY: Jossey-Bass.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. American Economic Review, 109(4), 1426-60.

Mishra, P., Koehler, M. J., & Kereluik, K. (2009). Looking back to the future of educational technology. TechTrends, 53(5), 49.

National Academies of Sciences, Engineering, and Medicine. (2018). *How people learn II: Learners, contexts, and cultures*. National Academies Press.

National Education Policy (2020). https://www.mhrd.gov.in/sites/upload_files/mhrd/files/nep/NEP_Final_English.pdf

Patel, A., Dasgupta, C., Murthy, S., & Dhanani, R. (2021). Co-Designing for a Healthy Edtech Ecosystem: Lessons from the Tulna Research-Practice Partnership in India. In Rodrigo, M. M. T. et al. (Eds.), Proceedings of the ICCE 2021 (pp. 589-598). Bangkok, Thailand: APSCE

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: Sage Publications.

Peng, H., Ma, S., & Spector, J. M. (2019). Personalized adaptive learning: an emerging pedagogical approach enabled by a smart learning environment. Smart Learning Environments, 6(1), 1-14. https://doi.org/10.1186/s40561-019-0089-y

Penuel, W. R., & Potvin, A. S. (2021). Design-based implementation research to support inquiry learning. International Handbook of Inquiry and Learning, 74-87.

Prekop, P., & Burnett, M. (2003). Activities, context and ubiquitous computing. Computer communications, 26(11), 1168-1176.

Sikandar, M. A., & Rahman, P. F. (2021). Edtech Start-ups in the education ecosystem in the post-Covid-19 era in India. Towards Excellence: Journal of Higher Education, UGC-HRDC, Gujarat University, India.

Sternberg, R. J. (2018). Context-sensitive cognitive and educational testing. Educational Psychology Review, 30(3), 857-884.

Zimmerman, H. T., Land, S. M., & Jung, Y. J. (2016). Using augmented reality to support children's situational interest and science learning during context-sensitive informal mobile learning. In Mobile, ubiquitous, and pervasive learning (pp. 101-119). Springer, Cham.

Liu, A. H. *et al.* (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Instructional Design on Data Visualization Model of Using AR Sandbox Apps in Learning and Teaching Geography

Aaron Hok-him LIU[a*] Percy Lai-yin KWOK[b]
[a] *Dept of Social Sciences, The Education University of Hong Kong, China*
[b] *Dept of Education Policy & Leadership, The Education University of Hong Kong, China*
*s1138614@s.eduhk.hk

Abstract: Visualization of spatial data powerfully enhances our understanding of Geography in which map reading is one of the difficult skills for secondary teachers to teach. The AR sandbox provides a new tangible way to overcome this learning obstacle. Since the COVID-19 suspension has restricted the usage of authentic teaching methods since 2020, virtualizing the tangible sandbox into the mobile AR Sandbox app is necessary to continue learning and teaching Geography. Despite the technological breakthrough, a new pedagogy is required for adopting the technologies into effective teaching. Past research bodies found that the best use of the AR sandbox requires good learning instructions. The researchers develops the new data visualization model to convert spatial learning into the new AR/VR learning environment. This paper aims to outline an instructional design of the AR sandbox and conjecture a conceptual IPAD framework of the AR sandbox in some online fieldtrips with a seamless approach in teaching some geospatial concepts of Geography in Hong Kong's K-12 classrooms. Finally, some implications will be drawn for further research.

Keywords: augmented reality, AR sandbox, data visualization, mobile AR, tangible AR

1. Introduction

The technologies of mixed reality have become ubiquitous in classroom teaching. The AR sandbox is one of the best examples of applying AR technologies to the innovative Geography learning environment. The AR Sandbox has been utilized through augmented contour lines projection on the authentic sand. The projection in the tangible sandbox changes with "hands-on" movement. Students need to learn the 3D landscape by touching it to visualize some sophisticated concepts and their interrelationships in learning Geography. However, due to school suspension of face-to-face classes, the strict COVID-19 prevention instructions place great obstacles to the physical contact between students. Therefore, the action researcher (the first author) and the university academic (the second author) have converted the display and functions of the tangible AR sandbox into a mobile AR app. The mobile AR function provides the same topographic display in a virtual mobile AR view. Students can keep their "hands-on" experiences in the mobile app-based virtual environment. The AR sandbox app acts as a seamless learning tool which provides a high-mobility AR display in Geography.

2. Literature Review

2.1 Past Research Directions in Sandbox

The AR sandbox is a hands-on tool (Vaughan *et al*., 2017) that simulates Kolb's (1984) "experimental learning" in higher education in map reading (Healey & Jenkins, 2000). Majgaard & Weitze (2020) inspire our approach to experiential learning in a visual reality environment. The augmented image projections can be reproduced many times, allowing repeated learning cycles and changes in difficulties through the simulation image. The "experiential learning model" applies to students learning through the AR sandbox.

A comparison between pre-post testing experiments showed that students' test scores improved after using the AR sandbox (Carrera & Bermejo Asensio, 2017). Students learn more in the AR sandbox (Woods *et al.*, 2016). However, some researchers have argued that an AR sandbox only stimulates interest in learning and does not learn through the AR sandbox (Giorgis *et al.,* 2017). Jackson *et al*. (2019) proved that the effect of learning depends on pedagogy and time. Woods *et al.* (2016) also found that trainers need to adjust their classes to use multiple AR sandboxes during the semester to perform well. A multi-institutional study of inquiry-based lab activities using the AR Sandbox has recently

indicated the benefits of using the AR sandbox without any learning instructions are not significant (McNeal *et al.*, 2019).

In Geography, data visualization (DV) is crucial. The map provides data visualization for our spatial thinking (Ellard, 2015). Understanding maps is an important method of learning Geography (Wellsted, 2006). In reality, map reading is not easy to teach (Jennings, 2011). Sánchez *et al.* (2016) suggested that the AR sandbox could be used for teaching and learning in secondary schools. Students can "experiment by changing the shape of the sand with their hands and experimenting with the changes in the topographic map." The use of 3D modellings does improve map-reading skills in K-12 Geography (Boardman, 1996). Digital maps may serve that purpose. However, most digital maps are web-based and viewed on a screen or tablet. The AR sandbox provides an authentic experience with sufficient modelling possibilities for teachers in normal lessons (Cárdenas-Delgado *et al.*, 2020).

This study in the paper focuses on pedagogical development and transformation of the 2015 version of AR Sandbox. It is designed to bridge the gap between 2D maps and 3D spatial concepts (Fisher *et al.*, 2019) and improve learning efficiency (c.f. Offermo, 2016).

2.2 Historical Development of AR in Hong Kong

The AR sandbox was introduced into Hong Kong by the first author in 2015 (reported on Singtao Daily, 2015.11.04). The AR sandbox has been well-promoted and widely used in secondary (and even primary) schools in Hong Kong (Lai, 2019). According to a local AR sandbox research, students' test scores had a 15% increase, and around 60% of students increased their length in answering after pre-test and post-test comparisons (Ku, 2017). Hence, the Education University of Hong Kong (EdUHK) conducted a pilot study and assisted the pilot school in Hong Kong in installing this teaching tool and a student at EdUHK carried out quantitative research in Hong Kong to study the usage of AR sandbox in schools (Hu, 2019). However, no local empirical research on pedagogical uses of AR sandbox has been conducted so far in Hong Kong. This motivates the first author of this paper to devise the most effective learning and teaching environment in Hong Kong (i.e. K-12 students in large classroom settings) through the AR sandbox. Owing to the school suspension under the pandemic attack, the first author has converted the AR sandbox's functions into an app-based environment to record learning and teaching progress through the functions of the AR sandbox.

3. Research Design

3.1 Emergence of the IPAD Framework in Learning and Teaching Geography

The conversion of the AR sandbox into an app-based environment under the emergence of the **IPAD** (Liu, 2016) framework is depicted in four successive phases in table 1.

Table 1. *Four Phases of the **IPAD** framework*

Phase 1	From reality to the tangible AR sandbox - **I**nformation gathering **(I)**
Phase 2	From tangible AR sandbox into the VR app environment - **P**rocess for visual enhancement (**P**)
Phase 3	VR app into mobile AR sandbox application - **A**pplying into seamless learning environment (**A**)
Phase 4	Bringing the enhancement learning environment into reality - **D**ata enhancement mobile learning (**D**)

This **IPAD** framework provides an innovative idea to integrate information for the researchers and users to develop and use the mobile AR sandbox app. during and beyond Geography lessons. Teachers and students may use the app to generate a virtual environment with the aid of AR images. The **IPAD** framework also helps consolidate the instructional design and draw pedagogical and learning reflections.

3.2 Instructional Design

The first author (the action researcher) uses a landslide to go through the **IPAD** framework in the case study. In the **IPAD** framework in figure 1, students are guided individually or in group settings to explore from a tangible sandbox in the real world to VR app environment. Then from that environment to the mobile AR sandbox app, the co-existence of AR/VR and real-world (R) helps students build / consolidate / revise some abstract or sophisticated geophysical concepts through data visualization.

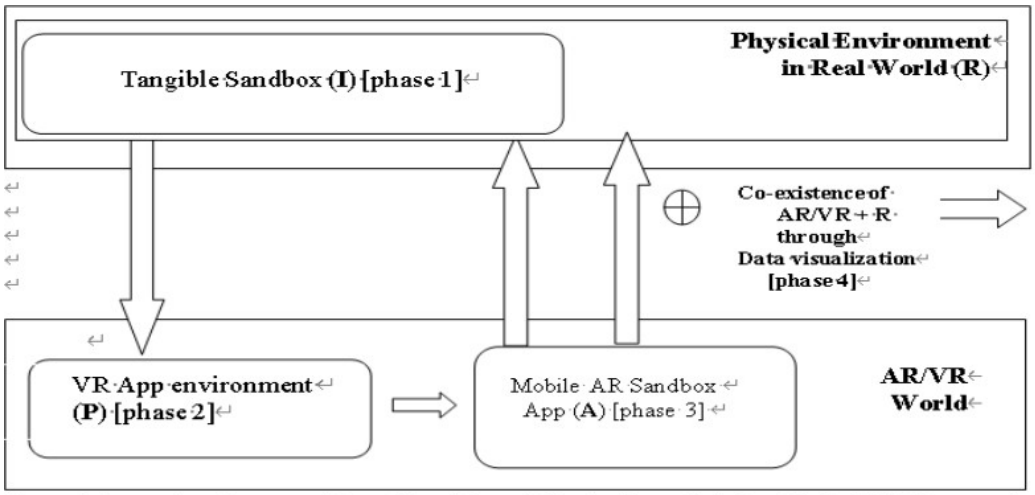


Figure 1. Interactions Between AR world and the real (physical) world depicted in the IPAD framework

3.3. Background of the AR Fieldtrip

This virtual field study is designed according to the three themes below:

1. **I**nformation: Exploring the Landscapes:
   Studying the area by the contour enhanced virtual 3D landscape.
2. **P**rocess and **A**pplication: Understanding the Risks
   Using AR technologies to add rainfall and reproduce the simulation of this event.
3. **D**ata enhancement: Managing the Hazards:
   Examining the usefulness of the check dam in the simulation.

The **IPAD** framework provides a way to organize the spatial data with AR/VR technologies. The literature review found that the AR sandbox lacks a pedagogical framework which is essential for being a better learning and teaching tool in Geography and Geoscience education. The **IPAD** framework requires teachers and students to gather the information (**I**), to process (**P**) the data and to apply (**A**) in real cases. Then, through the AR/VR technologies, the data enhancement (**D**) further enriches the visualization. It brings the map reading and virtual fieldtrip into high levels of conceptual understanding. It is not just a technological breakthrough but also a pedagogical renewal. Generally, AR visualization usually focuses on virtualizing the real world into the virtual world, which makes the virtual world real. But this study bridges the virtual world back to the real world. The virtual landscape displayed in the AR landscape creates realistic visualization for the users in figure 2.

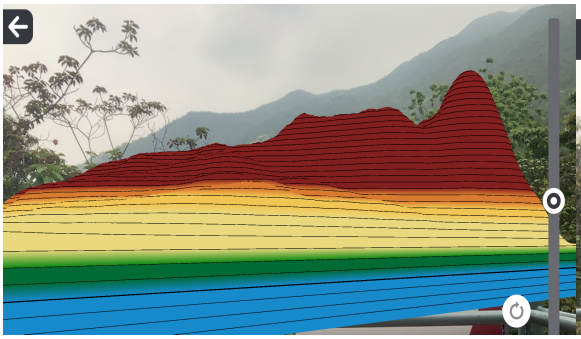


Figure 2. The AR Sandbox virtual terrain imposed in the AR background

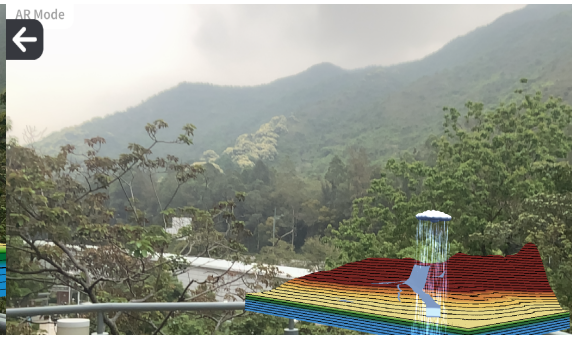


Figure 3. The rainfall creates visualized enhancement in students' learning

Furthermore, rainfall mode provides the chance for students to experience a real-life situation, i.e. landslide after heavy rainfall in figure 3. Learning becomes more dynamic through the AR rainfall function. An AR landslide happens in the mobile AR environment.

With this technological and pedagogical breakthrough, Geography teachers may explore more of the AR fieldtrips in the future. Students may use the virtual 3D space and impose the information into real-life learning. Furthermore, any images with predicted changes may also use this technique to enhance the visualization of their virtual prediction.

4. Conclusions and Implications

The pedagogical usage of tangible AR sandbox, mobile AR Sandbox app and **IPAD** framework are still under-researched in local and international contexts. More data will be collected to reshape learning Geography and diversify the uses of the AR/VR technologies in learning and teaching Geography through the **IPAD** framework. If 3D maps or 3D mappings are bing assessed in upcoming open local and international examinations, there will be an assessment reform. Indeed (mobile) AR sandbox has an enormous potential to help improve student learning, reflected from the above preliminary action research results.

Further research agendas include:

- Possible conceptual changes in student learning Geography aided by the mobile AR Sandbox app
- Contextualized facilitators for and barriers against the **IPAD** framework
- Cognitive models accounting for how students' individualized connections between AR/VR and real-world (R) in the **IPAD** framework deepen their geospatial learning and scaffolding supports for peer co-construction of new concepts or correction of misconceptions

References

Boardman, D (1996). Learning with ordnance survey maps. In Bailey, P., & Fox, P. (Eds.), *Geography teachers' handbook*. Sheffield: The Geographical Association.

Cárdenas-Delgado, S., Padilla Almeida, O., Loachamín-Valencia, M., Rivera, H., & Escobar, L. (2020, July). YaniWawa: An innovative tool for teaching using programmable models over augmented reality Sandbox. In *International Conference on Systems and Information Sciences* (pp. 378-388). Springer, Cham.

Carrera, C. C., & Bermejo Asensio, L. A. (2017). Landscape interpretation with augmented reality and maps to improve spatial orientation skill. *Journal of Geography in Higher Education*, *41*(1), 119-133.

Ellard, C. (2015). *Places of the heart: The psychogeography of everyday life*. Bellevue literary press.

Fisher, R, Heckbert, S., Maria, J., & Sutton, S. (2019). Augmenting physical 3D models with projected information to support environmental knowledge exchange. *Applied Geography*, *112*.

Giorgis, S., Mahlen, N. & Anne, K. (2017). Instructor-led approach to integrating an augmented reality sandbox into a large-enrollment introductory geoscience course for nonmajors produces no gains. *Journal of Geoscience Education*, *65*, 283–291.

Healey, M., & Jenkins, A. (2000). Kolb's experiential learning theory and its application in Geography in higher education. *Journal of Geography*, *99*(5), 185-195.

Hu, X. J. (2019). *Acceptance of AR sandbox: applicability of the unified theory of Acceptance and use of technology (UTAUT) to Hong Kong Geography teachers* (Dissertation). The Education University of Hong Kong.

Jackson, D., Kaveh, H., Victoria, J., Walker, A., & Bursztyn, N. (2019). Integrating an augmented reality sandbox challenge activity into a large-enrollment introductory geoscience lab for nonmajors produces no learning gains. *Journal of Geoscience Education*, *67*(3), 237-248.

Jennings, K. (2011). *Maphead: Charting the wide. weird world of Geography wonks*. Scribner's, New York, NY.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development* (Vol. 1). Englewood Cliffs, NJ: Prentice-Hall.

Ku, K. M. (2017). *Exploring the potential of augmented reality sandbox to effective strategies for pedagogical practice in secondary school Geography: A Hong Kong Study* (Dissertation), University of London.

Lai, C. H. (2019). *ARGEO sandbox educator guide*. Hong Kong: Classroom Centre.

Liu, H. H. A. (2016). The application of e-learning platform in Geography map reading: Practice and expectation（應用電子學習平台在地理科教導地圖閱讀：實踐及期望）[in Chinese]. In *GCCCE 2016 teachers' forum*. EDUHK, Hong Kong.

Majgaard, G., & Weitze, C. (2020). Virtual experiential learning, learning design and interaction in extended reality simulations. In *14th European Conference on Games Based Learning, ECGBL 2020* (pp. 372-379). Academic Conferences and Publishing International.

McNeal, K., Ryker, K., Whitmeyer, S., Giorgis, S., Atkins, R., LaDue, N., Clark, C., Stoltis, N., and Pingel, T. (2019). A multi-institutional study of inquiry-based lab activities using the augmented reality sandbox: impacts on undergraduate student learning. *Journal of Geography in Higher Education*, *44*(1), 85-107.

Offermo, R. (2016). *Augmented reality sandbox as a tool to facilitate learning for undergraduate students in the bachelors program of earth sciences* (Independent Project at the Department of Earth Sciences). Uppsala University: Sewden.

Sánchez, S.Á., Martín, L. D., Gimeno-González, M. Á., Martín-Garcia, T., Almaraz-Menéndez, F., & Ruiz, C. (2016). Augmented reality sandbox: a platform for educative experiences. In Proceedings of the *Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 599-602). Salamanca : Spain

Singtao Daily (2015.11.04). SWC AR sandbox teaches Geography（陳朱素華中學智能沙箱教地理）[in Chinese]. http://std.stheadline.com/yesterday/edu/1104go05.html

Vaughan, K. L., Vaughan, R. E., & Seeley, J. M. (2017). Experiential learning in soil science: use of an augmented reality sandbox. *Natural Sciences Education*, *46*(1), 1-5.

Woods, T. L., Reed, S., Hsi, S., Woods, J. A. & Woods, M. R. (2016). Pilot study using the augmented reality sandbox to teach topographic maps and surficial processes in introductory Geology labs. *Journal of Geoscience Education*, *64*(3), 199-214.

Wellsted, E. (2006). Understanding distant places. In Balderstone, D. (eds), *Secondary Geography handbook* (pp. 160-169). Sheffield: The Geographical Association.

Yang, E.F.Y. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Pupils' Learning Characteristics, Self-directed Learning Ability, and Mathematical Ability by Applying A Gamified Math Platform for Distance Learning

Euphony F. Y. YANG[a*], Tak-Wai CHAN[a]
[ab] *Graduate Institute of Network Learning Technology, National Central University, Taoyuan Country*
*euphony@ cl.ncu.edu.tw

Abstract: This study aims to examine a remote primary school teacher who collaborated with 42 university volunteer mentors by applying gamified mathematics platform "Math-Island" for distance learning to facilitate third-grade pupils' self-directed learning (SDL) and mathematical ability. This study uses Wilcoxon tests to analyze 21 weeks of Math Island learning data and mathematics abilities tests, and content analysis of 21 verbatim transcripts of teaching review meetings and 203 teaching records as evidence of the learning characteristics and the progress of pupils' self-directed learning ability. The result found that the pupils' learning characteristics related to volunteer mentors' teaching, classmates, and their own learning attitudes. The pupils under distance learning gradually developed their self-directed learning abilities in the second semester. Besides, the post-test of mathematics learning ability has improved significantly compared with the pre-test. Recommendations for future research are also presented at the end of this paper.

Keywords: Distance learning, Self-directed learning, Mathematical learning, Online mentor, Math Island

1. Introduction

The COVID-19 pandemic has shifted distance learning increasingly common in lives throughout the world (Fidalgo, Thormann, Kulyk, & Lencastre, 2020). Education is facing a difficult time, but also encountering a new opportunity. In remote schools, pupils often lack enough cultural stimulation and learning resources. During the COVID-19 pandemic, distance learning is necessary to be a good approach. Pupils can learn new knowledge by watching videos, such as massive open online courses (MOOCs), or by using digital technology to complete some questions after watching videos, such as online learning platforms like Khan Academy to support their learning process (Thompson, 2011; Vidergor & Ben-Amram, 2020). Individuals who learned and were responsible for their own planning, implementing, and evaluation of the effort were indicated as self-directed learning (SDL, Hiemstra,1994).

Distance learning students usually showed greater SDL (Khalid, Bashir, & Amin, 2020), and SDL may lead to successful learning performance. When students adopt distance learning, they may progress in structured learning without teachers' supervision (Zhu et al, 2022), but the critical point is using effective online learning platforms to do SDL. A well-structured learning platform can supplement students' learning information and help them pursue knowledge at their speed, engaging in the learning through immediate feedback. Previous studies had confirmed that timely and frequent feedback may contribute to online learners' performance (Goldsmith, 2014), while gamified learning platforms have a complete learning structure to increase students' learning interest and motivation (Yeh et al, 2019). Besides, the teacher can receive mentors' feedback and online log data to integrate the pupils' self-assessment features and ensure that pupils are properly engaged in learning (Schneider, & Council, 2021). Hence, to develop remote school pupils' SDL and improve their mathematical abilities, applying a gamified platform to support pupils' distance learning may be a good choice.

When pupils learn math under one-to-one distance learning with mentors' guidance may enhance their learning effect, but can pupils thus develop their SDL is still a question. According to the reasons above, a class teacher collaborated with volunteer mentors by applying the Math-Island to support the mentors' guidance for enhancing remote school pupils' mathematical learning and to see if

those pupils can foster their SDL ability through distance learning. Therefore, this study proposes three research questions as follows.

(1) What are the learning characteristics when pupils used the gamified learning platform for distance learning with the mentors' help?

(2) Can pupils develop self-directed learning abilities after using the gamified learning platform for distance learning with the mentors' help?

(3) Can "using the gamified mathematics platform for distance learning "improve the pupils' mathematics learning ability?

2. Methods

2.1 Research Design

This study conducted two sessions of distance learning once a week, each session is 40 minutes. The first semester had thirteen times of distance learning, while the second semester had ten times of distance learning. Before distance learning, a pre-test was conducted to understand the pupils' learning basis, and a post-test was also conducted at the end of the distance learning.

Pupils had to learn the Math-Island tasks the class teacher assigned. However, if some pupils learned faster, they could choose the learning tasks they felt interested, or go shopping the building on Math-Island. The volunteer mentors only taught the pupils when they made at least two mistakes. If the pupils were still wrong, the mentors may ask the pupils some questions as guidance or demonstrate how to solve word problems. If the pupils do not make mistakes, the mentors may just stand aside without giving any learning help in order to develop the pupils' SDL ability.

2.2 Participants

The participants in this study were a third-grade class teacher and all students, only 10 pupils (6 boys and 4 girls) from a remote primary school in northern Taiwan. There were 42 public university students who signed up online and volunteered to serve as online math mentors (6 males and 17 females in the first semester and 9 males and 10 females in the second semester), of which 16 volunteer mentors participated in more than half of service times, while 11 volunteer mentors only participated once.

Due to the Covid-19 pandemic, the third-grade students used the Math-Island platform to study during the after-school caring time. The mentors help pupils learn mathematics through one-to-one distance learning. The mentors played the role of asking mathematical questions rather than direct teaching instructors to scaffold the pupils' mathematical concepts and problem-solving, the pupils may thus internalize those mathematics questions into their learning and do mathematics correctly (Mason, 2000). The class teacher was responsible for the class order and solving the technological problems. She rarely intervened in mentors' instruction.

2.3 Implement Tools: Math-Island and Google Meet

In this study, pupils used the "Math-Island" to learn and practice mathematics. "Math-Island" is a gamified mathematics learning platform developed based on the curriculum guidelines in the mathematics field of the Taiwan Ministry of Education. It covers primary school mathematical concepts from first to sixth grades, which are presented through animated teaching videos and interactive games. Students can learn the mathematical concepts by themselves according to the guidance of concept maps.

After students complete the math tasks, the system would give corresponding rewards and feedback according to the correct answer rate, and the learning results can be transformed into a visual building appearance based on the correct answer rate and task completion rate. Besides, pupils can earn money and labor from completing math tasks to buy their buildings or ships. Pupils may plan and manage their learning by completing the mathematics tasks continuously to create the appearance of their island, and gradually accumulate their foundation of mathematics concepts, thereby improving their confidence and interest in mathematics learning. The class teacher can check pupils' learning history records to understand their learning progress and difficulties.

Each pupil used the online communication software Google Meet for one-to-one distance learning. The pupils needed to share their screen so that the volunteer mentor can see their learning task picture in Math Island, and the pupils can also see the teaching screen of the volunteer mentors in the same way.

2.4 Data Collection and Analysis

The purpose of this research is to explore the pupils' learning characteristics of using the Math Island learning platform for distance learning, whether volunteer mentors using the gamified mathematics platform for distance learning" enhance pupils' task numbers and correct rates, develop self-directed learning mode, as well as the mathematics learning progress. Four types of research materials were used to collect and analyze pupils' data in this study.

(1) LMS dashboard and log data: the LMS dashboard and log data on Math-Island were used to understand the pupils' learning characteristics, their correct rate, and the amount of learning tasks. The pupils' SDL ability means the number of learning tasks in which pupils kept their learning tasks on Math Island after distance learning without the class teacher assigning tasks.

(2) Transcripts of the briefing: in the teaching briefing, the online mentors would report students' learning performance and how they taught students that day. The teacher would provide feedback. Those processes were videotaped and transcribed to help understand pupils' learning characteristics.

(3) Teaching guidance records: the volunteer mentors had to fill out the teaching guidance record sheet after every teaching to help the class teacher understand students' myth concepts. This sheet was designed according to Bloom's (1956) teaching objectives and was divided into the cognitive, affective, and psychomotor domains. In other words, the volunteer mentors were invited to evaluate pupils' knowledge, learning attitudes, and skills in distance learning, the teaching strategies of volunteer mentors, as well as students' learning responses through online learning interaction. This document was used to analyze the mentors' teaching and the characteristics of pupils' learning behaviors.

(4) Pre-and post-test: This study uses the scores of " Test of Children Mathematics Achievement " developed by Xie (2014) as the pre and post-test tool to evaluate pupils' math learning ability and improvement individually. Pupils can listen to the explanation in the sample questions demonstration without reading long sentences during the test. The pupils listened to the question and replied with their answers in oral and written form. This standard test has 120 items. It targets intuitive mathematics and the law of mathematics, including zero concepts, addition concepts, carrying concepts and addition and subtraction, subtraction concepts, multiplication concept, division concepts, and mental arithmetic number lines for nine categories. Pupils who answer one question correctly can get one point. The internal consistency reliability coefficient of the test is .97, and the inter-rater reliability is above .90.

The study utilized three statistical analyses: descriptive analysis, Wilcoxon signed-rank test, and Pearson Correlation. All data were analyzed by using SPSS version 28 (SPSS Inc., Chicago, IL, USA). A Wilcoxon signed-rank test was used for analyzing pupils' learning tasks and correct rate, and pre and post-test scores for differences between the means in different periods and between SDL and distance learning. The Pearson Correlation was used to examine the relationship between math tasks and mathematical ability in post-test as the pupils' learning improvement. In addition, content analysis was used to analyze the qualitative data, such as meeting transcripts, and individual guidance records as supplement evidence of pupils' learning characteristics.

3. Result

3.1 Pupils' Learning Characteristics

This study intends to explore the learning characteristics of pupils who apply Math-Island for distance learning. The following introduces the number of math tasks the pupils completed, game engagement, and peers' influence.

As for the number of math tasks, the result shows that there are great differences among each pupil's learning task numbers. The pupils' average number of whole tasks is 219.7(SD=128.53). The pupil with the fewest participation did not have SDL, so she only did 82 math tasks, while the pupil

who had SDL and did the most math tasks is 449 tasks. Almost all pupils concentrated on the tasks of the third grade (M=107.3, SD=52.25) and the second grade (M=77.40, SD=48.66). However, the pupil with learning disabilities in the class mainly did the math tasks of the first and second grades, and only did four tasks of the third grade.

In addition, pupils rarely completed all tasks their teacher assigned during distance learning. the class teacher usually assigned pupils those math tasks below the third grade, so that pupils did fewer math tasks over the third grade. Only two high-achieving students completed tasks quickly and did nineteenth fourth-grade tasks and one task in fifth- and sixth-grade respectively.

Most pupils spend much more time when they encounter difficult mathematical concepts, such as fractions or two-step addition, multiplication, subtraction, and multiplication. Besides, pupils usually encounter difficulties with multiple-choice questions. A volunteer teacher indicated that "*When the student encounters a question with four options, he feels very annoying and doesn't want to read it.*" (MR121005CT), so the error rate is extremely high.

Although this study adopts the gamified learning platform for distance learning, students pay more attention to the math tasks the teacher assigned rather than playing games on Math-Island. Pupils usually buy new buildings and ships after or before the learning tasks. But the gameplay time decreased as time went by. And a special pupil did have any behaviors of gameplay. Only one pupil was interested in decorating her island. She can complete thirty math tasks to make money and change her buildings in her SDL. Her mentor encourages her to have more motivation in doing math tasks by using the strategy: "*Intersperse the questions with breaks, let her go to the Math-Island to buy things. She would come back and continue to do the math tasks for making money.*" (MR120307CT)

During distance learning, pupils may observe classmates' learning and compare mutual learning progress. A volunteer mentor responded that "*Ann saw that a high-achieving student in the class, Bob, solve the math island problems quickly. He would be a little impatient, and sometimes even asked Bob the answer.*" (MR120310CT) In other words, classmates' learning speed may put stress on pupils. Besides, the pupils would like to have the same teaching methods and tools as their classmates. The volunteer mentor mentioned that "*the pupils saw another volunteer mentor using small painters to teach her classmate, and she hoped I can use the same small painters to teach her.*" (MR120310CT)

3.2 Pupils' Task Amount and Correct Rates

This study investigates whether the pupils with those volunteer mentors' help can increase their correct rate in doing math tasks on Math-Island. Wilcoxon signed-rank test was used for comparing pupils' average math tasks for two different learning conditions, the number of math tasks of distance learning and the correct rate of SDL. The ranked pupils' mean number of math tasks in applying the gamified math platform for distance learning (6.43) had higher than the pupils in the SDL (3.44), which was also statistically significant ($Z = -2.395$, $p = .017 < .05$). These results indicate that pupils in distance learning with volunteer mentors' help can do more math tasks than they were in the SDL.

However, the volunteer mentor does not always speed up the pupils' learning. Sometimes the volunteer mentors may slow down the progress of the pupils' learning tasks when they evaluated that pupils do not understand certain mathematical concepts. The mentors may intend to explain and clarify students' mathematical concepts a whole course time on one math concept. Some volunteer mentors ask pupils to understand certain math concepts completely then they can go to the following math tasks causing pupils to spend too much time on one problem and do very few problems in total.

Therefore, this study also investigates whether the pupils with those volunteer mentors' help can raise pupils' correct rate in doing math tasks on Math Island. This study uses Wilcoxon signed-rank test as the statistical technique to compare the pupils' correct rate in the second semester for two different learning conditions, the correct rate of distance learning and the correct rate of SDL. The ranked mean scores (92.98) of the correct rate of the pupils in applying the gamified math platform for distance learning is significantly higher than that of SDL (85.56), Wilcoxon signed-rank test, $Z = -2.490$, $p < 0.05$. These results indicate that pupils in distance learning with volunteer mentors' help had a better correct rate than they were in SDL

According to the first author's participation observation and volunteer, mentors reflect their teaching in the teaching review meeting and guidance record, the volunteer mentors usually check pupils' answers before they press the confirm button. When the pupils give a wrong answer, volunteer mentors would stop them from pressing the confirm button, and then ask the pupils questions constantly

to clarify their mathematical concepts until they understand. Therefore, the accuracy rate is higher than SDL.

However, high-achieving students usually complete their answers directly and send their answers without waiting for the volunteer mentors to confirm. In contrast, pupils who have little confidence in mathematics usually wait for the volunteer mentors to check their answers.

3.3 The Development of Pupils' Self-directed Learning

This study used the nonparametric Wilcoxon Signed Rank Test for comparing pupils' math tasks between the first and second semester in the quantity of pupils' SDL by using the learning tasks on the Math Island pupils did on their own. The mean rank in the pupils' task amount in the second semester (M=29.50, SD=27.02) is higher than in the first semester (M=.90, SD=2.23). The Z value is -2.668 and the significance level is at a .008 significance level indicating a statistically significant SDL increased from the first to the second semester.

3.4 Pupils' Mathematics Learning Ability

Wilcoxon Signed Rank Test for comparing pupils' pre-test and post-test scores, the mean rank in the pupils' scores on the post-test (M=103.4, SD=18.37) is higher than the scores of the pre-test (M=76.6, SD=21.83). The Z value is -2.805 and the significance level is below .01. In other words, there was a significant difference between pupils' pre-test scores and post-test scores. In this study, those pupils who participated in the distance learning and were taught by volunteer mentors had a higher level of mathematics achievement from the post-test than the pre-test at a .005 significance level. After the given 13-week distance learning period, the post-test scores of the pupils in comparison with the pretest score increase significantly. Therefore, using the gamified mathematics platform for distance learning can improve the pupils' mathematics learning ability.

This study also uses Pearson Correction to examine the relationship between pupils' math tasks and their post-test mathematics scores. There was a strong and positive correlation between the completion amount of the third-grade math tasks (N=102) and post-test mathematics scores (M=103.40) with the values of r= 0.747, n= 10, and p= .013. This result indicates that the highly math-tasks completion was positively correlated with the post-test math performance. Therefore, the pupils doing more math tasks may enhance their mathematics learning performance.

4 Discussion

The purpose of this research is to foster pupils' SDL and mathematics abilities. Although the math tasks were assigned by the class teachers, the speed of learning is mainly determined by the pupils. However, when pupils encounter difficulties during learning, mentors may stop pupils from going to the next learning tasks and guide them in different ways. Sometimes mentors who lack teaching experience may result in pupils' ineffective distance learning and delay their learning speeds. Pupils may raise their correct rate but do fewer learning tasks.

The pupils who have better learning abilities may do more tasks. However, some high-achieving students don't need to do so many practice questions, they still have high math scores, and some low-achieving pupils with good mentors' guidance can do a lot of math tasks and develop their SDL.

Finally, although this study uses a game platform to support pupils' learning, but pupils did not engage in playing the game. This study observes that only a small number of pupils were affected by the mechanism of games and thus do a lot of math tasks. But they gradually reduce game participation over time and pay more attention to learning mathematics, which may relate to the purpose of classroom learning and teacher expectations. However, it requires further exploration and analysis.

5 Conclusion

To foster pupils' self-directed learning competencies, this study applied the mathematical the pupils were asked to do distance learning. This study examines pupils' learning characteristics, how the

volunteer mentors influence the pupils' learning, and whether pupils develop self-directed learning with the integration of distance learning through applying a gamified mathematical platform and improving their mathematical ability. The results showed that the pupils' learning characteristics about their learning variety such as math task completeness and game engagement, and were influenced by their classmates. Moreover, the volunteer mentors' teaching method may influence pupils' learning speed and correct rate than what pupils could do in self-directed learning independently. Furthermore, the pupils under distance learning gradually developed their self-directed learning ability in the second semester, and their mathematics learning ability in post-test also improved significantly and had a positive correlation with the practice amount of math tasks. Future research should focus on how the volunteer mentors teach the pupils and how to collaborate with the class teacher and change their teaching behaviors as they apply "Math-Island" to help pupils' distance learning.

Acknowledgments

The authors would like to thank the Ministry of Science and Technology of the Republic of China, Taiwan, for financial support (MOST 109-2511-H-008-011-MY3), and Research Center for Science and Technology for Learning, National Central University, Taiwan.

References

Fidalgo, P., Thormann, J., Kulyk, O., & Lencastre, J. A. (2020). Students' perceptions on distance education: A multinational study. *International Journal of Educational Technology in Higher Education, 17*(1), 1-18.

Hiemstra, R. (1994). Self-directed learning. In W. J. Rothwell & K. J. Sensenig (Eds.), The sourcebook for self-directed learning (pp. 9–20). HRD Press.

Khalid, M., Bashir, S., & Amin, H. (2020). Relationship between Self-Directed Learning (SDL) and Academic Achievement of University Students: A Case of Online Distance Learning and Traditional Universities. *Bulletin of Education and Research, 42*(2), 131-148.

Mason, J.(2000). Asking mathematical questions mathematically. *International Journal of Mathematical Educational in Science and Technology,31*(1),97-111.

Park, H., & Shea, P. (2020). A review of ten-year research through co-citation analysis: Online learning, distance learning , and blended learning. Online Learning, 24(2), 225-244. https://doi.org/10.24059/olj.v24i2.2001

Schneider, S. L., & Council, M. L. (2021). Distance learning in the era of COVID-19. *Archives of dermatological research, 313*(5), 389-390.

Thompson, C. (2011). How Khan Academy is changing the rules of education. Wired magazine, 126, 1-5.

Yeh, C. Y., Cheng, H. N., Chen, Z. H., Liao, C. C., & Chan, T. W. (2019). Enhancing achievement and interest in mathematics learning through Math-Island. *Research and Practice in Technology Enhanced Learning, 14*(1), 1-19.

Zhu, M., Bonk, C. J., & Berri, S. (2022). Fostering Self-Directed Learning in MOOCs: Motivation, Learning Strategies, and Instruction. *Online Learning, 26*(1), 153-173. DOI: 10.24059/olj.v26i1.2629

Moore, M. (1986). Self-directed learning and distance education. *International Journal of E-Learning & Distance Education/Revue internationale du e-learning et la formation à distance, 1*(1), 7-24.

Vidergor, H. E., & Ben-Amram, P. (2020). Khan academy effectiveness: The case of math secondary students' perceptions. *Computers & Education*, *157*, 103985.

Xie, R. S. (2014).學童數學成就測驗 [Test of children's mathematics achievement, 3-9 years]. Taipei, Taiwan: Psychological Publishing.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Learning Computational Thinking Efficiently with Block-based Parsons Puzzles

Jeff BENDER [a*], Alex DZIENA[a] & Gail KAISER[a]

[a]*Programming Systems Laboratory, Columbia University, USA*
*jeffrey.bender@columbia.edu

Abstract: To investigate learning system elements and progressions that affect computational thinking (CT) learning in block-based environments, we developed a Parsons Programming Puzzle (PPP) module within Scratch with scaffolding customized via a novel Blockly grammar. By varying the presentation and types of feedback encountered between- and within-subjects in a study of 579 adults, we identified features and scaffolding strategies that yield manageable cognitive load (CL), improved CT learning efficiency, and increased motivation, for a general populace. Findings indicate: 1) PPPs with feedback induce lowest CL; 2) an isolated palette, correctness feedback, and fading correctness feedback increase learning efficiency; 3) fading scaffolding can increase CT motivation. We analyze 12 conditions to provide insight to those developing block-based PPP systems with the aim to advance equitable CT education for all.

Keywords: computational thinking, parsons programming puzzles, scratch, motivation, cognitive load

1. Introduction

Increasingly, government and education leaders position computer science (CS) as a foundation useful for learning other disciplines, for empowering active citizenship, and for addressing inequalities (Blickstein, 2018). Computational thinking (CT) (Wing, 2006), whether characterized broadly as inclusive of the material and social elements of participatory computational literacy, or narrowly as the cognitive skills and knowledge necessary to reason effectively with a machine, has mobilized expanded access to CS (Proctor et al., 2018). Internationally, many schools have introduced compulsory coding (e.g. Whitehouse.gov, 2016; Hamilton-smith, 2016), and the CS exposure movement, inclusive of Code.org's Hour of Code (Code.org, n.d.) and many more (e.g. Girls Who Code, n.d., Black Girls Code, n.d.), reach students from schools that have not. This traction, while important, can limit focus on the sustained activity necessary to learn CT effectively (Blickstein et al., 2019).

The emerging discipline of learning engineering, which combines theory with data-driven analysis to develop educational methodologies that produce high-quality learning, offers the CS community an architecture to investigate efficient CT learning (Baker et al., 2022). This field encourages the instrumentation of learning environments, development of student learning models, and rich personalization supportive of equity. It guides us to design human-computer systems that harness and leverage the codependences between learning system, student, and teacher, to operationalize CT learning for all. This harmony among roles is important during an era of generational interdependency in which K-12 students and teachers alike are learning CT. In the U.S., 20% of CS teachers describe themselves as overwhelmed and 30% as under-qualified (Schaffhauser, 2018). The limited CT understanding can produce gaps in teacher pedagogical content knowledge, and testing-tuned curricula constrain time afforded to CT study (Celepkolu et al., 2020). While opportunities exist to infuse CT across curricula (c.f. Campos et al., 2017; Pollock et al., 2019), especially in STEM subjects (Cao et al., 2020), efficient CT learning remains crucial for universal access initiatives to succeed.

Though learning CT and coding can be difficult for novices (Anderson, 1982; Pea, 1986; Winslow, 1996), an approach offering efficiencies involves Parsons Programming Puzzles (PPPs), which are program completion exercises that facilitate learners practicing CT by assembling into correct order programming constructs that comprise examples of well-written code, typically focused on a single concept (Parsons et al., 2006). While PPPs originated in text-based formats (Half-Baked

Software, n.d.; Ihantola et al., 2010), researchers have introduced them to block-based environments, like Looking Glass and Scratch (Harms et al., 2016; Bender et al., 2021). Scratch, independently, has had success in attracting a community of over one million active users (MIT Media Lab, n.d.), but it is often used as a minimally guided entry to CT, leaving learners with misconceptions regarding sequences, Booleans, loops, and events (Franklin et al., 2020; Grover et al, 2015, 2017), and with inconsistent capability to demonstrate increased skill over time (Scaffidi et al., 2012). By balancing the agency Scratch offers learners, the integration of PPPs provides structure that can focus learning more narrowly, but efficiently, on the cognitive skill and knowledge elements of CT (Bender et al., 2022).

This approach does not offer the learner experience in the broader aspects of participatory computational literacy, however. Like the strategy in (Brusilovsky et al., 2018), in which learners extend their skills from code comprehension to construction by reading examples, viewing animations, and addressing misconceptions before trying PPPs, we propose learners might benefit from PPP practice in which scaffolding fades as CT concept familiarity develops. By decrementing supports provided by puzzle presentation and feedback, we can offer learners direct instruction to start (Kim et. al, 2005), then provide pathways toward open-ended CT via creation of personally meaningful artifacts, in alignment with Scratch's roots in constructionism (Brennan, 2013; Garner et. al, 2019).

To investigate this approach, we ran an online study targeting adults who were randomly selected into 1 of 12 conditions to learn the CT concept *looping* via 7 PPPs. Between conditions, we varied the presentation of programming constructs, the feedback types, and for control conditions, feedback activation or the CT concept. In 3 conditions, we faded scaffolding as learners progressed by suppressing correctness feedback and/or by presenting additional palette and block options. We investigate 3 research questions: after at most 56 minutes of puzzle solving, what are the effects of PPP variation on adult learner: **R1**) cognitive load (CL)?; **R2**); learning efficiency?; **R3**) CT motivation? Findings from 579 participants indicate **F1**) PPPs with feedback induce lowest CL; **F2**) an isolated palette, correctness feedback, and fading correctness feedback increases learning efficiency; **F3**) fading scaffolding can increase CT motivation. We first review related work and the software developed, then the study purpose, summative evaluation, and results, before previewing future work.

2. Related Work

Since PPP emerged as a new form of program completion problem in 2006, (Ericson et al., 2018) has assembled research results, (Du et al., 2020) has presented the variants in the literature, and (Bender et al., 2021) has documented strengths and weaknesses. Generally, PPPs offer learning efficiency gains because scaffolded support enables students to train in shorter time than via code writing exercises, while performing similarly on transfer tasks using either approach. PPPs also have led to more learning gain than alternative formats such as animations and annotated examples (Brusilovsky et al., 2018).

Our PPP integration with Scratch differs from previous PPP implementations by its embrace of economical game-based learning (GBL) instructional design, the variability of the scaffolding presented to the learner, and its embedding in a popular block-based environment. By leveraging game thinking (Kim, 2018), a blend of game design (Salen et al, 2003) with design thinking (Plattner, 2013), we facilitate the construction of engaging learning experiences with limited content development investment compared to serious games, as recommended in (Dickey, 2006). Educators can design motivating learning progressions that increase in difficulty while fading scaffolding and assessing implicitly, as in stealth assessment (Shute et al., 2021), as the learner masters individual CT concepts.

Like most PPPs, our implementation provides feedback as the learner positions programming constructs. We extend this with the configurability of per-construct points, a gameful scoring algorithm inspired by the longest common subsequence strategy in (Karavirta et al., 2012), target minimal moves needed to solve each puzzle alongside a move counter, and the inclusion of multiple feedback types. In addition to correctness feedback, we enable the customization of messages and actions triggered when certain puzzle solution conditions are (or are not) satisfied. This facilitates the design of intermediary objectives within puzzles, similar to subgoal labeling shown to help students solve PPPs (Morrison et al., 2016). The objectives constructed operate as test cases, like the execution-based approach described in (Helminen et al., 2013) and the auto-grader methodology in (Haldeman et al., 2018), though we perform static rather than dynamic analysis as the learner positions each construct, and provide dynamic

execution feedback concurrently via auto-execution of gameful animations in the Scratch stage. Using the feedback classification documented in (Raubenheimer et al., 2021), we view this feedback as constructivist because it is problem- and instance-oriented, which is a category of assistance associated with significantly lower student failure rates than alternatives, like solution- and theory-oriented.

We fortify the GBL features with those enabling the instructor to vary scaffolding, similar to the text/example/bug modulation in the BOTS study in (Zhi et al., 2018). Like text-based PPPs, the instructor can assemble constructs in an isolated palette for use during puzzle play. Alternatively, s/he can design a PPP requiring navigation through existing Scratch palettes (e.g. Control, Operators). Since the optionality might overwhelm, in a manner historically similar to sp/k (Holt et al., 1977) and more recently like (Cazzola et al., 2015; Rose et al., 2018), the instructor can select which blocks to enable in each palette, and further, which palettes to enable. This configurability facilitates the design of progressions in which learners encounter an isolated palette before transitioning to increasing numbers of palettes with increasing numbers of blocks enabled. This scaffolding fading can occur in the feedback dimension separately or simultaneously, as the instructor can also vary the correctness and objective feedback activation. As in (Dicheva et al., 2018), in which PPPs are used in exercises 1-2 and code writing is used in 3-4, the scaffolding variability allows for the introduction of CT concepts with constraints before relaxing those as learning proceeds to afford learners increasing agency.

Since the PPPs are in Scratch, learners who develop cognitive CT mastery as scaffolding fades become familiarized with the UI, enabling them to advance to epistemologically pluralistic elements of the curriculum in which they can use CT for open-ended creation (Turkle et al., 1990). Results from a related previous study indicated that the constraint of an isolated palette can increase learning efficiency and motivation, and the inclusion of correctness feedback can increase motivation (Bender et al., 2022). In the current study, we introduce new instructional design and feedback capabilities, and explore how CL, performance, learning efficiency, and motivation are affected by scaffolding variation.

3. Software Development

To investigate **R1-R3**, we developed an instructional design module using Blockly (Google for Education, 2018), and integrated the functionality with earlier modifications to Scratch that enable the design, play, and assessment of PPPs. While previous work introduced block points, scoring, progress bars, configurable blocks and palettes, and correctness feedback, the learning system could not fade feedback. Though feedback increased motivation in earlier studies, some participants noted anxiety, stress, worry, and anger when informed of mistakes move-by-move (Bender et al., 2022). To bridge the gap between correctness feedback and none except the animation, we defined a novel grammar and an objective editor for instructors to devise objectives with associated feedback messages and actions for each puzzle. Grammar details are available in this article's supplement site: https://bit.ly/3wdCowW.

Since teachers learn CT as they infuse it across subjects (Committee for the Workshops on CT, 2010), we designed the objective editor to be simple to use and include it in the concrete architecture in (Sulaiman et al., 2019) as a module exposed only to instructors. We frame it as a tool situated within teachers' practice intended to promote teacher learning, in the spirit of educative curriculum materials (Ball et al., 1996). The teacher should use pedagogical content knowledge to design useful objectives, such as the timely identification of repeating sequences and loop sentinels.

For example, consider a case in which a teacher introduces the CT concept *looping* to her students, after they have mastered *sequences*. If the puzzle designed requires a sprite to move in a path of a square, a student might begin by selecting a move-steps block, with the intention to later include three additional move-steps blocks, and four turn-left blocks, as shown in Figure 1a. However, the teacher can configure the puzzle to exclude the move-steps block initially, and use the objective editor to author guidance for the student to identify a repeating pattern s/he could embed in a control construct. Figure 1b presents an objective that might instead lead to the solution depicted in Figure 1c. First, the objective in 1b specifies that until the student uses the repeat block, she will receive feedback guiding her toward pattern identification; second, it records the feedback she will receive once she uses the repeat block; third, it designates three actions that will occur once the repeat is used: the move-steps and turn-left blocks will become enabled, and five points will be added to the student's objective score. An example of objective feedback during puzzle play in this study is show in the top-right of Figure 1d.

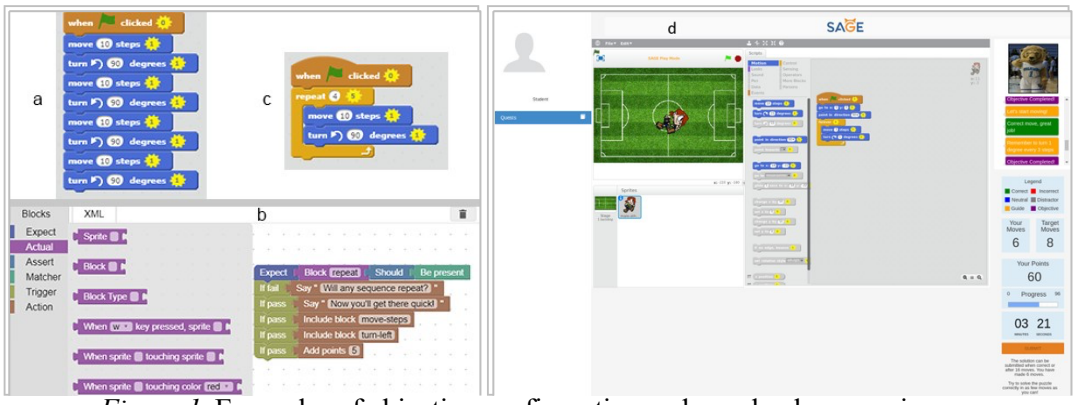


Figure 1. Examples of objective configuration and puzzle play experience.

4. Study Purpose

The software developed prepared us to investigate scaffolding variation. One study purpose was to identify which scaffolding approaches produce the lowest CL and most efficient CT learning. Given the limited skill increases over time and misconceptions developed by Scratchers discussed in section 1, the argument regarding the difficulties with minimally guided instruction in (Kirschner et al., 2006) appears applicable to CT learning in Scratch. By augmenting that learning experience with presentation and feedback modifications and comparing the outcomes, we aimed to contribute insight into the types of guidance that might help accelerate the proliferation of CT learning for the general populace.

A second study purpose was to explore the motivational impact of CT learning. In an era in which negative perceptions of CS persist (Schulte et al., 2005), self-directed learning in older adults is rising (Morrison et al., 2019), and deficits in adult CS knowledge limit parental involvement in children's CT learning (Bresnihan et al., 2019), insight into the attitudinal change provoked by a short CT learning intervention could inform future outreach. (Wang et al., 2015) emphasizes the crucial role family plays in encouraging women to participate in computing, (Ohland et al., 2019) offers arguments for educating parents on how to support CT learning, and (Sepúlveda-Díaz et al., 2020) offers guidance for supporting parents homeschooling their children in CT. We extend this work by examining whether the strategy of fading scaffolding leads to higher CT motivation. While a common approach, high-to-low scaffolding has not always yielded optimal outcomes, particularly in research on preparation for future learning (Chin et al., 2010). For example, (Blickstein et. al., 2018) found that open-ended instruction followed by detailed guidance led to higher learner activation in subsequent activities than the converse, indicating a carryover effect referred to as epistemological persistence.

We include 11 conditions on the CT concept *looping*, plus one concept control on *sequences*, and analyze the variation of presentation and feedback in 7 puzzles, each time-boxed for 8 minutes, that connect via a narrative with explicit goals related to playful animations. Based on the results in section 2, our hypotheses associated with **R1-R3** were: **H1**) PPPs with feedback yield lowest CL; **H2**) an isolated palette yields highest learning efficiency; **H3**) fading scaffolding yields highest CT motivation.

5. Summative Evaluation

5.1 Study Design

To test these hypotheses, we developed a quantitative experiment between-subjects, with some within-subject measurements. A study of perceived preferences for game elements in learning environments (Denden et al., 2018), as well as the review in (Santos et al., 2018) of the use of game elements for learning programming, inspired us to conduct a similar review of PPP elements in the PPP literature. However, we found limited research distinguishing which PPP elements prove most useful for learning (c.f. Ericson et.al, 2018, Bender et al., 2022). Consequently, we devised 12 study conditions detailed in Table 1 to ensure we could granularly analyze the effect of various PPP elements on learning outcomes

and help fill this gap in the literature. As documented in Table 2, participants followed a 10-step protocol which in part required them to respond to a validated: 1) CS CL component survey (CS CLCS) (Morrison et al., 2014); 2) intrinsic motivation Task Evaluation Questionnaire (TEQ) (SDT, 2022); 3) Computing Attitudes Survey (CAS) (Bockmon et al., 2020) (details at https://bit.ly/3wdCowW).

Table 1. *PPP scaffolding variation across 12 study conditions*

Cond.	CT Concept	Presentation	Feedback	Fading Scaffolding
C1	Looping	1-palette	Correctness	No
C2	Looping	Multi-palette	Correctness	No
C3	Looping	1-palette	Correctness + Objectives	No
C4	Looping	Multi-palette	Correctness + Objectives	No
C5	Looping	1-palette	Objectives	No
C6	Looping	Multi-palette	Objectives	No
C7	Looping	1-palette	3 Correctness, 2 Correctness + Objectives, 2 Objectives	Correctness feedback faded
C8	Looping	Multi-palette	3 Correctness, 2 Correctness + Objectives, 2 Objectives	Correctness feedback faded
C9	Looping	4: 1-palette, 3: multi-palette	3 Correctness, 2 Correctness + Objectives, 2 Objectives	Correctness feedback & 1-palette faded
C10	Looping	Multi-palette	None	No
C11	Looping	Multi-palette + distractors	None	No
C12	Sequences	1-palette	Correctness	No

Table 2. *Study protocol and data collected*

Step	Activity	Content	Data Collected
1	Registration	Credentials creation & condition assignment	Username & password
2	Background info	Demographic	Demographic, programming attitude, CAS, CT perceptions
3	Tutorial	8-minute video on the learning system & *looping*	N/A
4	Pretest (isomorphic)	7 multiple-choice (4-choice) *looping* questions	Pretest responses & score
5	CS CLCS	10 CL questions with 0-10 scaled responses	Pretest CL & IL/EL/GL components
6	Puzzles	7 puzzles on *looping* in 11 of 12 conditions; learning system behavior varies by condition; 7 puzzles on *sequences* in the 12th condition (control)	Per-puzzle time spent, time-stamped block moves & score, correctness, feedback log, self-explanations
7	CS CLCS	10 CL questions with 0-10 scaled responses	Puzzle CL & IL/EL/GL components
8	Posttest (isomorphic)	7 multiple-choice (4-choice) *looping* questions	Posttest responses & score
9	CS CLCS	10 CL questions with 0-10 scaled responses	Posttest CL & IL/EL/GL components
10	Concluding measurements	Motivation, programming attitude, learning system feedback, CT perceptions	TEQ, CAS, & programming attitude, CT perceptions

5.2 Participants

Using Prolific (Prolific, n.d.) we recruited 579 participants with varying degrees (57% high school, 27% undergraduate, 16% graduate), and a variety of self-reported programming experience (low: 50%; medium: 36%; high: 14%). 405 men, 167 women, and 2 non-binaries comprise the population sourced from 28 countries led by Poland (21%), Portugal (14%), and the U.K. (14%). Since the software development involved Scratch 2.0, which depends on Flash, a technology sunset at the start of 2021, learners needed to download a virtual machine we equipped to bypass Flash disablement. While necessary, this requirement introduced a risk to the external validity of the sample general population, as participants included only those sufficiently capable of installing virtualization software.

5.3 Data Collection & Processing

We created 7 surveys and instrumented the learning system to: 1) record puzzle play duration; 2) trace each block moved; 3) calculate score using an algorithm detailed in (Bender et al., 2021) that results in higher scores as construction nears the solution. To quantify the effect on CT learning, we also recorded self-reported CL. For learning to occur effectively, the CL of complex tasks should be reduced, though

the reduction need not occur in all three dimensions of CL (Sweller, 2010). The total number of interacting elements perceived by the learner determines intrinsic load (IL); the sometimes-impeding presentation of the content determines extraneous load (EL); and the instructional features necessary for schema construction determine germane load (GL). For example, conditions with an isolated palette and with blocks that become enabled incrementally as puzzle play unfolds aim to reduce EL otherwise required for interface navigation and block search to free learners' capacity to contend with GL.

To account for learners who compensate for an increase in CL by committing more mental effort, resulting in constant performance while load varies, we calculate instructional and performance efficiency (IE: learning process, PE: learning outcome) (van Gog et al., 2008). Previous studies have found lower PE for PPPs with randomly distributed distractor blocks compared to PPPs (Harms et al., 2016); higher IE for PPPs than for writing code (Bender et al., 2021); and higher IE for PPPs with 1 palette compared with multiple with distractors (Bender et al., 2022). We calculated IE and PE using both time and CL during training and transfer tasks. Since the data did not exhibit Shapiro-Wilk normality ($p<.05$), we used non-parametric statistics, like Kruskal-Wallis H and Mann-Whitney U between-subjects, and Wilcoxon within-subjects. For effect sizes, we used values in (Fritz et al., 2012).

6. Analysis & Results

6.1 Cognitive Load

We did not find significant differences between conditions in self-reported CL during the pre-test. Likewise, the posttest yielded no significant CL differences, suggestive of the acquisition of cognitive structures of equivalent expertise (van Gog et al., 2008). However, we did find significant differences between training conditions with moderate effect for GL ($H(11)=26.08$, $p=.006$, $\epsilon^2=.05$). Using a Bonferroni-adjusted alpha of .004 (.05/12), significant difference remained ($p=.019$) between conditions C4 ($M=4.90$) vs. C11 ($M=5.27$). This indicates participants training with multiple palettes and both correctness and objective feedback required less mental effort to contend with instructional features necessary for schema construction than those who received no feedback while navigating multiple palettes that included distractor blocks not part of the puzzle solution. To isolate the mediating variables, we analyzed treatment condition sets grouped by scaffolding variation and found a significant difference in GL with small effect between sets of conditions that did or did not receive objective feedback ($U(N_{objectives}=359, N_{no\text{-}objectives}=200)=29{,}306.5$, $z=-3.60$, $p<.001$, $r=.15$, $M_{objectives}=4.78$, $M_{no\text{-}objectives}=5.66$). We also found with small effect significant (GL) and moderate (overall CL) differences between those who received any feedback and those who did not ($U(N_{feedback}=461, N_{no\text{-}feedback}=98)=18{,}537.5$, $z=-2.79$, $p=.005$, $r=.12$, $M_{feedback}=4.95$, $M_{no\text{-}feedback}=5.78$; $U(N_{feedback}=461, N_{no\text{-}feedback}=98)=19{,}860.0$, $z=-1.88$, $p=.060$, $r=.08$, $M_{feedback}=4.65$, $M_{no\text{-}feedback}=4.97$). These results offer evidence that training with feedback limits the GL and CL experienced by the learner, supportive of **H1**.

6.2 Performance

During the transfer phase, the treatment population solved significantly more posttest than pretest questions correctly with small effect ($z=3.4$, $p<.001$, $r=.14$, $M_{pretest}=6.0$, $M_{posttest}=6.25$). Like in (Harms et al., 2016; Ericson et al., 2018; Bender et al. 2021, 2022), we did not find transfer performance disparity between PPP conditions. During training, however, we found significant differences between conditions in the time spent solving with relatively strong effect ($H(11)=39.29$, $p<.001$, $\epsilon^2=.20$), and with moderate effect, block moves made ($H(11)=113.8$, $p<.001$, $\epsilon^2=.07$) and puzzles solved ($H(11)=74.46$, $p<.001$, $\epsilon^2=.13$). Notable time-saving conditions, both with an isolated palette and feedback, are C1 ($M=21.6$) and C7 ($M=20.8$), which required significantly less time than C10 ($M=26.7$, C7 vs C10: $p=.033$) and C11 ($M=29.0$, C1 vs C11: $p<.001$) with multiple palettes and no feedback. Examining conditions sets led to significant differences with small effect between sets with an isolated palette and multiple palettes ($U(N_{1\text{-}palette}=211, N_{multi\text{-}palette}=348)=28{,}926$, $z=-4.21$, $p<.001$, $r=.18$, $M_{1\text{-}palette}=23.0$, $M_{multi\text{-}palette}=2.2$), with and without correctness feedback ($U(N_{correctness}=359, N_{no\text{-}correctness}=200)=29{,}188$, $z=-3.67$, $p<.001$, $r=.16$, $M_{correctness}=24.0$, $M_{no\text{-}correctness}=26.7$), and with and without fading scaffolding ($U(N_{fading}=155, N_{no\text{-}fading}=404)=27{,}032$, $z=-2.50$, $p=.012$, $r=.11$, $M_{fading}=23.5$, $M_{no\text{-}fading}=25.5$). This

indicates an isolated palette, correctness feedback, and fading scaffolding lead to less training time.

Condition sets with the lowest number of block moves were those with an isolated palette and with objective feedback ($U(N_{1\text{-palette}}=211, N_{\text{multi-palette}}=348)=31{,}653$, $z=-2.73$, $p=.006$, $r=.12$, $M_{1\text{-palette}}=78$, $M_{\text{multi-palette}}=84$; $U(N_{\text{objectives}}=359, N_{\text{no-obectives}}=200)=20{,}825$, $z=-8.24$, $p<.001$, $r=.35$, $M_{\text{objectives}}=74$, $M_{\text{no-objectives}}=95$). Those without objectives solved significantly more puzzles correctly with small effect, as did those with correctness feedback and fading scaffolding ($U(N_{\text{objectives}}=359, N_{\text{no-obectives}}=200)=28{,}819$, $z=-3.91$, $p<.001$, $r=.17$, $M_{\text{objectives}}=2.6$, $M_{\text{no-objectives}}=3.3$; $U(N_{\text{correctness}}=359, N_{\text{no-correctness}}=200)=42{,}220$, $z=3.49$, $p<.001$, $r=.15$, $M_{\text{correctness}}=3.1$, $M_{\text{no-correctness}}=2.5$; $U(N_{\text{fading}}=155, N_{\text{no-fading}}=404)=37{,}576$, $z=3.70$, $p<.001$, $r=.16$, $M_{\text{fading}}=3.4$, $M_{\text{no-fading}}=2.7$). These results indicate that objective feedback leads to the fewest puzzles solved and correctness feedback and fading scaffolding lead to the most.

6.3 Efficiency

While we did not find significant PE differences between treatment conditions, when using time spent as an estimate of mental effort, we found significant IE differences with moderate effect ($H(10)=27.73$, $p=.002$, $\epsilon^2=.05$), with condition pairs C1 vs. C11 ($p=.034$) and C7 vs. C11 ($p=.002$) remaining significant after Bonferroni adjustment. This indicates that training with an isolated palette with correctness feedback ($M_{C1}=.29$) and more so with an isolated palette with correctness feedback fading ($M_{C7}=.46$) leads to more efficient CT learning than with multiple palettes without feedback and with distractors ($M_{C11}=-.34$). When grouping condition sets by number of palettes, inclusion of correctness feedback, and fading scaffolding, we identified with a small effect significant differences for the first two and a moderate difference for the third ($U(N_{1\text{-palette}}=211, N_{\text{multi-palette}}=348)=41{,}474.0$, $z=2.57$, $p=.010$, $r=.11$, $M_{1\text{-palette}}=.15$, $M_{\text{multi-palette}}=.09$; $U(N_{\text{correctness}}=359, N_{\text{no-correctness}}=200)=40{,}025.0$, $z=2.25$, $p=.024$, $r=.10$, $M_{\text{correctness}}=.06$, $M_{\text{no-correctness}}=.11$; $U(N_{\text{fading}}=155, N_{\text{no-fading}}=404)=34{,}279.0$, $z=1.74$, $p=.082$, $r=06$, $M_{\text{fading}}=.11$, $M_{\text{no-fading}}=.04$). These results support **H2** and reveal additional learning efficiency opportunities via the use and fading of correctness feedback in addition to an isolated palette.

6.4 Motivation

6.4.1 Quantitative Results

To analyze motivation quantitatively, we scored the TEQ, and calculated the within-subject change in CAS scores, programming attitude, and CT perception, from study start to end. Condition sets with significant findings with small effect included those that varied by objective feedback for the interest/enjoyment subscale ($U(N_{\text{objectives}}=359, N_{\text{no-objectives}}=200)=94{,}948.0$, $z=-3.13$, $p=.002$, $r=.13$, $M_{\text{objectives}}=3.73$, $M_{\text{no-objectives}}=4.2$) and by objective and correctness feedback for the perceived competence subscale ($U(N_{\text{objectives}}=359, N_{\text{no-objectives}}=200)=32{,}801.5$, $z=-1.79$, $p=.074$, $r=.08$, $M_{\text{objectives}}=2.80$, $M_{\text{no-objectives}}=3.15$; $U(N_{\text{correctness}}=359, N_{\text{correctness}}=200)=40{,}229.0$, $z=2.30$, $p=0.21$, $r=.10$, $M_{\text{correctness}}=3.06$, $M_{\text{no-correctness}}=2.70$). Though narrowly missing significance, the objective feedback condition sets also showed noteworthy differences in the pressure/tension subscale ($U(N_{\text{objectives}}=359, N_{\text{no-objectives}}=200)=39{,}638.5$, $z=1.94$, $p=.052$, $r=.08$, $M_{\text{objectives}}=3.74$, $M_{\text{no-objectives}}=3.46$). These results indicate correctness feedback increases perceived competence, while objective feedback reduces interest/enjoyment, lowers perceived competence, and increases pressure/tension.

From the CAS scoring, we found significant within-subject change for the factors of: 3) importance, 4) problem solving – strategies, and 6) personal interest for the treatment population (3: $z=-5.3$, $p<.001$, $r=.22$, $M_{\text{start}}=3.94$, $M_{\text{end}}=3.76$; 4: $z=5.62$, $p<.001$, $r=.24$, $M_{\text{start}}=3.44$, $M_{\text{end}}=3.56$; 6: $z=4.28$, $p<.001$, $r=.18$, $M_{\text{start}}=3.42$, $M_{\text{end}}=3.61$). This decrease in importance and increase in problem solving strategies and personal interest largely held across conditions. For the gender equity (2) and gender bias factors (5), however, only the conditions with fading scaffolding yielded significant differences (2: $z=2.35$, $p=.019$, $r=.10$, $M_{\text{start}}=4.47$, $M_{\text{end}}=4.53$; 5: $z=-1.97$, $p=.049$, $r=.08.$, $M_{\text{start}}=2.01$, $M_{\text{end}}= 1.93$), indicating an attitudinal improvement in gender equity and reduction in gender bias.

Additionally, we found many significant programming attitude and CT perception improvements from study start to end per condition set; selected results are presented in Table 3. Overall, the quantitative motivation results provide partial support for **H3**, as fading scaffolding led to improved programming attitude and CT perception scores, but the lack of longitudinal data represents a threat to internal validity, since we cannot confirm the change measured at study conclusion persists.

Table 3. *Within-subject attitude and CT perception changes. *p<0.05, **p<0.01*

-- Progrmaming is… --	1-palette	Correctness Feedback	Fading Scaffolding
fun	M=.71**	M=.65**	M=.64**
enjoyable	M=.82**	M=.69**	M=.53*
easy to start	M=.94**	M=.67**	M=.73**
-- CT Perception --	--	--	--
I would recommend children in my family attend a CT camp or after-school program	no significant change	M=.35**	M=.42*
I would ask an employer for CT training	M=.31*	M=.39**	M=.71**

6.4.2 Qualitative Results

We studied motivation qualitatively by requiring that participants describe their attitude toward programming at study-end, as well as their perspectives on presentation and feedback scaffolding. Those with low self-reported prior programming experience often reflected surprise: 1) "It is perhaps easier than I thought once you get to grips with how it works and the different ideas and ways of doing it, and it is more accessible to the average person than I thought."; 2) "Starting it is much easier than I thought… it can be more fun after this practice I feel more motivated to find some time and finally give it a proper try." Those with higher prior experience advised on use cases: "this format is good for younger people to learn and get introduced to it. I think is needed to be a class in school about it."

Those who received correctness feedback reported the utility: "It helped me know when I was straying from course and re-evaluate my decisions. Very important and useful." Those receiving only objective feedback also perceived value: "The feedback was very motivating for me and each "correct answer" was giving me a lot of joy. I found this very helpful to know if I was heading in the right direction or missing out anything important, and the dynamically enabled blocks helped me make sure I was doing things in the right order." A C1 participant, with an isolated palette and correctness feedback, perceived challenge decreasing as s/he solved puzzles: "At the beginning of the task, I was preoccupied with how to navigate the system, but when I was familiar… I could easily solve the looping puzzles."

Many responses support **H3**, as they indicate fading scaffolding provides motivating challenge across puzzles, while increasing agency. Several participants in C9, in which both presentation and feedback scaffolding faded, reflected inspiration: 1) "When you start by having a lot of support and then it gets reduced, you need to rely on your skills and understanding to keep learning, and when you see that the program is working fine, it is more rewarding."; 2) "the ones with the same palette help at a very early stage to familiarize with the concepts, the ones with a different palette broke this familiarity and helps further by making you think more; 3) "The single palette made it easier to focus on how to organize the blocks, instead of wasting time looking for them. Having them easily accessible at first made it easier to find them and use them in the later tasks when the blocks were spread across multiple palettes." Some with fading feedback, however, noted difficulties: "I found the lack of feedback later on really hard to overcome. It helped me immensely at the beginning. Even though I was more familiar with the concepts and blocks etc, the complexity of puzzles was also increasing so I really struggled." One participant in C8, with multiple palettes and feedback fading, recommended offering learner-configurability: "Create a settings where you can turn on or turn off the feedback for the puzzle. This way, if the student wants to try solve the puzzle without help, they can turn off the feedback. However, if they find it too difficult and needs help, then they can turn on the feedback for guidance."

7. Conclusion & Future Work

To illuminate how learning system elements and progressions affect CT learning in block-based environments, we developed Blockly-scaffolded, configurable PPP functionality within Scratch that enables variation of presentation and feedback types. In a between- and within-subjects study of 579 adults, we enumerated several features and scaffolding strategies suitable for a general populace. PPPs with feedback yield lowest CL; an isolated palette, correctness feedback, and fading correctness feedback results in the highest learning efficiency; and fading scaffolding can increase CT motivation. The analysis offers PPP developers and instructors insight to advance efficient CT education for all.

While these results expose opportunities to advance CT learning via augmentations to popular block-based environments, we caution that developing effective technical interfaces between these systems is non-trivial given current implementation stratification. To echo and apply the call for standardization to advance progress in learning engineering in (Baker et al., 2022), we advocate for CT education data standards supportive of scaling consistent, quality data capture using extendable data models facilitating customization but rooted in commonality. Such anchoring could accelerate efforts to extend functionality in future work, for example by simplifying the way the Blockly objective editor could be integrated with the Scratch PPP evaluation engine to drive execution-based feedback like the js-parsons method described in (Helminen et al., 2013). It could also lead to data-driven generation of concept inventories and misconception maps, an approach proposed by EvoParsons developers in (Bari et al., 2019), that could better equip educators with the operational learning definitions necessary to infuse CT across curricula. With continued study, we aim to identify paths toward reliably efficient, effective, and equitable CT learning that builds bridges between cognitive, situated, and critical CT.

Acknowledgements

The Programming Systems Lab is supported in part by DARPA N6600121C4018, NSF CCF-1815494 and NSF CNS-1563555.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*.
Baker, R. S., Boser, U., Snow, E., & McNamara, D. (2022). Learning engineering. *Technology, Mind, Behavior*.
Ball, D., & Cohen, D. (1996). Reform by the book? *Educational Researcher, 25*(9), 6-14.
Bari, A., Gaspar, A., Wiegand, R., Albert, J., Bucci, A., & Kumar, A. (2019). EvoParsons. *GPEM*, 213-244.
Bender, J., Zhao, B., Dziena, A., & Kaiser, G. (2022). Learning computational thinking efficiently. *ACE*.
Bender, J., Zhao, B., Madduri, L., Dziena, A., Liebeskind, A., & Kaiser, G. (2021). Integrating Parsons Puzzles with Scratch. *ICCE.*
Black Girls Code. (n.d.). *Women of color in technology*. Retrieved from https://www.blackgirlscode.com/
Blickstein, P. (2018). *Pre-college computer science education.* Retrieved from https://goo.gl/gmS1Vm
Blickstein, P., & Hejazi, M. S. (2019). Computing education. In *Computing Education Research* (pp. 56-78).
Blikstein, P., Gomes, J., Akiba, H., & Schneider, B. (2017). The effect of highly scaffolded versus general instruction on students' exploratory behavior and arousal. *TKL*, 105-128.
Bockmon, R., Cooper, S., Gratch, J., & Dorodchi, M. (2020). Validating a CS attitudes instrument. *SIGCSE*.
Brennan, K. (2013). *Best of both worlds: issues of structure and agency in computational creation.* MIT.
Bresnihan, N., Srong, G., Fisher, L., & Lynch, Á. (2019). Increasing parental involvement in CS education. *CSE*.
Brusilovsky, P., Malmi, L., Hosseini, R., Guerra, J., Sirkiä, T., & Pollari-Malmi, K. (2018). An integrated practice system for learning programming in Python: design and evaluation. *RPTEL, 13*(1), 1-40.
Buffum, P. S., Lobene, E. V., Frankosky, M. H., Wiebe, E. N., & Lester, J. C. (2015). A practical guide to developing and validating CS knowledge assessments. *SIGCSE*, (pp. 622-627).
Campos, A., Rodrigues, M., Signoretti, A., & Amorim, M. (2017). piBook. *CSE*, (pp. 179-195).
Cao, L., Rorrer, A., Pugalee, D., Maher, M. L., Dorodchi, M., Frye, D., Barnes, T., Wiebe, E. (2020). Work in progress report: a STEM ecosystem approach to CS/CT. *ACM SIGCSE*, (pp. 999-1004).
Cazzola, W., & Olivares, D. M. (2015). Gradually learning programming supported by a growable programming language. *IEEE Transactions on Emerging Topics in Computing*, 404-415.
Celepkolu, M., O'Halloran, E., & Boyer, K. (2020). Upper elementary middle grade teachers' perceptions, concerns, and goals for integrating CS into classrooms. *SIGCSE*, (pp. 965-970).
Charters, P., Lee, M., Ko, A., & Loksa, D. (2014). Challenging stereotypes and changing attitudes. *SIGCSE*.
Chin, D., Dohmen, I., Cheng, B., Oppezzo, M. C., & Schwartzz, D. (2010). Preparing students for future learning with teachable agents. *Educational Technology Research and Development*, 649-669.
Code.org. (n.d.). *Hour of Code*. Retrieved from https://hourofcode.com/
Committee for the Workshops on Computational Thinking. (2010). *Report of a workshop on the scope and nature of computational thinking.* National Academies Press, Washington, D.C.
Denden, M., Tlili, A., Essalmi, F., & Jemni, M. (2018). Does personality affect students' perceived preferences for game elements in gamified learning environments? *Advanced Learning Technologies*, (pp. 111-115).
Dicheva, D., & Hodge, A. (2018). *Active learning through game play in a data structures course.* SIGCSE.
Dickey, M. D. (2006). "Ninja looting" for instructional design. *SIGGRAPH.*
Du, Y., Luxton-Reilly, A., & Denny, P. (2020). A review of research on parsons problems. *ACE*, (pp. 195-202).
Ericson, B., & Rick, J. (2018). Evaluating the efficiency and effectiveness of adaptive parsons problems. *ICER*.

Franklin, D., Salac, J., Thomas, C., & Krause, S. (2020). Eliciting student Scratch script understandings. *SIGCSE*.
Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates. *Experimental Psychology, 141*(1), 2-18.
Garner, J., Denny, P., & Luxton-Reilly, A. (2019). Mastery learning in CS education. *ACE*, (pp. 37-46).
Girls Who Code. (n.d.). *Girls Who Code*. Retrieved from https://girlswhocode.com/
Google for Education. (2018). *Bockly*. Retrieved from https://developers.google.com/blockly/
Grover, S., & Basu, S. (2017). Measuring student learning in introductory block-based programming. *SIGCSE*.
Grover, S., Pea, R., & Cooper, S. (2015). Designing for deeper learning in a CS course for MS students. *CSE*.
Haldeman, G., Tjang, A., Babes-Vroman, M., Bartos, S., Shah, J., Yucht, D., & Nguyen, T. (2018). Providing meaningful feedback for autograding of programming assignments. *SIGCSE*, (pp. 278-283).
Half-Baked Software. (n.d.). *Hot Potatoes*. Retrieved from http://hotpot.uvic.ca/
Hamilton-smith, L. (2016). *Learning Curve: Coding classes to become mandatory in Queensland schools*. https://www.abc.net.au/news/2016-11-17/coding-classes-in-queensland-schools-mandatory-from-2017/
Harms, K., Chen, J., & Kelleher, C. (2016). Distractors in parsons problems decrease learning efficiency for young novice programmers. *International Computing Education Research*, (pp. 241-250).
Harms, K., Rowlett, N., & Kelleher, C. (2015). Enabling independent learning of programming concepts through programming completion puzzles. *Visual Lanugages and Human-Centric Computing*, (pp. 271-279).
Helminen, J., Ihantola, P., Karavirta, V., & Alaoutinen, S. (2013). How do students solve parsons programming problems? *Learning and Teaching in Computing and Engineering*, (pp. 55-61).
Holt, R. C., Wortman, D. B., Barnard, D. T., & Cordy, J. R. (1977). SP/k. *Comms. of the ACM , 20*(5), 301-309.
Ihantola, P., & Karavirta, V. (2010). Open source widget for parson's puzzles. *Innovation and Technology in CSE.*
Karavirta, V., Helminen, J., & Ihantola, P. (2012). A mobile learning application for parsons problems. *CER*.
Kim, A. J. (2018). *Game thinking.* gamethinking.io.
Kim, T., & Axelrod, S. (2005). Direct instruction. *The Behavior Analyst Today, 6*(2), 111-120.
Kirschner, P., Sweller, J., & Clark, R. (2006). Why minimal guidance during instruction does not work. *Educational Psychologist, 41*(2), 75-86.
MIT Media Lab. (2022). Retrieved from Scratch Statistics: https://scratch.mit.edu/statistics/
Morrison, B., Dorn, B., & Guzdial, M. (2014). Measuring cognitive load in introductory CS. *ICER*, (pp. 131-138).
Morrison, B., Ericson, B., & Guzdial, M. (2016). Subgoals help students solve Parsons problems. *SIGCSE*.
Morrison, D., & McCutheon, J. (2019). Empowering older adults' informal, self-directed learning. *RPTEL*, 1-16.
Ohland, C., Ehsan, H., & Cardella, M. E. (2019). Parental influence on children's CT in an informal setting. *ASEE.*
Parsons, D., & Haden, P. (2006). Parson's programming puzzles. *ACCE.*
Pea, R. D. (1986). Language independent conceptual "bugs" in novice programming. *ECR*, 25-36.
Plattner, H. (2013). *An introduction to design thinking: process guide.* Institute of Design at Stanford.
Pollock, L., Mouza, C., Guidry, K. R., & Pusecker, K. (2019). Infusing CT across disciplines. *SIGCSE*.
Proctor, C., & Blikstein, P. (2018). How broad is computational thinking? *Society of the Learning Sciences.*
Prolific. (n.d.). Online participant recruitment for surveys and market research: https://www.prolific.co/
Raubenheimer, G., Jeffries, B., & Yacef, K. (2021). Pedagogical feedback in programming learning. *ACE*.
Rose, S., Habgood, J., & Jay, T. (2018). *Pirate plunder: game-based CT using scratch blocks.* ACPIL.
Salen, K., & Zimmerman, E. (2003). *Rules of Play: Game Design Fundamentals.* MIT Pess.
Santos, A., Souza, M., Dayrell, M., & Figueiredeo, E. (2018). A systematic mapping study on game elements and serious games for learning programming. *Computer Supported Education*, (pp. 328-356).
Scaffidi, C., & Chambers, C. (2012). Skill progression demonstrated by users in Scratch. *HCI, 28*(6), 383-398.
Schaffhauser, D. (2018). *CS education week kicks off, but som teachers don't feel ready*. https://thejournal.com/articles/cs-education-week-kicks-off-but-some-teachers-dont-feel-ready.aspx
Schulte, C., & Magenheim, J. (2005). Novices' expectations and prior knowledge of soft. dev. *CER*.
SDT. (n.d.). Self-determination Theory: https://selfdeterminationtheory.org/intrinsic-motivation-inventory/
Sepúlveda-Díaz, C., Suardro Rojas, E., Simmonds, J., Guitierrez, F. J., Hitschfeld, N., Casanova, C., & Sotomayor, C. (2020). Lessons learned from introducting preteens in parent-led homeschooling to CT. *SIGCSE*.
Shute, V., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C., Kuba, R., Liu, Z.,Yang, X., & Sun, C. (2021). Maximizing learning without sacrificing the fun. *Computer Assisted Learning*, 127-141.
Sulaiman, J., Dziena, A., Bender, J., & Kaiser, G. (2019). SAGE-RA: a reference architecture to advance the teaching and learning of CT. *Embedding AI in Education Policy and Practice for Southeast Asia*.
Sweller, J. (2010). *Cognitive Load Theory: Recent Theoretical Advances.* Cambridge University Press.
Turkle, S., & Papert, S. (1990). Epistemological pluralism. *Women in Culture and Society*, 128-157.
van Gog, T., & Paas, F. (2008). Instructional efficiency. *43*(1), 16-26.
Wang, J., Hong, H., Ravitz, J., & Ivory, M. (2015). Gender differences influencing pursuit of CS. *ITiCSE*.
Weintrop, D., & Wilensky, U. (2017). Between a block and a typeface. *IDC*, (pp. 183-192).
Whitehouse.gov. (2016). *CS4All* https://obamawhitehouse.archives.gov/blog/2016/01/30/computer-science-all
Wing, J. (2006). Computational Thinking. *Communications of the ACM, 49*(3), 33-35.
Winslow, L. E. (1996). Programming pedagogy—a psychological overview. *SIGCSE Bulletin*, 17-22.
Zhi, R., Lytle, N., & Price, T. (2018). Exploring instructional support design for K-12 CE. *ITiCSE*, (pp. 747-752).

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

For People and Planet: A Pilot Study of an Educational Mobile Game on the Sustainable Development Goals

Maria Mercedes T. RODRIGO*, Louise Marie C. TULAYBA, Marianne U. LIM, Marlon C. MORALEJO, Ananea ARCEGA, Maria Celeste A. MARASIGAN-LASCANO & Allyah Elizabeth ORIO

Ateneo de Manila University, Philippines
*mrodrigo@ateneo.edu

Abstract: *For People and Planet: An SDG Adventure* refers to a freely available Android-based narrative adventure game and teacher resource pack that helps learners see the United Nations Sustainable Development Goals (SDGs) in their day-to-day lives. In this paper, we describe the results of a pilot study of the game with 29 middle school Science students from the Philippines. The purpose of the study was to determine whether the students gained new knowledge from the game and how they characterized the game experience. The study found that learning was uneven with most students achieving quiz scores of 65% or higher but with one cohort averaging only 40%. Despite the unevenness of the knowledge assessment, students agreed that the game made them feel competent, gave them a sense of flow, and was immersive, albeit not very challenging. Opportunities for improvement include the creation of versions of the game for other platforms, deployment of the game in other subject areas, and further editing of the teacher resource pack.

Keywords: Sustainable Development Goals, Mobile Game, STEM education, Philippines, Pilot Study

1. Introduction

In 2015, United Nation Member States adopted the seventeen (17) Sustainable Development Goals (SDGs) that aimed to end extreme poverty, improve everyone's quality of life, and to protect the natural environment from degradation by 2030 (UN, 2016). To fulfill these goals, all sectors of society had and continue to have a role to play. Schools were responsible not just for educating students about these SDGs but also for guiding their formation towards the practice of sustainability in their everyday lives (Filho, et al., 2019, Nousheen, Zai, Waseem, & Khan, 2020).

Teaching about the seventeen SDGs can be daunting as it requires adjustments in existing curricula and teaching methods (Williamo et al, 2018, Hensley, 2020). Furthermore, traditional methods of teaching tend to have low engagement among young learners in particular (Prensky, 2006). One innovative approach that has been shown to be an efficient and engaging method of explaining environment and sustainability related concepts is game-based learning (Katsakiali and Mustafee, 2012, Bevilacqua, et al., 2015).

For People and Planet: An SDG Adventure refers to an Android-based narrative adventure game that attempts to address some of these needs. Contextualized in the Philippines, *For People and Planet* addresses Filipino learners specifically and attempts to help them see the SDGs in their day-to-day lives. This paper describes a pilot study conducted with this game and attempts to describe both the pilot study design process and the pilot study's outcomes. We wanted to know if this innovative approach delivered was indeed an effective and engaging method of teaching and learning, i.e. whether they learned something from playing the game and whether they enjoyed the learning experience. The specific research questions we attempted to answer were:

- To what extent was the game able to fulfill the learning goals of the lessons?
- How did the students characterize the game experience?

2. *For People and Planet: An SDG Adventure*

As mentioned in the introduction, *For People and Planet: An SDG Adventure* is a narrative-based adventure game for the Android phone or tablet that helps learners see the SDGs in their day-to-day lives. Its design is described in detail in Rodrigo, Diy, and colleagues (2021). The game is intended as supplementary material to middle school lessons on SDGs. The design of *For People and Planet* echoes the themes of contextualization and activity-based tasks. All activities are grounded in the Philippine context by selecting situations that are typical of the Philippines and are linked to the Philippine basic education curriculum's Most Essential Learning Competencies (MELCS). The game is accompanied by a teacher resource pack that contains sample lesson plans, sample assessments, slide decks, and supplementary materials. This teacher resource pack is discussed in detail in (Rodrigo, Torres, et. al, 2021).

In the game, players assume the role of a middle school student in a rural community in the Philippines and embark on five (5) stories that can be played in any order. Each player is shown how the community encourages sustainability and maintains their environment through errands that the player accomplishes. Each story covers an aspect of life and ties it to one or more SDGs. In *What's For Lunch?*, the player is asked to buy food for the family's lunch (See Figure 1). In *Flood Fighters*, the player accompanies the grandmother and learns about disaster risk reduction and disaster risk management, both at the community and at the household levels. In *A Walk in the Park*, the player goes to a neighboring town with friends to learn about clean energy and wastewater treatment. In *Work, Work*, the player visits the community enterprise where the grandmother works. In *Learning is for Everyone*, after attending a theater workshop in school, the player befriends a student with low vision.

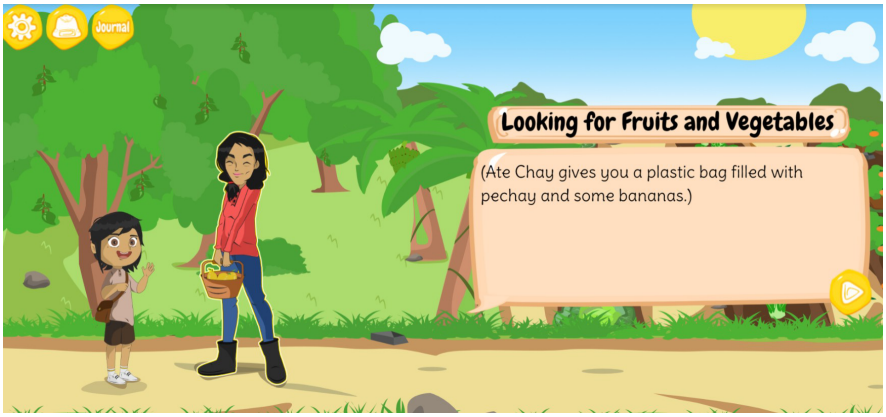


Figure 1. Ate Chay gives the player a basket of fruits and vegetables after the player completes the quest.

Within each story are mini-games that players have to accomplish in order to complete game quests. Mini-games were designed as simple, fun ways to provide more interaction and to break the monotony of reading character dialog and educational text. Some mini-games also provide an alternative way to present certain lessons. For example, in the ocean cleanup minigame, the player must collect as much trash as possible within the allotted time (See Figure 2) while in the Go-Bag mini-game, players must collect items such as flashlights and radios, needed for a go bag.

3. Pilot Study

This section describes the design and implementation of the pilot study including the participants, the lesson plan preparation, and the assessment design. The protocols for the pilot study were submitted to the University Research Ethics Office. The protocols were deemed to be exempt from review because this was research that was "conducted in established or commonly accepted educational settings, involving normal educational practices, such as … research on the effectiveness of … instructional techniques, curricula, or classroom management methods." (R. Cruz, Personal Communications). Data from all respondents was anonymized for this report.



Figure 2. Ocean Cleanup Mini-game.

3.1 Description of Teachers and Students

The pilot study was conducted in a private, all-boys school located in Metro Manila, Philippines. A total of four teachers (all female) and the Science subject area coordinator (male) participated in the study, now co-authors of this paper. On average, they had 8 years of experience teaching science at the grade school level. By the time they participated in this study, they had been teaching online for around 16 months because of the COVID-19 pandemic. They had previously undergone training related to teaching with technology. Training topics included G Suite for Education, Google Workspace, Robotics (Lego Education WeDo 2.0 and Lego Mindstorms EV3) and various apps for online classes.

A total of 29 students from grades 3 to 6 participated in the study, ranging in age from 8 to 11 years old. Of these 29, only 26 completed the pilot study.

As this was a pilot study as opposed to a full implementation, the teachers recruited student participants from among their classes based on the students' willingness to participate in after-school sessions, ownership or access to an Android device, and permission from their parents or guardians to join the study. Their participation was voluntary and had no impact on their final grades. Both the students and their parents gave written assent and consent to participate respectively. Furthermore, the study protocol was reviewed and approved by the University Research Ethics Office.

3.2 Lesson Plan Preparation

To prepare for the pilot study, the teacher-volunteers and the Science subject area coordinator met with the game developers for familiarization with the game and to discuss the purpose of the study. Since the subject area coordinator had the most comprehensive knowledge of the topics within the overall Science curriculum, he was the one who suggested which of the stories within the games would best fit the curriculum of each grade level.

For Grade 3, for example, the story "What's for Lunch?" was chosen because it is aligned with the third quarter lesson on food, water, and balanced diets. For Grade 6, the story "Flood Fighters" was chosen because it was aligned with the second quarter lesson on disaster preparedness, risk management, typhoons, and public storm signals.

After the meeting, the teachers downloaded and played the game on their devices in order to assess its age-appropriateness and to also foresee possible sources of difficulties on the students' end. The teachers also studied the resource pack (Rodrigo, Torres, et al, 2021). The teachers used the resource pack's references, sample assessment tools and other helpful information as they crafted their lesson plans. Teachers adopted most of the learning objectives articulated in the resource pack, adding to them as needed, e.g. the Grade 3 teacher added learning objectives regarding food production while the Grade 6 teacher added affective components that best fit the lesson as well as the game.

Each teacher created a lesson that blended with the curriculum, depending on the students' prior knowledge. In the case of Grade 6, the students had already taken lessons on typhoons and disaster preparedness prior to the use of the game. The teacher therefore created a lesson that leveraged on this knowledge and used an inquiry-based approach to teaching science called the Inquiry-Concept

Formation-Information Approach or I-C-I. This method helps students develop higher-order cognitive skills by going beyond the texts and look for answers to problems through well planned discussions and investigations.

During the Inquiry Stage, the students were asked to play the story “Flood Fighters" for 30 minutes. Afterwards, the teacher asked several questions which scaffolded Concept Formation, asking questions such as *What was the first activity that you were asked to do?* (Answer: Help in the relief efforts of the barangay) or *What did you do as part of the relief effort?* (Answer: Identify appropriate clothes for donation, classify relief goods).

The students answered based on their experiences within the game and the answers served as an anchor for the Information Stage in which key points of the lessons were restated. For example, teachers could reiterate that preparation for disaster is key and that before a disaster strikes, we should make sure we have the supplies and equipment in case there is no electricity, no communication, or no means to travel to buy what we need.

3.3 Assessments

At the end of the lesson, students were asked to take two types of assessments: a knowledge assessment and selected statements from the Game Experience Questionnaire (GEQ; IJsselsteijn, de Kort, & Peols, 2013). All the knowledge assessments included 5 questions that tested the students’ comprehension of the content of the game

The GEQ questionnaire on the other hand, was an instrument that measured the players’ experience. The complete GEQ consists of 33 statements that measured seven (7) components of a player’s experience: competence, sensory and imaginative immersion, flow, tension/annoyance, challenge, negative affect, and positive affect. From these 33, we selected 23 statements, making sure that all seven components were represented. Participants indicated their level of agreement (strongly disagree to strongly agree) with each of these statements.

At the end of the GEQ, the research team also included open ended questions asking what students liked most and least about the game, what about the game made learning the topic fun, interesting, or easy to learn, and what about the game made the topic boring or difficult to learn.

4. Results

In this section, we summarize the results of the pilot study into three sections: results of the knowledge assessments, results of the GEQ, and qualitative responses and observations.

Average student performance varied from grade level to grade level. Grade 5 students averaged the highest at 92% while the Grade 4 students did the poorest at 40%. The Grade 3 cohort scored 80% while the Grade 6 cohort scored 65%. The results imply that the game was a learning experience for most students, but it perhaps did not appeal to others. Why exactly the Grade 4 cohort scored so badly is unclear. We speculate that the students did not take the assessment seriously as it had no bearing on their final grades.

Despite the uneven knowledge assessment results, the GEQ results showed that the students regarded the game well. The scores for competence, sensory and imaginative immersion, flow, and positive affect were between “Agree” and “Strongly Agree.” The game did not elicit tension and annoyance nor negative affect. The challenge score, though, leaned towards “Disagree” which implied that the game was not very challenging.

The qualitative responses of the students and the teachers’ observations corroborated the GEQ results. One teacher noted that students arrived early during the game orientation session and expressed excitement to play the game. Most were able to install the game without difficulty but one student did not have access to an Android device. The game development team therefore prepared a Windows version of the game for him. While he was able to download and install the game, doing so required assistance from the development team.

One theme that emerged was that the game made learning fun. Students liked learning about how to help people and the environment. They also liked learning about the importance of work-life balance. One student said he liked how the game “has a real world relevance and … is fun.”

Students repeatedly said that they liked the mini-games in particular. The mini-games were interactive and made learning easy. Important lessons such as what to do in the event of a disaster were made interesting and fun through these activities. Teachers appreciated that the mini-games were set in the Philippine context, which made them relatable.

The mini-games also helped students appreciate some of the processes that they were not completely aware existed. For example, students are not aware of food production. The Grade 3 teacher said that students "think that food is always ready since they can find [it] inside their refrigerator." The mini-games in "What's for Lunch" help them that before the food becomes available in groceries and markets, all of the ingredients needed to make our healthy meals were harvested by farmers and fisherfolk. The game helps them value the hard work of farmers and fisherfolk which may make them become more responsible consumers.

Teachers observed that there was a good match between student learning preferences and the game format. One teacher said that students, "do not want the usual pen and paper activity. They want to be engaged and to explore. They easily get bored and they always look forward to do something new and unusual." The game therefore fit this preference.

Students also appreciated the game was intuitive. One of the teachers observed that the students needed a minimum of coaching, a few clarifications at most, in order to navigate their way through the game. The tasks in the game were clearly defined and achievable because instructions were clear. "The directions were easy to follow, and the flow of the game was easy to understand," one student said.
Students were eager to participate in discussions that followed gameplay. Teachers noted that it was easy to elicit responses from the students and that students shared their thoughts freely.
Some students did say that some of the text tended to be quite long. One student said, "Although I know that it is important and part of the learning process, I find the long sentences in the introduction a bit dull." Other students said they wanted more challenge and that the design was "a bit too cartoonish."

5. Conclusions and Recommendations

We embarked on this pilot study to answer to answer two main questions: To what extent was the game able to fulfill the learning goals of the lessons? How did the students characterize the game experience?

We found that the game succeeded in being a learning experience for most students, though one cohort was less successful. If the game is less effective for some learners, finding out why is something that could be explored in follow up studies, or during actually in-class deployment.

We also found that, in general, the students enjoyed the game. The characteristics that contributed to the positive game experience included the high level of interactivity of the mini-games, the valuable lessons that the game contained, and general ease of use.

The pilot test also helped the team identify some ways forward. From a technical perspective, the game development team needs to make an iOS version of the game available as well as a more easily installable Windows version.

In terms of potential for use, teachers saw how the game can be an effective way to introduce students to the SDGs, not just within Science classes but also in other subjects. For example, the games about social responsibility, food supply, sustainable cities (urban/rural communities) could be used as part of their Social Studies lessons. Some of the scenarios presented could be the springboard for essays and discourse in English or Mother Tongue. The teacher's responsibility is to make sure that the planned activities are age-appropriate.

In order to maximize this potential, teachers will need to familiarize themselves with the game in order to find alignment between the game's contents and their lessons. The resource pack was very helpful in this regard, but teachers observe that the resource materials are quite text-heavy and need to be trimmed.

Acknowledgements

We thank our design and development team composed of Johanna Marion R. Torres, Janina Carla M. Castro, Abigail Marie T. Favis, Ingrid Yvonne Herras, Francesco U. Amante, Hakeem Jimenez, Juan

Carlo F. Mallari, Kevin Arnel Mora, Walfrido David A. Diy, Jaclyn Ting Ting Lim, & Ma. Assunta C. Cuyegkeng, Ma. Rosario Madjos, Jose Isidro Beraquit, Rommel Ian De La Cruz, Fr. Johnny Go, SJ, Tomas Falgui, Dan Mark Restoles, and Toby Rubio. This work was made possible through the grant entitled, "For People and Planet: Gamified Tutorials for Middle School on Sustainable Development" from the University Research Council of the Ateneo de Manila University.

References

Bevilacqua, M., Ciarapica, F.E., Mazzuto, G., Paciarotti, C. (2015). Cook & Teach: Learning by playing. Journal of *Cleaner Production. 106*: 259-271.

Filho, W. L., Shiel, C., Paço, A., Mifsud, M., Ávila, L. V., Brandli, L. L., Molthan-Hill, P, Pace, P., Azeiteiro, U. M., Vargas, V. R., & Caeiro, S. (2019). Sustainable Development Goals and sustainability teaching at universities: Falling behind or getting ahead of the pack?. *Journal of Cleaner Production*, *232*, 285-294.

Hensley, N. (2020). Educating for sustainable development: Cultivating creativity through mindfulness. *Journal of Cleaner Production*. 243.

IJsselsteijn, W. A., De Kort, Y. A., & Poels, K. (2013). The game experience questionnaire. Technische Universiteit Eindhoven. Available online: https://research.tue.nl/en/publications/the-game-experience-questionnaire

Katsaliaki, K., and Mustafee, N. (2012). A survey of serious games on sustainable development. In: *Proceedings of The 2012 Winter Simulation Conference*, 300–312. Edited by C.Laroque et al., Piscataway, New Jersey: IEEE.

National Research Council. (2000). *Inquiry and the National Science Education Standards : A Guide for Teaching and Learning*. National Academies Press.

Nousheen, A., Zai, S. A. Y., Waseem, M., & Khan, S. A. (2020). Education for sustainable development (ESD): Effects of sustainability education on pre-service teachers' attitude towards sustainable development (SD). *Journal of Cleaner Production*, *250*, 119537.

Prensky, M. (2006). Don't bother me, Mom, I'm learning! How computer and video games are preparing your kids for 21st century success and how you can help. St. Paul: Paragon House.

Rodrigo, M. M. T., Diy, W. D., Favis, A. M. T., Amante, F. U., Castro, J. C. M., Herras, I. Y., Mallari, J. C. F., Mora, K. A., Torres, J. M. R., Cuyegkeng, M. A. C. (2021). A RECIPE for teaching the sustainable development goals. In M. M. T. Rodrigo, S. Iyer, & A. Mitrovic (Eds.), *Proceedings of the 29th International Conference on Computers in Education* (pp. 451-465). Asia-Pacific Society for Computers in Education.

Rodrigo, M. M. T., Torres, J. M. R., Castro, J. C. M., Favis, A. M. T., Herras, I. Y., Amante, F. U., Jimenez, H., Mallari, J. C. F., Mora, K, A., Diy, W. D., Lim, J. T. T. & Cuyegkeng, M. A. C. (2021). For People and Planet: Teachers' Evaluation of an Educational Mobile Game and Resource Pack. In M. M. T. Rodrigo, S. Iyer, & A. Mitrovic (Eds.), *Proceedings of the 29th International Conference on Computers in Education* (pp. 657-666). Asia-Pacific Society for Computers in Education.

UN. (2016). *The Sustainable Development Goals Report*. NY: United Nations. Retrieved from https://unstats.un.org/sdgs/report/2016/The%20Sustainable%20Development%20Goals%20Report%202016.pdf.

Willamo, R., Helenius, L., Holmström, C., Haapanen, L., Sandström, V., Huotari, E., Kaarre, E., Värre, U., Nuotiomäki, A., Happonen, J., Kolehmainen, L. (2018). Learning how to understand complexity and deal with sustainability challenges – A framework for a comprehensive approach and its application in university education. *Ecological Modelling. 370*: 1-13.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

An Analysis of Filipino Primary School Learners' Game Experience and STEM Interest within Minecraft

Jonathan CASANO [a*], Maria Mercedes T. RODRIGO [a]

[a]*Ateneo de Manila University, Philippines*
*jcasano@ateneo.edu

Abstract: What-If Hypothetical Implementations using Minecraft (WHIMC) is a set of Minecraft worlds that students can explore to learn more about astronomy, geosciences, and ecology. In this paper we present an attempt to analyze the game experience and STEM interest answers of students who played WHIMC as guided by instructions in custom teacher-created learning modules that incorporate its use. We describe the process of remodeling the game experience questionnaire of IJsselsteijn et al. (2013) and the STEM interest questionnaire of Lent et al. (2008) to fit the context of WHIMC. Open-ended questions were added to the GEQ to facilitate further interpretation of the GEQ answers. The GEQ was administered once (after playing WHIMC) and the SIQ was given twice (before and after playing WHIMC) both as out-of-game tests. There was a noticeable neutrality in the SIQ answers of the low-performers, particularly in the dimensions of Self-Efficacy, Choice Goals and Outcome Expectation. The neutrality in the Self-Efficacy dimension shifted towards a higher degree of agreement in the post test. High-performing students showed consistently higher levels of agreement to the SIQ items compared to low-performers. In terms of the GEQ answers, the high- and low-performing groups share the same pattern of scores for each experience dimension. Low-performing students had higher scores in the Challenge dimension due to feeling time pressure. Open-ended answers showed that game- and learning-related WHIMC features were most liked, module-related difficulties were the most disliked, and game-related features were most instrumental to learning. High-performers reported that module-related aspects made the learning more difficult than it should be and low-performers said the technical bugs were the primary hurdles that made learning difficult.

Keywords: WHIMC, STEM interest, Philippines, Game Experience, Minecraft

1. Introduction

The What-If Hypothetical Implementations using Minecraft (WHIMC) project is a set of Minecraft worlds that learners can explore as supplementary activities to learn more about science, mathematics, engineering, and technology in general and astronomy, geosciences, and ecology specifically. WHIMC makes use of Minecraft's natural physics and mechanics to simulate science-related concepts. Finally, WHIMC immerses learners in simulations of conditions on certain exoplanets and on alternate versions of Earth, logging both the ways in which learners traverse these words and the observations that learners make during their explorations.

When engaging learners in open-ended environments such as Minecraft, assessment of learning and the learning experience is always a challenge (Park, 2019; Short, 2012). How can we determine what effect playing Minecraft has on the acquisition of knowledge or skills? Do students enjoy the game experience? Are there some types of learners who benefit from the game format more than others? In this paper we present an analysis of the game experience and STEM interest answers of students who played WHIMC as guided by instructions in custom teacher-created learning modules that incorporate its use.

Our formal research questions are the following: **RQ1**: How can game experience and STEM interest be measured in open-ended learning environments such as Minecraft? **RQ2**: What effect does

WHIMC game play have on students' STEM interest scores? **RQ3**: What is the difference between the game experience scores of the high- and low-performing students?

2. Review of Related Literature

Measuring the effects of games, especially those that are open-ended and less structured games such as Minecraft, has been a topic of continuing research (Šajben, 2020). Educators who use Minecraft for their classes use a variety of assessment methods to ascertain the effect of the game on student learning and the learning experience (Ekaputra, 2013). Educators in Malaysia, for example, implemented a Minecraft module on fractions that allowed students to produce what they called fraction models, i.e. structures or objects that represent fractions (Tangkui and Keong, 2021). Using pre- and post-tests, they found that students who used the Minecraft-based module showed greater achievement than those who learned fractions using traditional teaching methods. They attributed the success of the module to greater interactivity and greater engagement. Researchers also made use of self-reports and teacher observations (Overby, 2015). In an Australian high school, students were guided through a problem-based learning module on designing and building an ideal house (Callaghan, 2016). At the end of the activity, students were asked to respond to a student feedback survey regarding the use of Minecraft in general and as an educational tool. These assessments concluded that students were engaged in an authentic activity that promoted collaboration and the development of problem-solving skills.

In this paper, we present the outcomes of a Minecraft-based learning activity. We focus specifically on the effect of Minecraft on STEM interest.

3. Data Collection

Four schools from the Philippines were invited to identify topics in their respective academic level's curriculum where the use of WHIMC could be integrated. Teachers were asked to design learning modules that come with out-of-game assessments. After going through a review process to check for curriculum alignment, feasibility and quality, these modules were then tested out on their students. Table 1 summarizes the demographic details of the respondents.

Table 1. *Summary of demographic details of respondent groups*

Respondent Group		Module Implementation Dates	Total Student Respondents	Year Level and Subject	Age Range
PH School 1	Class 1	9/6 – 9/17, 2021	16	Grade 6	10 – 12 y/o
PH School 2	Class 2	10/6, 2021	31	Grade 4 & 5	9 – 11 y/o
PH School 3	Class 3	11/8, 2021	49	Grade 11	13 – 16 y/o
PH School 4	Class 4 to Class 12	between 12/15, 2021 and 3/16, 2022	233	Grade 7 & 8	10 – 14 y/o

3.1 Data Collection Design, GEQ and SIQ

During module implementation, the students were first asked to answer the STEM Interest Questionnaire (SIQ). After answering the SIQ, students were allowed access to the WHIMC worlds and were asked to follow the instructions described in the learning modules prepared by the teachers. After going through the module. The students were asked to answer the SIQ again together with the Game Experience Questionnaire (GEQ). Lastly, the students were asked to answer the out-of-game assessment questions that come with the teacher-created learning modules.

The Game Experience Questionnaire (GEQ) used in this study is an abridged version of the instrument developed by IJsselsteijn et al. (2013) to measure the factors in a game that contribute to an *engaging gameful experience* described across 7 dimensions of the player experience namely, **I**mmersion [I], **F**low [F], **C**ompetence [C], **P**ositive Affect [P], **N**egative Affect [N], **T**ension [T] and **C**hallenge [CH].

Four (4) open-ended questions were appended to the GEQ. These questions were, *What was your favorite part of the module and why?, What was your least favorite part of the module and why?*, *What about WHIMC made the topic fun, interesting, or easy to learn?* and *What about WHIMC made the topic boring and/or difficult to learn?*

The STEM Interest Questionnaire (SIQ) used in this study is likewise an abridged version of an original Student Interest and Choice in STEM (SIC-STEM) questionnaire developed by Roller et al. (2018) which was based on the Social Cognitive Career Theory (SCCT) questionnaire of Lent, Brown and Hackett (2008) that the following dimensions: **S**elf-**E**fficacy (SE), **O**utcome **E**xpectation (OE), **I**nterests (I), **C**hoice **G**oals (CG) and **C**hoice **A**ctions (CA).

3.2 Performance Categories

Using the scores gathered from the out of game assessments, the respondents were grouped into high- and low- performers via the method described in the work of Conati and Kardan (2013). Following this method, the scores of the respondents were first verified as being normally distributed (**PH School 1**: $\bar{x} = 40.18$, $\tilde{x} = 42.5$, $\kappa = -0.68$, $S_{KP} = 0.51$, **PH School 2**: $\bar{x} = 12.78$, $\tilde{x} = 10.02$, $\kappa = -0.35$, $S_{KP} = -0.32$, **PH School 3**: $\bar{x} = 61.9$, $\tilde{x} = 65$, $\kappa = -1.18$, $S_{KP} = -0.41$, **PH School 4**: $\bar{x} = 82.22$, $\tilde{x} = 80$, $\kappa = -1.05$, $S_{KP} = -0.11$) and performance was defined using the standard deviation and means. High-performers (HP) are students who received a total assessment score (s) that is greater than or equal to one standard deviation above the mean ($HP = s \geq \bar{x} + 1\sigma$) and low-performers (LP) are conversely defined as ($LP = s \leq \bar{x} - 1\sigma$).

4. Analysis and Results

4.1 Analyzing the GEQ Answers

A total of 593 rows of GEQ answers were considered in the analysis. A coding process was conducted to categorize the answers to the open-ended questions. These categories that surfaced out of the coding process were Game, Learning Content, Module/Teacher, Technical A (game constraints) and Technical B (game bugs).

Three coders were invited to categorize a total of 2372 open-ended answers. Each coder was given access to a spreadsheet where only the class numbers and corresponding open-ended answers were presented and was instructed to code independently. 1347 answers (56.79%) were unanimously coded. 986 answers (41.57%), where 2 coders were in agreement followed the majority rule. The coders convened to reach a consensus for the remaining 39 answers (1.64%) that have been coded differently by each coder.

4.1.1 Insights from the Answers to the Player Experience Items

The answers to the GEQ were aggregated per performance category to create Figure 1.

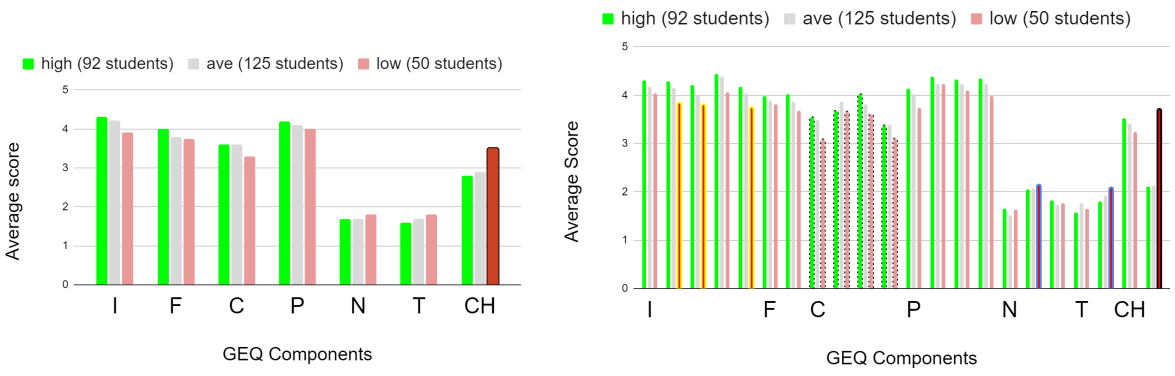


Figure 1. (left) GEQ component averages of the performance groups, (right) GEQ component player experience breakdown per performance group.

Low-performers found the game *moderately* aesthetically pleasing and impressive compared to the high-performers that found the experience *fairly* pleasing and impressive (red bars with solid yellow outline, 1 scale higher). The low-performers slightly outscored the high-performers in two items namely, *I found it tiresome* (N2) and *It felt frustrating* (T2) (red bars with blue outline), and greatly outscored the high-performers in the item *I felt time pressure* (CH2) (red bar with black outline). These observations were used to complement the insights drawn from the answers to the open-ended questions discussed in the next subsection.

4.1.2 Insights from Answers to the Open-ended Questions

Figure 3 shows a frequency distribution of the coded answers per performance category. We show how these findings relate to what was found in the analysis of the answers to the player experience dimensions above.

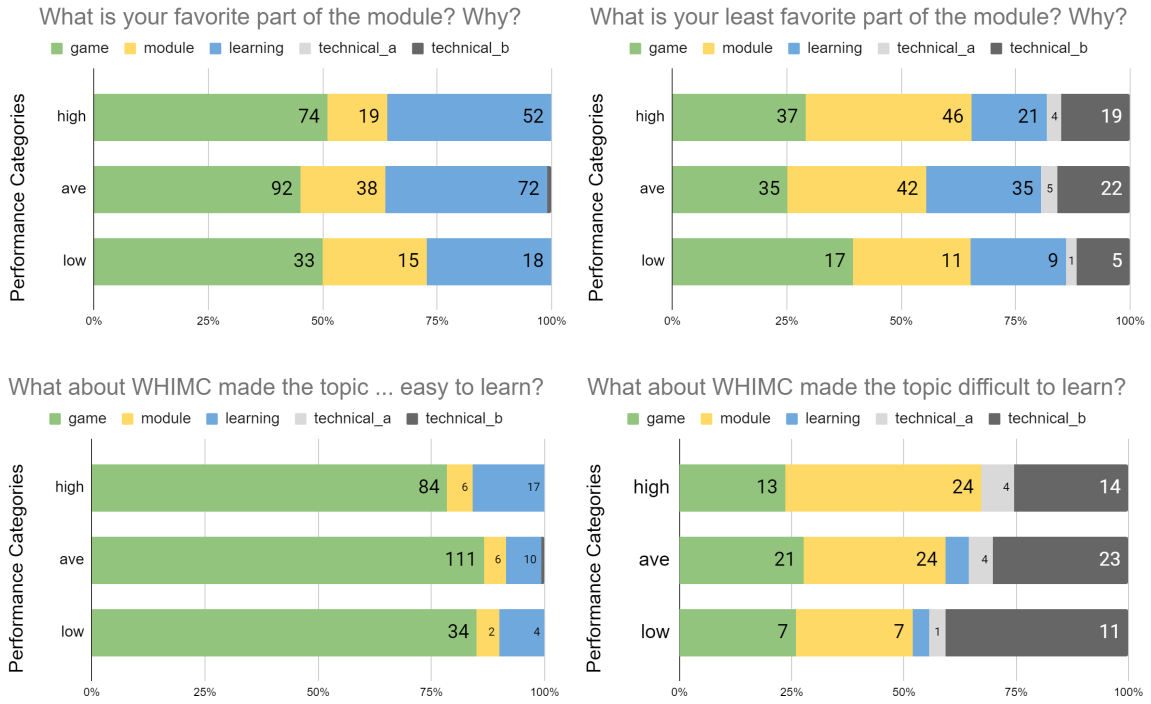


Figure 3. Answers to the open-ended questions appended to the GEQ.

In terms of most liked features, across the performance categories, students mentioned features related to the game as their most liked part of the experience. They were particularly pleased that the game allows for interaction with their classmates beyond the usual classroom discussion. Several students even noted that the session "*flew by so fast* (Flow F1) *because we were having fun* (Positive Affect P2) *exploring together* (Immersion I4)". This shared exploration might have been a factor that supported the observed high scores in the Immersion, Flow and Positive Affect dimensions (Figure 1). The students also liked meeting and receiving information from the NPC's about their work and the attributes of the altered Earths (44 out of 142 Learning answers or 30.0%) and how the simulations gave "*an interactive form to the information [they] normally would find in the textbooks*".

In terms of least-liked features, majority of answers for this question fell into the categories of *Game* and *Module*. Students felt the repetitiveness of making observations was tedious and boring. In terms of the *Module*, students felt that the tasks contained in the modules were too strict such that they did not blend well with the affordances of an open-world game such as Minecraft which they assumed should be centered around free exploration.

Across all performance categories, game related features were dominantly seen as the factor that contributed to fun, interest and ease of learning and what made the game boring or difficult to learn were features relating to Modules (inherent complexity of the lessons) and technical bugs.

4.2 Analyzing the SIQ Answers

As mentioned in the Methodology, the respondents were instructed to answer the SIQ twice. Once before playing WHIMC and another round after playing WHIMC. Figure 4 presents the divergent stacked bar charts showing the aggregated responses of high- and low-performers for each question in the SIQ.

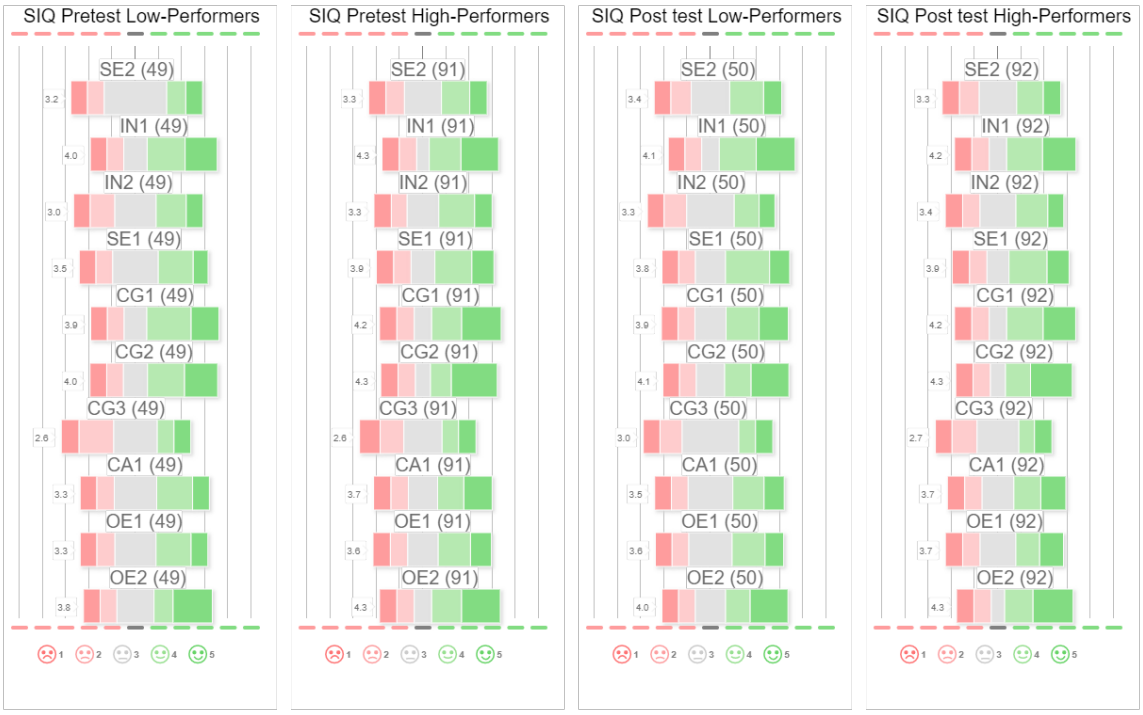


Figure 4. Divergent stacked bar charts representing the pretest and post test SIQ answers of high- and low-performing students.

4.2.1 Insights from the SIQ answers

At a glance, low-performers felt more neutral in most SIQ items than high-performers (gray bars). This is true for both the pretest and the post test. This neutrality is more pronounced in the questions pertaining to Self-Efficacy, Choice Goals and Outcome Expectation. This shows that the high-performing students generally believe in their ability to do STEM-related work and have more conviction in the significance of STEM concepts and skills in their future career.

The pretest and post test scores of the high-performing students stayed consistent. Most of the shift in the level of agreement was observed in the reported SIQ scores of the low-performing students. For instance, we note that when comparing the Pretest and Post-test graphs of the low-performing students, the neutralities for SE1 and SE2 have shifted towards a higher degree of agreement. This might mean that the WHIMC experience gave the low-performing students some degree of confidence in their ability to learn Science concepts while recognizing that it is also challenging.

Low-performers *enjoy solving Science and Math problems* (IN2) less than high-performers. High-performers also showed high agreement to CA1, OE1 and OE2 compared to the low-performers who were more neutral in their agreement to these items. These observations hold true for both the pretest and the post test.

5. Conclusion

In this paper we demonstrated how Game Experience and STEM interest may be measured in open-ended learning environments such as Minecraft by using WHIMC as a testing platform, and having students answer the GEQ and SIQ questionnaires after interacting with WHIMC [**RQ1**].

We note the following effect of using WHIMC in the STEM interest scores of the student respondents. There was a noticeable neutrality in the SIQ answers of the low-performers, particularly in the dimensions of Self-Efficacy, Choice Goals and Outcome Expectation. This shows that low-performers have a level of uncertainty with regards to their perceived ability to do STEM related tasks and well as to the significance of STEM in their future careers. In the post test however, the neutrality in the dimension of Self-Efficacy has decreased and shifted towards higher degree of agreement, showing that exposure to WHIMC might have equipped students with some level of esteem in doing STEM related tasks. High-performing students showed consistently higher levels of agreement to the SIQ items compared to low-performers [**RQ2**].

In terms of the GEQ answers, the high- and low-performing groups share the same pattern of scores for each experience dimension. The high-performers in the dimensions where a high score is desired such as Immersion, Flow, Competence and Positive Affect scored slightly higher than the low-performers, implying that the high-performing students had a slightly better overall game experience. Upon closer examination, low-performers felt time pressure with respect to the tasks contained in the modules. Perhaps giving these students more time to execute the instructions in the modules would be beneficial [**RQ3**].

Acknowledgements

The authors thank H Chad Lane and Jeff Ginger for their enthusiastic collaboration, Dominique Marie Antoinette Manahan, Maricel Esclamado, Mikael William Fuentes, and Ma. Rosario Madjos for their support, the Ateneo Laboratory for the Learning Sciences, the Ateneo de Manila University, our funding agency Department of Science and Technology (DOST) for the grant entitled, "Nurturing Interest in STEM among Filipino learners using Minecraft.", and our Monitoring Agency The Philippine Council for Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD).

References

Callaghan, N. (2016). Investigating the role of Minecraft in *educational learning environments*. *Educational Media International*, 53(4), 244-260.

Ekaputra, G., Lim, C., & Eng, K. I. (2013). Minecraft: A game as an education and scientific learning tool. *ISICO* 2013, 2013.

IJsselsteijn, W. A., De Kort, Y. A., & Poels, K. (2013). The game experience questionnaire.

Lent, R. W., & Brown, S. D. (2008). Social cognitive career theory and subjective well-being in the context of work. *Journal of Career Assessment*, 16(1), 6-21.

Overby, A., & Jones, B. L. (2015). Virtual LEGOs: Incorporating Minecraft into the art education curriculum. *Art Education*, 68(1), 21-27.

Park, J., Kim, S., Kim, A., & Mun, Y. Y. (2019). Learning to be better at the game: Performance vs. completion contingent reward for game-based learning. *Computers & Education*, 139, 1-15.

Roller, S. A., Lampley, S. A., Dillihunt, M. L., Benfield, M. P., & Turner, M. W. (2018, June). Student attitudes toward STEM: A revised instrument of social cognitive career theory constructs (fundamental). In 2018 *ASEE Annual Conference & Exposition*.

Šajben, J., Klimová, N., & Lovászová, G. (2020). Minecraft: Education Edition as a Game-Based Learning in Slovakia. *In the Proceedings of 12th annual International Conference on Education and New Learning Technologies* (pp. 1-8).

Short, D. (2012). Teaching scientific concepts using a virtual world: Minecraft. *Teaching Science*, 58(3), 55-58.

Tangkui, R. B., & Keong, T. C. (2021) Improving Achievements in Fractions: A Conceptual Framework for Digital Game-Based Learning Using Minecraft. *International Journal on E-Learning Practices (IJELP),* 4, 76-91

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Ludonotation: A Visualization System of Recording and Analyzing Interactive Strategic Gaming Behaviors

Ju-Ling SHIH
Graduate Institute of Network Learning Technology
National Central University, Taiwan
juling@cl.ncu.edu.tw

Abstract: Interactive strategic games (ISG) are dynamic and involve complex gaming behaviors and oftentimes intrinsic interrelated gaming strategies. Students' interactions would influence group dynamics, and players with different personality traits would behave differently in the game, and vice versa. This study develops a visualization tool: Ludonotation, a notation system of ludology, to record and analyze interactive gaming behaviors. The presentation of Ludonotation@ISG and its functions are accompanied by the example of the analysis of the house-made strategic game <Fragrance Channel v2.0>; the conceptual model is adaptive to other game genres. The paper shows how the gaming strategies of the players with different personality traits were modelled, and to understand the relationships between interactive gaming behaviors and gaming strategies.

Keywords: Ludonotation, Interactive Strategic Games, Gaming Behaviors, Gaming Strategies

1. Introduction

Interactive strategic games (ISG) are dynamic, and involve complex gaming behaviors and often times intrinsic interrelated gaming strategies. In most game analyses, players' single behaviors are identified individually and independently. Behavior features are extracted and the procedural data are analyzed only to predict their outcomes. Nevertheless, depicting their processes and defining their behavior patterns in social interactions, especially those interrelated, interactive, and intrinsic behaviors across players, become important to understand what people would do in the game, especially when people are with distinct personality traits and goals.

In order to do that, a visualization tool: Ludonotation, a notation system of ludology, is introduced to record and analyze interactive gaming behaviors both physical and conceptual. Ludology is game studies which rooted in anthropology and examining aspects such as game designs, player interactions, and social roles of games. Ludonotation is to visualize players' gaming strategies hidden from the interactive gaming behaviors that assist us to understand, respect, communicate, and work with others, either in or outside of the game.This paper presents Ludonotation@ISG and its functions with the example of the analysis of the house-made strategic game <Fragrance Channel v2.0>.

2. Literature review

2.1 Interactive Strategic Games

Strategies can be defined in various ways depending on the genres of games. To avoid being distracted to the long list of game genres, this paper places its focus on the interactive strategic games (ISG) that has fundamental features in varying degree: a) Players progress based on strategic decisions, not luck; b) Players have equal knowledge to play; c) Play is based on multiple decisions a person could make on each turn with possible advantages and disadvantages each time; d) Players can plan strategies that will take multiple turns to complete fully; e) Players can replay the game many times and have a different experience each time; f) Winning and losing is specific and achievable.

The ISG designed in this study, <Fragrance Channel v2.0>, has a couple more features that defines it as a scenario-based ISG (e.g. Covaci, Ghinea, Lin, Huang, & Shih, 2018; Shih, Huang, Lin, & Tseng, 2017): a) A thematic scenario and narrative that encompass certain degree of conflicts. The scenario often comes along with narratives and conflicts so that players are engaged in the game environment. Conflicts are the source of competition in which players compete over limited options, choices, and resources. b) Multiplayers that take unequal roles in the game. The inequality motivates players to use strategies to win the game. c) The round-based game with certain degrees of free-talk game sections. This free-talk section is designed to encourage social interactions that allow players to negotiate for reaching common goals or achieving mutual agreements. d) Interactions that are generated from the free-talk section for negotiation or so forth. The negotiation also affects players' strategic planning and decision-making. e) Open ending, in which the victory is defined by principles instead of the achievement of any single player. Game outcome answers the objective of game that fit the theme and solve conflicts of the triggering events. Games can end in a variety of ways by achieving a stated condition described by the designer. These features bring ISG its complexity and adhesiveness.

In ISG, players use strategies as their decisions and choices in the game where the outcome depends not only on their own actions but on the actions of others (Kugler, Kausel, & Kocher, 2012). ISG are often dynamic that do not rely on luck but decision-making skills in determining the course and outcome of the game. These decisions do not occur by accident as game designers create games purportedly to generate specific actions, behaviors, and strategies.

2.2 Behavior Analysis

Game-based learning (GBL) involves two types of behaviors: learning behaviors and gaming behaviors. Whereas learning behaviors are those actions outside of the game but using game as learning medium, gaming behaviors are those actions happen inside of the game according to the mechanisms. Learning behaviors are done for the learning goals, and gaming behaviors are for gaming tasks. Learning is for improvement and gaming is for winning. Since the goals and behaviors are different between learning and gaming, researchers should make clearer definitions to the two.

Most GBL behavior analysis are about individual behaviors in the whole process with or without peers (e.g. Huang, Cheng, Huang, & Teng, 2018; Prada, & Paiva, 2009; Worth, & Book, 2014); seeing group behaviors as a whole in terms of learning styles, personalities, learning achievements, etc. (e.g. Buckley, & Doyle, 2017). So far, there is no study that looks into individual player's interactive behaviors in the group, nor generate identifiable social interactive behavior patterns of individuals. Also, most studies in GBL focus on analyzing learning behaviors with the game but not gaming behaviors in the game (e.g. Yang, Chang, Hwang, & Zou, 2020); and some studies mix learning and gaming behaviors in their coding system (e.g. Hou & Keng, 2021).

Other than looking at probabilities, frequencies, ratios, factors, we are looking for behavior patterns. Behavior documentation should be done in multimodal forms, such as text, audio, video, as well as system logs. Multimodal analytics would also involve facial expression recognition, body gesture recognition, social distance and space movements, and so forth (e.g. Spikol, 2017).

It also requires a visualization tool so that the complex interrelationship of player interactions can be seen and transmitted. To understand the relationships between individual behaviors, Lag Sequential Analysis (LSA) can be used (e.g. Hou, 2012); yet only an extended version of Time Sequence Analysis (TSA) can show a series of intersected actions done by several players, and the chain of every individual's actions in every round of the game. The major difference is that the later shows the real-time stamps of every event instead of a mined data calculation. There is only one study mentioning game behavior strategies (e.g. Azhar, Mclennan, & Reif, 2005).

3. Research Methods

3.1 Game Design

This research develops strategic complex board game <Fragrance Channel v2.0> in which the players role-play European countries, namely England, Netherland, Spain, and Portugal, each completes respective spice trade missions in the Great Voyage time. The players have to obtain spices from not

only their own colonies but also others which cause cooperation (trades) and competition (battles) to happen. Smart phone is integrated in the physical board game which extends the functions of the game mechanisms including documenting and exchanging gaming parameters and values. The database documented every player's gaming behaviors which can be used for consecutive time sequence behavior analysis. In this dynamic game, players have a lot of interpersonal interactions and strategic planning. Other than moving their ships to designated locations, they manage to find out the best approach to complete their missions faster than other players and win the game.

3.2 Research Design

This research documented 8 games for about two hours each with four players in a game. Students were identified with the personality trait using PDF personality scale. In order to more clearly describe the following discussion, animal nicknames are used to refer to the personality traits. Nevertheless, students were too young to possess dominance (Tiger) trait. Therefore, Tiger trait is neither included in the analysis nor the discussion in this study.

Document codes were inscribed as below for narrative purposes and served as evidences for analysis. For example, 814-M-01 refers to game session of "date -time (morning/afternoon)-session number"; and P1-B-Peacock refers to "player number-gender(boy/girl)-personality".

After every game, 20-minute focus group was conducted with every gaming group to review the gaming process together. All gaming behaviors and dialogues were transcribed and coded for analysis. Two categories of codes were used. Gaming Behavior codes (GB) are to document the players' extrinsic gaming movements that are observable. Gaming movements in <Fragrance Channel v2.0> were transcribed into ten codes. Movements with higher code values mean more aggressive game behaviors, such as: 10-attack, 9-upgrade weapon; reversely, lower code values mean more conservative game behaviors, such as: 5-inbound without spices, 2-trade, 1-maintanence, etc. Code 6-move placed in the middle of the chart is simply neutral in terms of gaming attitudes.

Simultaneously, and Human Interaction codes (HI) are to document players' dialogues between each other to identify their intrinsic gaming strategies that are not evident from observations. Focus groups are performed with the players to review the process together to avoid over-speculation of the researchers. Since the scenario of <Fragrance Channel v2.0> embeds conflict-resolution and problem-solving process, this research adopts Spivack and Shure's (1974) Preschool Interpersonal Problem-Solving Test (PIPS) about social problem-solving strategies to generate the human interaction data. Interactions that would happen in the games were listed and categorized into six interactive strategies including lie, prosocial strategies, neutral, controversial strategies, affective, and change goals. Higher code values are more pro-social to other players, such as 12-temptation, 11-lie, 10-ask for consent, 9-share and give advice; lower code values are more anti-social, such as 6-self-centered, 5-threaten, 4-criticize, 3-plead for pity. In the middle is 7-neutral communication.

4. Ludonotation

The game research analysis representations that are known today mostly exhibit only a piece of the information, a variable that influences the whole study, or a route that one has taken. It hardly takes a comprehensive view of a gaming scene, especially when the subject of the study is complex, dynamic, and interconnected. There is a need to generate a new tool for this specific purpose which can more perfectly capture the milieu.

Just like Labanotation as a dance equivalent to music notation, Ludonotation is a notation system meant to serve the recording and analysis in game studies. In order to allow researchers to visualize the whole detailed game process, just like the music notes shown in Grand Staff that goes on and on until the end of the music, Ludonotation is also an incessant representation that documents and shows the complete game process for the total game time (also see Figure 1)

In the notation, it is suggested to comply around 13 codes to connote notes like those in the Grand Staff, which can be topped up to two coding systems like the upper and lower staff. With needs to present several parties or groups working in parallel, a symphony score can be useful just as several music instruments play together in a piece of music. The coordination of the four players becomes an

ensemble and the interactions would be composed into distinct music. It is supposed that certain behavior patterns can be read as particular melodies that can be quickly inferred in Ludonotation.

Ludonotation needs several elements in the notation system to document sufficient information identified by respective symbols (also see Figure 1). The five fundamental requirements include: players, player features, player behaviors in codes, player behavior direction to its target player, and flow sequence of behaviors in regard of time. Players' actions and interactions are documented on and between the lines according to the values of the codes. The frequencies of each action can be noted with the time signatures. Players are identified by colors, and personality traits, or any other players' features, are identified by geometric shapes. Black lines pointing to four directions identified by quadrants symbolizing the target players each behavior aiming to. It opens up the possibilities to illustrate the social space created in the game. For adaptation of Ludonotation for other game genres, maybe different symbols would be used to suffice its nature and needs.

Lag Sequential Analysis is often used to identify behavior patterns according to the frequencies between any two behavioral movements. However, in the strategic game, behavior patterns appear in a time span normally involve two movements and up of one player or between players across rounds (Shih, Chiu, & Lin, 2022). The string of single player's movements composes gaming strategies that might dynamically change in accordance to the movements of other players that occur in between the plotted string. In other words, streaming linkage between single players' cross-round behaviors and causal-effect relationships between players' behaviors are beyond what sequential analysis can capture. Other than asking the probability between behaviors, it is more important to recognize the dynamic matrix across time.

Therefore, an example of the use of Ludonotation@ISG was presented with <Fragrance Channel v2.0> that gives a comprehensive view of the game that strengthens the need of displaying in parallel the gaming behavior chart and the human interaction chart to see the flow of games over time.

5. Research Results

The research analysis results will be presented with the excerpts from games in the experiment to showcase what players with different personality traits do, how they influence each other, and what their gaming strategies are by observing and analyzing their gaming behaviors and human interactions. The two coding results were displayed in a Ludonotation which is a lengthy time sequence chart so the researchers can visualize the game process. The following section chooses one game excerpt to illustrate: a) the gaming behavior patterns of the players with different personality traits; b) the gaming stratagems used in the strategic game; c) the use of Ludonotation and its functions. Various stratagems were found in the excerpts (Table 1).

Table 1. *Stratagems used by players in <Fragrance Channel v2.0> Game 815-A-02*

Players	Chapter	Stratagems
P1-B-Chameleon (P1)	I-1 II-4 III-5 VI-4	Deceive the heavens to cross the sea. Hide a knife behind a smile. Tossing out a brick to get a jade gem. Inflict injury on oneself to win the enemy's trust.
P2-B-Chameleon (P2)	I-3	Kill with a borrowed knife.
P3-B-Owl (P3)	I-4 VI-6	Wait at leisure while the enemy labors. If all else fails, retreat.
P4-B-Koala (P4)	I-3 III-5	Kill with a borrowed knife. Tossing out a brick to get a jade gem.

Figure 1 is the excerpt of 815-A-02. In the beginning of this game, P1, P2, P4 made coalition in the beginning and use attack to exchange spices. P1 and P4 has a lot of conversations on lending (HI8), sharing and giving advice (HI9), and asking for consent (HI10) about exchange spices as a gesture of "tossing out a brick to get a jade gem" (Stratagem III-5) and "hiding a knife behind a smile" (Stratagem II-4). Since P3 played as an independent country, that he "Wait at leisure while the enemy labors" (Stratagem I-4). He made sure he ended in the port (GB5, GB7) in every round to avoid being attacked by using "if all else fails, retreat" (Stratagem VI-6) even though these behaviors cost him action points.

However, in Round 13, P1 broke his contract with others and move (GB6) to start point after he obtained his spices. It is the act of "deceiving the heavens to cross the sea" (Stratagem I-1) by lying to others to achieve his own goal. P2 criticized (HI4) him, and agitatedly reminded P4 of P1's betrayal; this is when he used "kill with a borrowed knife" (Stratagem I-3). Then P4 moved (GB6) to attack (GB10) P1 by "tossing out a brick to get a jade gem" (Stratagem III-5). P1 was good at communication and altered his strategies for his own benefits according to the conditions due to his personality traits.

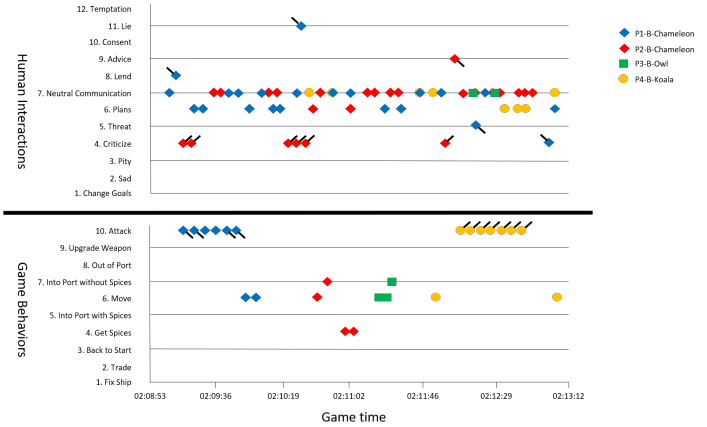


Figure 1. Ludonotation stave for strategic game excerpt 815-A-02 Round 13

After the analysis, preliminary conclusions are made to identify some behavior chains to be possible gaming stratagems (Table 2).

Table 2. *Behavior chains of possible stratagems used in the game*

Stratagems	Behavior chains	Behavior codes
I-1	Pro-social — Lie/temptation — Anti-social	HI8/9/10 — HI11/12 — HI4/5
I-3	Pro-social — (others) attack	HI9/10/11/12 — HI4/I5; GB10
I-4	No anti-social interactions.	No HI2/3/4/5
I-6	Talk to A — Attack B	HI4/5 (to A) — GB10 (to B)
II-3	(Others attack each other) — (No pro-social or anti-social actions)	(others) GB10 — (no) HI4/5/8/9/10
II-4	Lie/temptation — Anti-social	HI11/12 — HI4/5
III-5	Pro-social/lure — (others) Pro-social	HI8/9/10/11/12 — (others) HI8/9/10
VI-4	(others) Anti-social — Sad/Pity	(others) HI4/5; GB10 — HI2/3
VI-6	(others) Anti-social — Avoid/Hide	(others) HI4/5 — HI1; GB3/5/6/7

In order to be able to visualize the possible pattern of the behaviors, they are noted in Ludonotation. Figure 2 are the nine stratagems that are found in <Fragrance Channel v2.0>. One can see that behaviors done by one player, though with other behaviors in between the stratagem-oriented behaviors, can be identified as using the stratagem. It is to certain degree that the stratagems might be overlapped, abandoned, ignored, or with variants, that have to be acknowledged and identified by players. What's tricky and difficult to note is those stratagems that specify the absence of behaviors, such as "no lying", "no attacking", "no trading", etc.

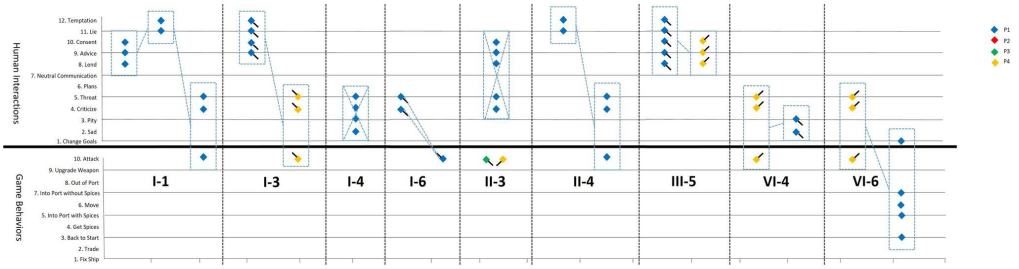


Figure 2. Behavior Chains of Stratagems on Ludonotation

6. Conclusion

Ludonotation is a means for the documentation of a game, an adjunct to video, and as a tool for game analysis. The analysis of the gaming process can help to diagnose the behavior patterns of specific personality traits; providing possible advices can be valuable and major contribution of this study.

The prospective development of Ludonotation shall continue to grow into a more sophisticated version that can serve more variation of strategy games, or even other genres of games. Such a notation system shall have flexibility for users to add, select, or modify for specific purposes. With the assistance of computer, digitized notation system using automated detection function with encoding module, annotation and visualization functions as well as big data mining ability will all contribute to the enhancement of the elicitation of behavior patterns.

Using Ludonotation is to conduct game analysis with anthropological perspective. The understanding to the complete gaming process, including the players, objects, mechanisms, and intentions, is the comprehension to the small world. Such system can be beneficial for education, and future studies of games and human behaviors.

Acknowledgements

This study is supported in part by the Ministry of Science and Technology, Taiwan, under MOST-104-2628-S-008-002-MY, MOST-108-2511-H-008-016-MY4.

References

Azhar, S., Mclennan, A., & Reif, J. H. (2005). Computation of Equilibria in Noncooperative Games. *Computers and Mathematics with Applications, 50*(2005), 823-854.

Buckley, P., & Doyle, E. (2017). Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market. *Computers & Education, 106*(2017), 43e55.

Covaci, A., Ghinea, G., Lin, C.-H., Huang, S.-H., & Shih, J.-L. (2018). Multisensory games-based learning - lessons learnt from olfactory enhancement of a digital board game. *Multimedia Tools and Applications, 77*(3), 21245–21263.

Hou, H. T. & Keng, S. H. (2021). A Dual-scaffolding framework integrating peer-scaffolding and cognitive-scaffolding for an augmented reality-based educational board game: An analysis of learners' collective flow state and collaborative learning behavioral patterns. *Journal of Educational Computing Research, 59*(3), 547-573.

Hou, H. T. (2012). Exploring the behavioral patterns of learners in an educational massively multiple online role-playing game (MMORPG). *Computers and Education, 58*(4), 1225-1233.

Huang, H. -C., Cheng, T. C. E., Huang, W. -F., & Teng, V. -I. (2018). Impact of online gamers' personality traits on interdependence, network convergence, and continuance intention: Perspective of social exchange theory. *International Journal of Information Management, 38*(2018), 232-242.

Kugler, T, Kausel, E. E., & Kocher, M. G. (2012). Are groups more rational than individuals? A review of interactive decision making in groups. *WIREs Cognitive Science, 3*(4), 471-482.

Prada, R., & Paiva, A. (2009). Teaming up humans with autonomous synthetic characters. *Artificial Intelligence, 173*(2009), 80-103.

Shih, J.-L, Chiu, M.-M, & Lin, C.-H. (2022). Personalities, Sequences of Strategies and Actions, and Game Attacks: A Statistical Discourse Analysis of Strategic Board Game Play. *Computers in Human Behavior, 133,* 107271.

Shih, J.-L., Huang, S.-H., Lin, C.-H., & Tseng, C.-C. (2017). STEAMing the Ships for the Great Voyage: Design and Evaluation of a Technology-integrated Maker Game. *Interaction Design and Architecture(s) Journal (IxD&A), 34*, 61-87.

Spikol, D. (2017). Using Multimodal Learning Analytics to Identify Aspects of Collaboration in Project-Based Learning. CSCL, (June), 263-270. https://doi.org/10.22318/cscl2017.37

Worth, N. C. & Book, A. S. (2014). Personality and behavior in a massively multiplayer online role-playing game. *Computers in Human Behavior, 38*(2014), 322-330.

Yang, Q. F., Chang, S. C. Hwang, G. J., & Zou, D. (2020). Balancing cognitive complexity and gaming level: effects of a cognitive complexity-based competition game on EFL students' English vocabulary learning performance, anxiety and behaviors. *Computers & Education,148(2020)103808, 1-21*.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Design a Robot as a Double with Micro-expression to Participate in a Virtual Situational Learning Environment and its Effect on Students' Learning Performance

Vando Gusti Al HAKIM[a], Su-Hang YANG[b], Jen-Hang WANG[c*], Yi-Jing LI[a], Yi-Hsin CHEN[a] & Gwo-Dong CHEN[a]

[a]*Department of Computer Science and Information Engineering, National Central University, Taiwan*
[b]*Department of Hospitality Management, Chien Hsin University of Science and Technology, Taiwan*
[c]*Research Center for Science and Technology for Learning, National Central University, Taiwan*
*harry@cl.ncu.edu.tw

Abstract: One way to encourage students to keep on practicing and studying learning materials is to let them be aware of how they will perform if they are placed in the situational evaluation. Thus, there should be a mechanism to act as a double that can be trained and revised by students so they can watch, monitor, and learn their own performances. Besides, emotional intelligence has gained a lot of attention due to its potential in maintaining situational interactions. Such a skill can detect the micro-expression to grasp the true feelings and hence students can take the appropriate responses. Therefore, this study proposes a learning approach, where students can collaboratively design the robot as their double with micro-expression to participate in situational evaluation tests. To evaluate the effectiveness of the proposed approach, a quasi-experiment was conducted with 90 undergraduate students who enrolled in a Hospitality Japanese course in Taiwan. Three classes were assigned randomly with different learning approaches. The experiment results indicated that the groups with a robot as a double showed significant positive effects in terms of learning motivation and anxiety. Furthermore, a group of students using a double robot with micro-expression exhibited better learning achievement, but their anxiety was also increased. Finally, the limitations of the study and some suggestions for improvement of this proposed approach are provided.

Keywords: Robots for Learning, Virtual Situational Learning Environment, Double, Emotional Intelligence, Micro-expression, Human-Robot Interaction, AI and Robots in Education, Situated Learning, Collaborative Learning, Scaffolding Tool

1. Introduction

One way to encourage students to keep on practicing and studying learning materials is to let them be aware of how they will perform if they are placed in a situational evaluation (Al Hakim, Yang, Liyanawatta, Wang, & Chen, 2022). The situational evaluation here refers to the process of judging actors based on their interaction with other actors, simulated agents, or companions within a specified time during present in a virtual situational learning environment (VSLE). For instance, when students are playing roles as magicians, they have to cast the spell properly during the evaluation so that the scenario can proceed to the next stage of the journey (Al Hakim, Yang, Wang, Yen, Yeh, & Chen, 2021). Therefore, it is necessary to provide a test mechanism to let students watch, know, and learn their own performances that can be designed and revised by themselves. By doing so, students can reflect on their current performance and keep on improving until passing the situational evaluation tests.

In filmmaking, a double is a being who represents the original actor of the film (Piotrowska, 2013). It can create the illusion that makes the spectators or audiences believe that it is their favorite actor who performs the show (Smith, 2004). Thus, the double can be employed as a scaffolding tool that can be trained by students to show their learning outcomes and make the double behave like themselves. On the other hand, emotional intelligence ability has gained a lot of attention due to its potential to guide thinking and behavior, distinguish between different expressions and label them appropriately, and

adjust emotions to adapt to environments (Colman, 2015). Such an ability can detect the micro-expression, i.e., involuntary emotional leakage that happens within a fraction of a second and exposes the true emotions of people (Ekman, 2009; Ekman & Friesen, 1969). It is important to grasp micro-expression in particular situations and hence people can take appropriate action to respond to the interlocutors (Soltani, Zarzour, & Babahenini, 2018).

Therefore, this study aims to create a learning environment, where students can collaboratively train or design the double with micro-expression to participate in situational evaluation tests. Moreover, such a learning environment allows students in small groups to immerse themselves with a double to present in a virtual scenario with situational evaluation. Because students show their presentations in front of the audience, they might be driven to revise their own presentation until it is appropriate to the desired design (Chen, Fan, Chang, Chang, & Chen, 2018). The motivation for learning is important because it plays a critical role in the improvement of learning performance (Murphy & Alexander, 2000). Besides, adequate practice in training or designing the double might decrease the anxiety level of students during presentations, which becomes an important issue in a situational learning environment (Liu, Lin, Wu, Chen, & Chen, 2017). To sum up, evaluating the effectiveness of the proposed learning environment to enhance students' learning performance in terms of learning achievement, motivation, and anxiety is worth investigating.

2. Literature Review

2.1 Emotional Intelligence in VSLE

Emotional intelligence is "*the ability to monitor one's own and others' feelings and emotions, to discriminate among them, and to use this information to guide one's thinking and actions*" (Salovey & Mayer, 1990, p. 189). With respect to learning, Perera and DiGiacomo (2013) claimed that emotional intelligence is a good predictor of students' academic performance. Emotional intelligence allows students to have better adaptability to new situations and research has demonstrated a positive relationship between emotional intelligence and achievement of students in a dynamic workplace that requires interactions (Cleary, Visentin, West, Lopez, & Kornhaber, 2018).

When stimulated, human faces will show minor subconscious changes without the control of thoughts within around 0.5 seconds, which are recognized as "micro-expressions" (Zhou, Song, Chen, Chen, Ben, & Cao, 2022). This concept of micro-expression was first discovered by Haggard and Isaacs (1966) and viewed as cues for repressed emotions. Ekman and Friesen (1969) defined micro-expressions as a brief facial movement revealing a true emotion that a person tries to conceal. Micro-expressions are even claimed that as the most promising approach to detect deception, as they usually occur in high-stake situations (Ekman, 2009). With these promising features, a study about micro-expression is worth investigating.

Recent advances in the fields of digital reality have provided new opportunities to let a robot and students become actors and have an authentic experience by immersing themselves in a VSLE with immediate evaluation (Al Hakim et al., 2021; Al Hakim, Yang, Liyanawatta, Wang, Ku, Zhuang, & Chen, 2022). In an authentic environment, students are usually positioned as actors in a particular scenario with learning content and context to interact and solve contextual problems (Barab, Gresalfi, & Ingram-Goble, 2010). They might be role-played as employees in a simulated workplace to emotionally interact with simulated agents and make required decisions that reflect real cases to satisfy the agents (Ke, Moon, & Sokolikj, 2020; Chang, Kao, Hwang, & Lin, 2020). Consequently, emotional intelligence's ability to grasp the real feelings of people through micro-expression is essential to maintaining situational interactions.

2.2 Robots as a Scaffolding Tool to Participate in Situational Evaluation Tests

Currently, a large number of learning activities using robots occur inside the classroom. Robots can play many important roles in learning, including as scaffolding tools to support collaborative learning processes (Okita & Clarke, 2021). Collaborative group strategies can promote individual, group, and community cognitions (Zheng, Yang, Cheng, & Huang, 2014). The robots have characteristics that can

be useful for learning, such as flexibility, repeatability, ability to present data, interactivity, appearance, and movement to support instructional goals (Chang, Lee, Chao, Wang, & Chen, 2010).

Atmatzidou and Demetriadis (2012) reported that the students had a positive attitude towards the collaborative learning activities with the LEGO robot, in which they felt that collaborative works become better and interesting, and made high-achievements responsibly to transferring knowledge (as experts) to other groups members, and hence enhanced students' learning motivation. Moreover, Al Hakim et al. (2020) claimed that designing the robot's emotion expressions (sound and facial expression), movements, and speaking features based on the students' preferences can increase their learning interest and concentration. Therefore, a robot can be used as a double that can be designed and revised by the students to represent their learning outcomes (Piotrowska, 2013; Smith, 2004) until the double passes the situational evaluation tests.

Considering the benefits of the robot as a scaffolding tool, this study proposes a learning approach to investigate its impact on the learning performances of students in a VSLE. This study proposes the following hypothesis: learning through design a robot as a double with micro-expression to participate in a virtual situational learning environment can improve students' learning performance. In general, there are two stages in VSLE, one is designing stage and the other is performing stage (Al Hakim, Yang, Liyanawatta, Wang, & Chen, 2022). Initially, before performing in VSLE, all groups with different learning approaches need to construct the scenarios and interactions based on textbook context and content in the designing stage. Three learning approaches are compared, and their effects on students are assessed. The first approach is a micro-expression double robot, wherein a group of students can train or design the robot as their double with micro-expression and practice situational evaluation tests. The second approach involves a double robot as in the first approach but without the micro-expression. The last one uses a conventional instruction approach, wherein a group of students practice situational evaluation tests with their classmates. Moreover, an experiment is conducted in the course of Hospitality Japanese for university students to examine the following questions.

1. Do students learning with the micro-expression double robot approach show better learning achievements than those with a double robot approach or conventional instruction?
2. Do students learning with the micro-expression double robot approach show higher learning motivation than those with a double robot approach or conventional instruction?
3. Do students learning with the micro-expression double robot approach show lower anxiety of learning than those with a double robot approach or conventional instruction?

3. Development of the Micro-expression Double Robot Approach

Fig. 1 illustrates the system architecture. The controller for the whole system, developed on the Android platform, was installed on tablet control. The VSLE program was installed on a personal computer with an NVIDIA graphics card and running on Windows 10. Such a program was used to capture the images of the actors and immerse them into a virtual situational learning scenario, where immediate evaluation and interactive elements were included (e.g., sound effects, the virtual scoring system, and the dynamic scenario). The aforementioned system used WampServer v2.4.9 as the application server and MySQL v5.6.17 as the database server to store all the information.

ZENBO robot from ASUS was employed in this study. The Android-based robot had one screen and two wheels with various LED colors to support emotional actions (i.e., facial expressions, wheel lights, and action). It was 62 cm tall with a round body and had a multimedia deployment feature as well as Wi-Fi connectivity. Furthermore, Microsoft Azure Speech Cognitive Services and DialogFlow services were employed in the robot to manage the speech-to-text and conversation capability, respectively. The robot, tablet control, VSLE program, and database were connected through socket communication.

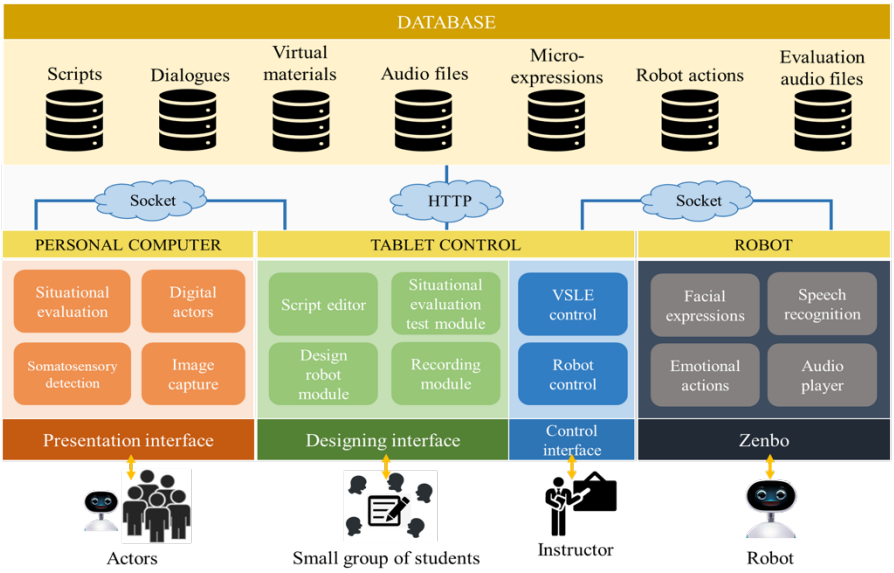


Figure 1. Architecture of the system.

In this study, a micro-expression double robot approach was developed by referring to the interactive situated learning approach with robots proposed by Al Hakim, Yang, Liyanawatta, Wang, & Chen (2022) and considering ICAP (interactive, constructive, active, passive) framework proposed by Chi and Wylie (2014) (see Fig. 2).

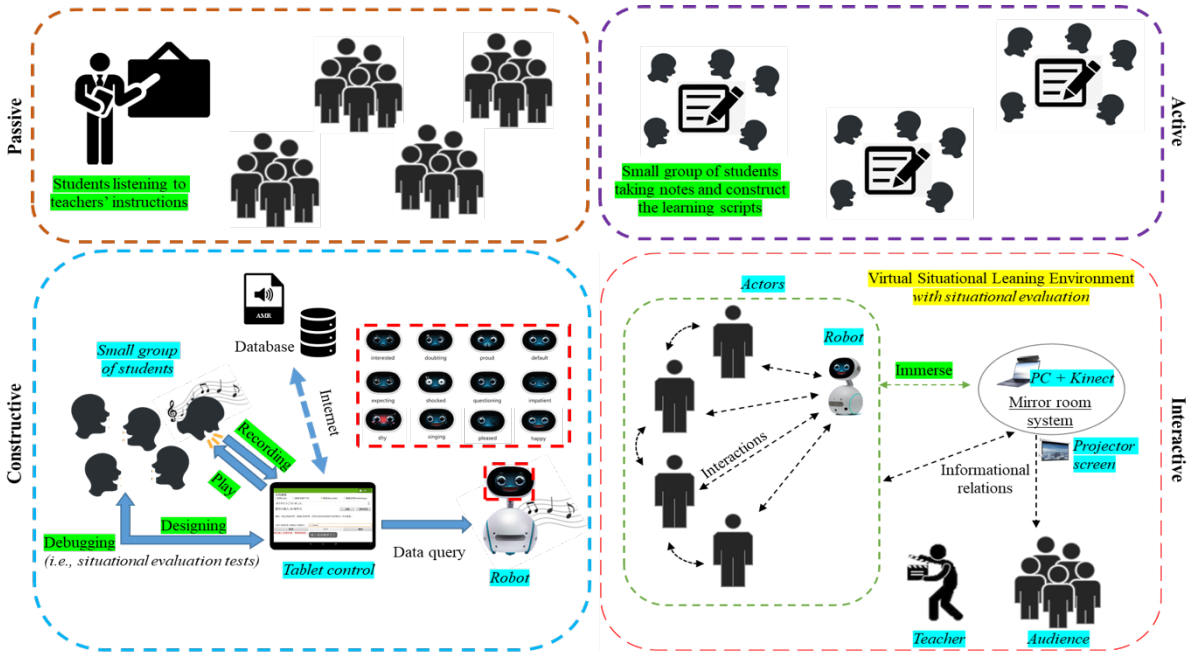


Figure 2. Micro-expression double robot approach.

In the beginning, the schoolteacher gave instruction about the course objectives and taught the content of the textbook (i.e., reserving a seat, ordering meals, and payment in Japanese restaurants), and explained the meaning of each sentence. The schoolteacher also taught students about the emotional interactions in particular situations based on micro-expressions defined in Ekman (1999), helped students in practicing the learning materials, and assisted them to learn pronunciation correctly. Afterward, a description of testing or debugging the situational evaluation tests using a micro-expression double robot was demonstrated, including the tutorial of presenting in VSLE.

Next, students were divided into small groups (5 to 7 members in each group) to construct learning scripts. The format of learning scripts proposed by Chen, et al. (2018) was adopted and modified by researchers and schoolteachers to guide a small group of students while designing the interaction

of particular scenarios based on textbook content and contexts. The students designed their presentation with a double and themselves. In this study, the students and a robot were positioned as waitresses and customers in a Japanese restaurant, respectively. The group of students discussed and pondered what action or dialogue that might be performed by the robot as a customer with specific requests in real-life situations. At the same time during scriptwriting process, they also needed to consider the scenario and interaction based on context of robot's emotional responses.

After the learning script was completed, the students can collaboratively work to construct the double robot's micro-expressions and actions using Designing Interface on tablet (Fig. 3 left). Students design or train the robot to perform for them and make the robot behave like themselves, so they can reflect and revise their own presentation. Such an interface also provides a tutorial on how to pronounce a sentence and be able to record voice, replay, and validate whether it can pass the situational evaluation tests (Fig. 3 right). Subsequently, after the learning script was completely digitized into the system, students can debug and practice their designed script with the robot before presenting it in the VSLE.

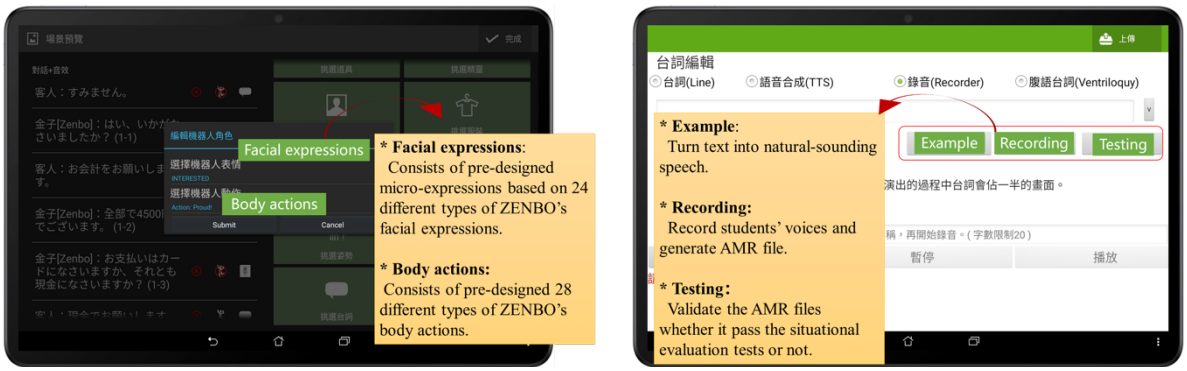


Figure 3. Designing interface in tablet control.

After finishing the designing stage, all of small groups take turns to rehearsal and present in VSLE with their designed micro-expression double robot, while the other classmates as the audience can watch the actors' performances. In this stage, the teacher situationally controlled the flow of the designed script through the Control Interface on the tablet. Moreover, this interface managed which subtitles were displayed on the screen to be read by the actors. During presentation, the robot is positioned as an interlocutor who is role-playing as a restaurant customer that replays and deploys the emotions, actions, and voices of students based on the designed script (see Fig. 4 left) but in random ways as real-life situations. For instance, the robot requests a beefsteak with a specific cook level, a cup of green tea, and a pudding as a meal, beverage, and dessert, respectively. Fig. 4 (right) shows that when the students correctly performed the interactions, the robot provided positive, happy feedback and praised them. The VSLE then changed the scenario and increased the score. Conversely, when student actors failed in the interaction, the robot provided unhappy feedback and the VSLE decreased the score.

ID	Actor	Conversation	Next ID
1	Robot	Please give me a beefsteak.	2
		Please give me an Indian curry rice.	6
		Please give me a belly pork satay.	7
2	Student	How about the cook-level of the beefsteak?	3
3	Robot	Rawly cooked.	4
		Fully cooked.	5
4	Student	Understood, a mediumly cooked beefsteak. What do you want for the drink?	8
5	Student	Understood, a fully cooked beefsteak. What do you want for the drink?	8
6	Student	Understood, an Indian curry rice. What do you want for the drink?	8
7	Student	Understood, a belly pork satay. What do you want for the drink?	8
8	Robot	Please give me orange juice.	9
		Please give me a cup of green tea.	10
9	Student	Understood, a cup of orange juice. What do you want for dessert?	11
10	Student	Understood, a cup of green tea. What do you want for dessert?	11
11	Robot	Please give me cupcake.	12
		Please give me pudding.	13
12	Student	Understood, a cupcake. Please wait a moment.	
13	Student	Understood, a pudding. Please wait a moment.	

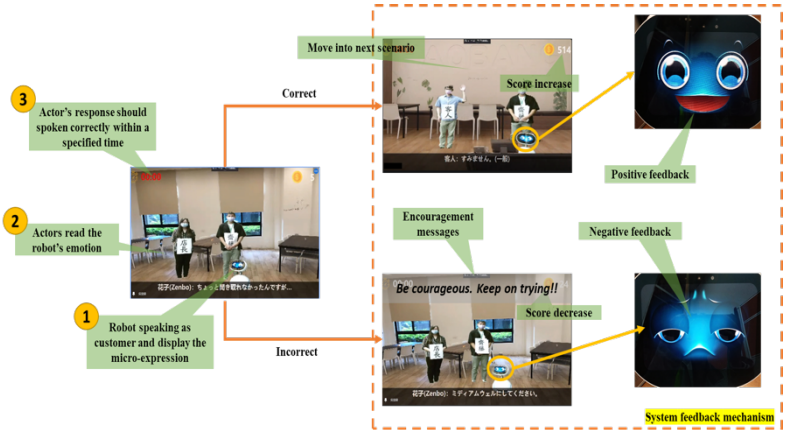


Figure 4. Designed script and illustration of interaction.

.

4. Experiment

4.1 Participants and Learning Materials

The experiment was conducted in a course of Hospitality Japanese at a university in Taiwan. Three classes were randomly selected as the experimental A, B, and control groups, respectively. All of the students in groups were around 21 years old and taught by the same instructor. Before performing in VSLE, all groups were further randomly divided into smaller groups (five to seven students in a small group) and went to the designing stage with different learning approaches to construct the scripts.

Experimental group A containing 30 students (18 females, 12 males) used the micro-expression double robot. Another class with 30 students (16 females, 14 males) selected as the experimental group B learned with a double robot approach. The control group consisting of 30 students (19 females, 11 males) used the conventional instruction approach. Details of the experimental A, B, and the control groups are shown in Table 1, respectively.

Table 1. *Feature of learning approaches in the study.*

Group	Design the learning scripts	Construct a robot as a double	Construct a double with micro-expression
Experimental group A	✓	✓	✓
Experimental group B	✓	✓	
Control group	✓		

This study compiled the materials taken from the Hospitality Japanese course textbook into a learning script. All three groups used the same learning script. The learning script was developed by the schoolteacher with more than 10 years' experience in teaching the subject. Additionally, the learning script plots was based on the Transformational Play proposed by Barab et al. (2010) to make the developed learning script convey the narrative structures and character development.

4.2 Procedure and Measurement Tools

In this study, experiment was conducted for two months. A pre-test with total scores ranging from 0 to 100 points developed by two schoolteachers was administrated at the beginning of the experiment to evaluate the prior knowledge level of students, along with pre-questionnaires to evaluate their prior perceptions in terms of learning motivation and anxiety. The post-test questions were similar to but different from the pre-test questions to keep the consistency of test. The pre- and post-test questions were designed to measure learning achievement in terms of sentence translation, pitch accent, and situational responses. Regarding the interview, it was conducted in the final stage to obtain the perceptions of students. The experimental procedure was shown in Figure 5.

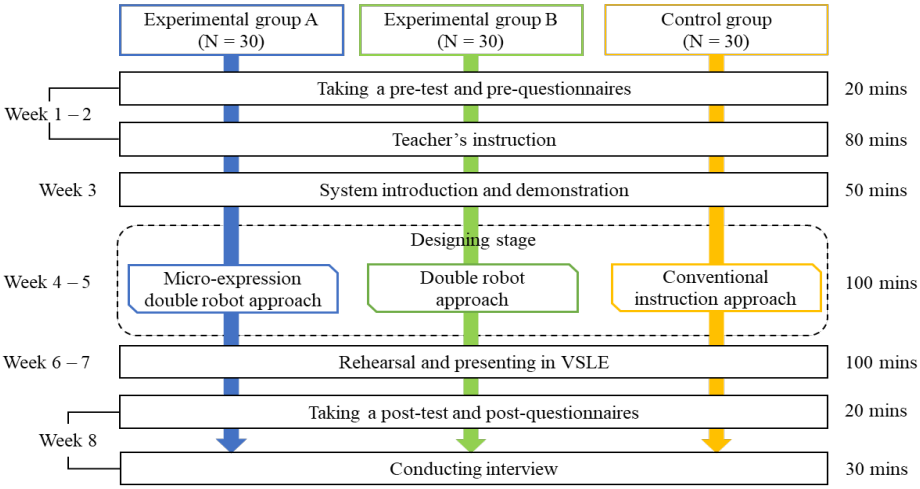


figure 5. Experimental procedure

A single factor covariate analysis (ANCOVA) was employed in this study to evaluate the learning achievements, motivation, and anxiety among the three groups based on the obtained test scores and questionnaire responses. The questionnaires with a five-point Likert scale were used to evaluate the learning motivation and anxiety of students.

The questionnaire adopted to measure the participants' learning motivation was developed based on Wang and Chen (2010). It consists of six items (e.g., "I want to do my best in this class because it is important to show my ability to my classmates, teacher, or the others") and has been reported with a reliability coefficient (Cronbach alpha) of 0.79 (N = 115). With regard to the measurement of the learning anxiety, the foreign language anxiety scale modified by Yang, Chang, Hwang, and Zou (2020) was employed. It consists of eleven questions (e.g., "I feel confident when I speak in a foreign language") with the Cronbach alpha value (0.90) proposed by the original study (N = 51). Moreover, at the end of the experiment, ten students were randomly selected from each group for the interview to have understanding how they perceive the learning approach in detail. The interview questions were based on the study of Al Hakim, Yang, Liyanawatta, Wang, & Chen (2022) and were modified according to the context of this study (e.g., "When using the learning approach, which part did help you study?").

5. Results and Discussion

5.1 Analysis of Learning Achievements

The assumption of the homogeneity regression slopes was met ($F = 2.05$, $p > .05$), implying that the ANCOVA was able to be conducted. The adjusted mean of the experimental A, B, and control groups were 72.34, 68.85, and 68.55, respectively, and a significant difference among the three groups was found ($F = 15.43$, $p < .05$), indicating that the three groups had significant differences in the post-test scores. By applying pairwise comparisons, it was found that experimental group A outperformed all groups. The results mean that the group of students with the robot as their double with micro-expression (experimental group A) had significantly higher learning achievements than those either with the double robot without micro-expression (experimental group B) or the conventional instruction (control group) in VSLE. Additionally, the effect size of the micro-expression double robot approach reach to a large level with partial eta squared (η^2) = .276 (Cohen, 1988).

Students can learn better when the micro-expression were added to the double robot in VSLE. Because the students in the micro-expression double robot group should trained the robot and need to obey the emotional interactions to pass the evaluation, they outperform those in the other groups in terms of learning achievements. They should be able to perceive the micro-expression of the double robot and then have appropriate reaction related to the situation encountered, as stated by the students in the interviews. Based on the study of Perera and DiGiacomo (2013) and the results of our experiment, we believed that grasping the situation encountered through emotional intelligence revealed in micro-expression plays an important role in the successful performance of students. This finding is in line with the studies by Cleary et al. (2018) and Lam & Kirby (2002); they argued that emotional interaction ability had a positive influence on emotional awareness and management of individual cognitive-based performance. Thus, collaboratively designing or training the robots as an actor's double with micro-expression in a situational learning environment can enhanced students' learning achievements.

5.2 Questionnaires Analysis for Learning Motivation and Anxiety

The responses to the questionnaire were analyzed with ANCOVA to compare the students' learning motivation and anxiety in three different learning approaches. For learning motivation, the homogeneity of regression ratings among the three groups was $F = 1.17$ ($p > .05$), revealing the assumption of ANCOVA was fulfilled. The adjusted mean of the post-learning motivation measurement for the experimental group A was 3.83, the experimental group B was 3.68, and the control group was 3.07. The results of ANCOVA demonstrated that there was a significant difference

among the three groups' learning motivation ($F = 6.324$, $p < .05$, $\eta^2 = .135$). The pairwise comparisons showed that experimental groups A and B outperformed the control group, while a significant difference between the experimental group A and B was not exist. The results implies that in the milieu of VSLE, both group of students using a double robot, in spite of with or without micro-expression, had higher learning motivation than the students in the conventional instruction group.

Based on the interviews of the students, it was inferred that the situational evaluation for learning performance provided by the VSLE could drive them to design or train the double robot properly. Because they had to present and complete the learning tasks, they took responsibility for mastering the interaction and revising the double robot until passing the situational evaluation tests. This learning process might increase the frequency of practice and willingness to learn (Atmatzidou & Demetriadis, 2012). Moreover, the curiosity and fantasy brought out by the robot during situational evaluation can enhance students' learning motivation (Chang, Lee, Chao, et al., 2010; Gordon, Breazeal, & Engel, 2015). This could be one of the reasons why experimental groups A and B attained better learning motivation than the control group.

Regarding learning anxiety, several previous studies (Al Hakim, Yang, Liyanawatta, Wang, Ku, et al., 2022; Shiomi, Iio, Kamei, Sharma, & Hagita, 2015), reported that adopting robots in learning activities can decrease students' learning anxiety. The adjusted mean scores of the experimental A, B, and control groups were 3.67, 4.24, and 2.47, respectively. The F-value ($F = 46.351$, $p < .05$, $\eta^2 = .534$) implies that significant differences exist among three groups in terms of learning anxiety with a large effect size. Based on presented pairwise comparisons, it was found that the experimental group A and B outperformed the control group. The results reveal that the students with a double robot, either with or without micro-expression, had lower learning anxiety than those practicing with their peers before performing in VSLE. We believed that adequate practice in the designing stage can enhance students learning experience in VSLE (Al Hakim, Yang, Liyanawatta, Wang, & Chen, 2022; Al Hakim, Yang, Wang, et al., 2021), and hence they became more prepared and less anxious during the performing.

On the other hand, an interesting result was found that the experimental group B significantly outperformed the experimental group A in terms of decreasing learning anxiety. This finding may happen due to the reason that the students' group with the micro-expression double robot was faced with a situation where the students had to deal with the uncertain emotions of their double robot as an interlocutor in VSLE, as revealed in the interview. Although the unexpected responses from the robots is claimed as non-judgemental during interactions (Alemi, Meghdari, & Ghazisaedy, 2015; Chang, Lee, Chao, et al., 2010), in this study, the micro-expression exposed by the double robot in the group A may increase the anxiety of students. This might be attributed to that the students of group A tried to read the unexpected micro-expression of the robots and manage the situation appropriately in order to pass the situational evaluation during the VSLE performance, therefore, increasing their anxiety.

6. Conclusion, Limitations, and Future Works

This study proposed a micro-expression double robot approach, where students can collaboratively design the robot as their double with micro-expression to participate in the situational evaluation. The major role of the robot in this study is to play as a scaffolding tool to construct the student actor's double that can be revised, and at the same time, as the evaluation tool that enables students to reflect based on their design (i.e., micro-expressions, actions, and students' utterances). After finishing constructing the scenario and interaction as well as training the double robot, both of the double robot and students perform in a VSLE with immediate evaluation.

Based on the experimental results, it is concluded that the students collaboratively designing a robot as their double to complete the situational evaluation showed significant positive effects in terms of learning motivation and anxiety. Furthermore, students' learning achievements were significantly improved when micro-expression was employed in the double robot, but it also increased their learning anxiety. Our study provides a novel area of insight into the use of robots in learning, by integrating pedagogical approaches, reality technology, humanistic design, and practices, to create an immersive intelligent learning environment with a robot as a scaffolding tool that can show learning outcomes and foster emotional intelligence.

However, although the results of this study indicated that the proposed approach is effective, there were several limitations that should be addressed. First, comparing different instructional media with conventional instructional approach, which is categorized as a media comparison study (Surry and Ensminger, 2001), has been claimed as a weakly designed research in the field of educational technology and so could have flaws in the research design itself. Therefore, in the future, authors would discussed how three groups in this study overcame the limitation of the media comparison study using alternative research approach, and thus, our research findings would be more significant for the field. Second, this study conducted an experiment on a Hospitality Japanese course for university students in Taiwan, which might not be proper to apply the findings on different ages, cultures, or learning capabilities of students without further experiments. Future study considering gender and personality is worth investigating since such features can affect students' perception of specific learning environment elements (Denden, Tlili, Essalmi, Jemni, Chen, & Burgos, 2021). Third, the measurement tools were determined by two experienced schoolteachers based on their past experience, which did not measure and analyze the speaking ability, cognitive load, and emotional intelligence level of students. In the future, the authors would cooperate with schoolteachers to construct more appropriate measurement tools. Moreover, since this learning activity involves social recognition by presentation, measuring and analyzing the responsibility, ownership, and autonomy of students in small groups are also interesting topics for further study. Finally, we also plan to improve the robot program that can read and record the micro-expression of students. By doing so, students can more easily design the robot and simultaneously allow themselves to practice emotional interaction.

Acknowledgements

We would like to thank all the teachers, students, and researchers who involved in this experiment. This study was supported in part by the Ministry of Science and Technology (MOST) of the Republic of China (R.O.C.) under contract numbers MOST 110-2511-H-008-004-MY3.

References

Al Hakim, V. G., Yang, S. H., Liyanawatta, M., Wang, J. H., & Chen, G. D. (2022). Robots in situated learning classrooms with immediate feedback mechanisms to improve students' learning performance. *Computers & Education, 182,* 104483.

Al Hakim, V. G., Yang, S. H., Liyanawatta, M., Wang, J. H., Ku, Y. H., Zhuang, Y., & Chen, G. D. (2022, July). Robot as a Ventriloquist Doll in a Virtual Situational Learning Environment to Facilitate Learning Through Self-Dialogue. In *2022 IEEE 22nd International Conference on Advanced Learning Technologies.* IEEE.

Al Hakim, V. G., Yang, S. H., Tsai, T. H., Lo, W. H., Wang, J. H., Hsu, T. C., & Chen, G. D. (2020, July). Interactive Robot as Classroom Learning Host to Enhance Audience Participation in Digital Learning Theater. In *2020 IEEE 20th International Conference on Advanced Learning Technologies* (pp. 95-97). IEEE.

Al Hakim, V. G., Yang, S. H., Wang, J. H., Yen, C. C., Yeh, L., & Chen, G. D. (2021, November). Robot with Embodied Interactive Modes as a Companion Actor in Journey of Digital Situational Learning Environment and its Effect on Students' Learning Performance. In *Proceedings of the 29th International Conference on Computers in Education* (ICCE) (pp. 441-450). Asia-Pacific Society for Computers in Education.

Alemi, M., Meghdari, A., & Ghazisaedy, M. (2015). The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics, 7*(4), 523-535.

Atmatzidou, S., & Demetriadis, S. N. (2012, July). Evaluating the role of collaboration scripts as group guiding tools in activities of educational robotics: Conclusions from three case studies. In *2012 IEEE 12th International Conference on Advanced Learning Technologies* (pp. 298-302). IEEE.

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational play: Using games to position person, content, and context. *Educational Researcher, 39*(7), 525-536.

Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society, 13*(2), 13-24.

Chang, C. Y., Kao, C. H., Hwang, G. J., & Lin, F. H. (2020). From experiencing to critical thinking: A contextual game-based learning approach to improving nursing students' performance in Electrocardiogram training. *Educational Technology Research and Development, 68*(3), 1225-1245.

Chen, G. D., Fan, C. Y., Chang, C. K., Chang, Y. H., & Chen, Y. H. (2018). Promoting autonomy and ownership in students studying English using digital comic performance-based learning. *Educational Technology Research and Development, 66*(4), 955-978.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, *49*(4), 219-243.

Cleary, M., Visentin, D., West, S., Lopez, V., & Kornhaber, R. (2018). Promoting emotional intelligence and resilience in undergraduate nursing students: An integrative review. *Nurse Education Today*, *68*, 112-120.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Colman, A. M. (2015). *A dictionary of psychology.* Oxford, UK: Oxford University Press.

Denden, M., Tlili, A., Essalmi, F., Jemni, M., Chen, N. S., & Burgos, D. (2021). Effects of gender and personality differences on students' perception of game design elements in educational gamification. *International Journal of Human-Computer Studies*, *154*, 102674.

Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), *Handbook of cognition and emotion*. Sussex: Wiley.

Ekman, P. (2009). Lie catching and microexpressions. In C. Martin (Ed.), *The philosophy of deception* (pp. 118–133). Oxford: Oxford University Press.

Ekman, P., & Friesen, W. V. (1969). Nonverbal leakage and clues to deception. *Psychiatry, 32*(1), 88-106.

Gordon, G., Breazeal, C., & Engel, S. (2015). March). Can children catch curiosity from a social robot?. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 91–98). New York: ACM.

Haggard, E. A., & Isaacs, F. S. (1966). Micromentary facial expressions as indicators of ego mechanisms in psychotherapy. In C. A. Gottschalk & A. Averback (Eds.), *Methods of research in psychotherapy*. New York: Appleton-Century-Crofts.

Ke, F., Moon, J., & Sokolikj, Z. (2020). Virtual reality–based social skills training for children with autism spectrum disorder. *Journal of Special Education Technology*.

Lam, L. T., & Kirby, S. L. (2002). Is emotional intelligence an advantage? An exploration of the impact of emotional and general intelligence on individual performance. *The Journal of Social Psychology, 142*(1), 133-143.

Liu, Y. T., Lin, S. C., Wu, W. Y., Chen, G. D., & Chen, W. (2017). The digital interactive learning theater in the classroom for drama-based learning. In *Proceedings of the 25th International Conference on Computers in Education* (ICCE) (pp. 784-789). Asia-Pacific Society for Computers in Education.

Murphy, P. K., & Alexander, P. A. (2000). A motivated exploration of motivation terminology. *Contemporary Educational Psychology, 25*(1), 3–53.

Okita, S. Y., & Clarke, S. N. (2021). Robots and Agents to Support Collaborative Learning. In *International Handbook of Computer-Supported Collaborative Learning* (pp. 407-424). Springer, Cham.

Perera, H. N., & DiGiacomo, M. (2013). The relationship of trait emotional intelligence with academic performance: A meta-analytic review. *Learning and Individual Differences, 28*, 20-33.

Piotrowska, A. (2013). The horror of a doppelganger in documentary film. *New Review of Film and Television Studies, 11*(3), 302-313.

Salovey, P., & Mayer, J. D. (1990). Emotional intelligence. *Imagination, Cognition and Personality*, *9*(3), 185-211.

Shiomi, M., Iio, T., Kamei, K., Sharma, C., & Hagita, N. (2015). Effectiveness of social behaviors for autonomous wheelchair robot to support elderly people in Japan. *PloS One, 10*(5), e0128031.

Smith, J. (2004). Seeing double: stunt performance and masculinity. *Journal of Film and Video, 56*(3), 35-53.

Soltani, M., Zarzour, H., & Babahenini, M. C. (2018, March). Facial emotion detection in massive open online courses. In *World Conference on Information Systems and Technologies* (pp. 277-286). Springer, Cham.

Surry, D. W., & Ensminger, D. (2001). What's wrong with media comparison studies?. *Educational Technology*, *41*(4), 32-35.

Wang, L. C., & Chen, M. P. (2010). The effects of game strategy and preference-matching on flow experience and programming performance in game-based learning. *Innovations in Education & Teaching International, 47*(1), 39–52.

Yang, Q. F., Chang, S. C., Hwang, G. J., & Zou, D. (2020). Balancing cognitive complexity and gaming level: Effects of a cognitive complexity-based competition game on EFL students' English vocabulary learning performance, anxiety and behaviors. *Computers & Education, 148*, 103808.

Zheng, L., Yang, J., Cheng, W., & Huang, R. (2014). Emerging approaches for supporting easy, engaged and effective collaborative learning. *Journal of King Saud University-Computer and Information Sciences*, *26*(1), 11-16.

Zhou, Y., Song, Y., Chen, L., Chen, Y., Ben, X., & Cao, Y. (2022). A novel micro-expression detection algorithm based on BERT and 3DCNN. *Image and Vision Computing*, *119*, 104378.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Achievement, Behaviors, and STEM Interest of Frustrated and Bored Learners Using Minecraft

Maricel A. ESCLAMADO[ab*], Maria Mercedes T. RODRIGO[a] & Jenilyn A. CASANO[a]

[a]*Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, Philippines*
[b]*University of Science and Technology of Southern Philippines, Philippines*
*maricel.esclamado@obf.ateneo.edu

Abstract: In this paper, we analyze the relationship between behavioral and achievement outcomes and self-reported frustration and boredom levels of students using the What-If Hypothetical Implementations using Minecraft (WHIMC). We examine in-game data, out-of-game assessment data, self-reported frustration and boredom, and results of the STEM interest questionnaire (SIQ) from 175 Grade 8 learners from a school in the Philippines. We find that learners who are frustrated tend to disengage from the game. Learners who are bored tend to do less well on post-game assessments. While students' STEM interest generally increased, SIQ scores of frustrated and bored learners tended to remain unchanged. Among low-performing learners who did not express frustration or boredom, SIQ scores increased, implying that the absence of these states was beneficial to this category of learners. This study contributes to what is known about how affect relates to outcomes such as test scores, in-game behaviors, and attitudes. It also raises questions about how games can best be used to pave the way towards greater student engagement in STEM.

Keywords: Minecraft, WHIMC, Frustration, Boredom, Behaviors, STEM interest, Philippines.

1. Introduction

Games used for education can act as rich primers for engaged learning and were reported effective in improving learning outcomes (Zhonggen, 2019). A popular type of game used in education is open-ended games. Open-ended games like Minecraft provide players autonomy on how to do the tasks in the learning environment, thus there is no prescribed learning sequence and analysis of behavior becomes significantly important (Käser & Schwartz, 2020). Minecraft is an open-ended, sandbox-type video game first launched in 2009 (Bitner, 2021). It is open-ended in that it gives players freedom in exploring the environment and players can also mine resources and craft or build objects, often in collaboration with other players. With the open-ended nature of these games, it is expected that the players will have positive and negative emotional states. Frustration and boredom are some of the most frequent emotional states in open-ended learning and are associated with reduced learning outcomes (Sabourin & Lester 2014).

Frustration is defined as an emotion that results from interference that prevents achievement of a goal (Rajendran, Iyer, Murthy, Wilson, & Sheard, 2013). Boredom, on the other hand, occurs when individuals experience low arousal and dissatisfaction or disinterest in response to low arousal (Vogel-Walcutt, Fiorella, Carper, & Shaltz, 2012). Previous work has found that frustration, boredom, and other negative emotions are associated with negative learning and behavioral outcomes (Tze et al., 2015; Liu et al., 2013; Sabourin & Lester, 2014). For example, frustrated students tend to have a history of incorrect actions and help requests. Bored students tend to work slowly (Baker et al, 2012), have more idle time which is also an indicator of off-task behavior (Pardos et al., 2014), give up on a task (Andres et al., 2015), or engage in non-learning behaviors such as gaming the system (Baker et al., 2010).

In this paper, we analyze the relationship between frustration and boredom and learner in-game behaviors of students in Minecraft, an open-ended, less-structured game. We also investigate the relationship between learners' self-reported frustration and boredom with their knowledge assessments and STEM interest. This study sought to answer the following research questions:

RQ1: How do the behaviors of the learners who self-reported frustration and boredom differ from those who did not?

RQ2: What is the relationship between learners' self-reported frustration and boredom and their assessment outcomes?

RQ3: How did the STEM interest of the learners who self-reported frustration and boredom change after using WHIMC?

RQ4: How did the STEM interest of high- and low-performing learners who self-reported frustration and boredom change after using WHIMC?

2. WHIMC

What-If Hypothetical Implementations in Minecraft (WHIMC) is a set of simulations that learners can explore in order to learn more about science, technology, engineering, and mathematics (STEM). It is composed of a Rocket Launch Facility on Earth, Lunar Base LeGuin, a Space Station, alternate versions of Earth, exoplanets, and others. The Rocket Launch Facility is modeled after that of NASA. Here, learners can visit mission control and Mars rover test sites as well as talk to simulated NASA scientists. Lunar Base LeGuin is a tutorial level in which learners practice using in-game science tools to measure and record temperature, oxygen, pressure, and wind speed. The Space Station includes an unaltered version of Earth in which learners practice making different kinds of observations. The Space Station is also the jump-off hub from which learners travel to the different worlds.

The alternate versions of Earth present learners with opportunities to observe the planet under altered conditions. Although the worlds are fictional, they are created in consultation with scientists: They accurately depict conditions on Earth under these circumstances. For example, what if Earth had no moon (See Figure 1)? There would be no seasons, days would be shorter, and winds would be stronger. What if Earth was tilted by 90 degrees, that is, with the North Pole pointed towards the sun? There would be very different day/night cycles. What if Earth had a slightly colder sun? Water might only be able to exist in liquid form in a limited strip of green, and this is where we would all be forced to live.



Figure 1. Earth with No Moon

In each of these alternate Earths, learners explore the terrain, describe the environment, report observations about how life on Earth is affected by these circumstances, and possibly create habitats that will enable them to survive. By immersing learners in these activities, WHIMC hopes to generate interest in and excitement for STEM among participating learners.

3. Review of Related Literature

Game-based learning environments have become popular learning tools because of their potential to deliver knowledge while keeping the experience fun for the learners (All et al., 2016). Their ability to strike a balance between teaching/learning and gaming is a characteristic leveraged by educators to

support the learning process (Prensky, as cited in Beserra et al., 2014). Learners are likely to exhibit a wide range of emotions when interacting with game-based learning environments and the role of affect in this context has been the focus of continuing research (Sawyer et al., 2017; Sabourin & Lester, 2014). Studies have shown interesting relationships between affect and learning. For example, confusion, which is considered a positive emotion, can benefit learning when it is resolved or partially resolved (Zhang et al., 2021). On the other hand, negative emotions such as frustration and boredom have the potential to encourage unproductive behaviors like gaming the system and can thus have detrimental effects to learning (D'Mello & Graesser, 2011; Andres et al., 2015).

Frustration and boredom have been found to be associated with off-task behavior (Baker et al., 2010). Off-task behaviors are characterized by learners disengaging from the learning environment or task (Baker, 2007). These behaviors vary among learning environments and the learning goals they are trying to achieve. For example, Sabourin et al. identified (1) moving a task-related object to an unrelated location and (2) spending too much time in a location irrelevant to a task within Crystal Island as off-task behaviors (Sabourin et al., 2011). Baker (2007) operationalized a learner's idle time as an indicator of off-task behavior in a Cognitive Tutor software. He presented a machine-learned model for detecting off-task behavior by considering a learner's idle time along with other features. Idle time was considered to be all actions taken beyond a set cut-off time to perform a certain task. In game-based learning environments, students disengaging from a task but not necessarily from the learning environment also indicates off-task behavior. The students take a break from the difficult learning task and focus on the game-based aspects of the learning environment. Prior work found that this behavior is negatively correlated with positive learning outcomes but may also aid students in regulating negative affect and returning to the learning task (Sabourin & Lester, 2014).

In this paper, we look at the in-game behaviors of students who self-reported frustration and boredom in Minecraft, an open-ended, less-structured game. We also investigate the relationship of learner self-reported frustration and boredom to their knowledge assessments and STEM interest.

4. Data Collection

The data analyzed for this paper consisted of in-game data, out-of-game assessment data, and results of the game experience questionnaire (GEQ) and STEM interest questionnaire (SIQ) from 175 Grade 8 learners, 13 to 15 years old, from a school in the Philippines. Partner teachers from the school developed two (2) learning modules that integrated the use of WHIMC in classroom lessons about biotic/abiotic components of the environment and adaptation.

WHIMC is instrumented to collect data about learner actions as they use the game. WHIMC logs learner positions in intervals of 6 seconds as they explore the game, or whenever the learner makes an observation or access a science tool. Details of the interaction logs included the timestamp, coordinates, the world explored, observations made, and the science tool used. Time sequences of learner explorations done in Minecraft WHIMC of each learner were generated based on the interaction logs. The time sequence indicates the different worlds explored for the whole duration of the module implementation and the time spent on each world. The total time spent is computed by adding all the time spent in each exploration in the time sequence.

Knowledge assessments were done after every module. The assessments consisted of questions about the module topic and short essays. Questions for the short essays were the following:

Module 1 – "How can human activities affect the balance of ecosystems?"

Module 2 – "As people have moved from place to place, they have often brought plants and animals with them. How might the introduction of a new species of plant and animal in an area have disastrous effects on organisms already in that area?" and "How is biodiversity important in an organism's survival?"

The GEQ assessed players' levels of immersion, challenge, and negative / positive affect among others. The GEQ instrument used in this study is an abridged version of the instrument developed by IJsselsteijn et al. (2013). Two (2) items from GEQ: "*I felt frustrated"* and "*I felt bored*", were used for the data on student self-reported frustration and boredom respectively.

The SIQ is an instrument that is used to describe and measure the propensity of pursuing (or not pursuing) a career in STEM (See Table 1). Questions in the SIQ assessed students' self-efficacy as well as how they regarded the relevance and usefulness of science.

Table 1. *Items in SIQ*

SCCT Constructs		SIQ Item
Self-Efficacy (SE)	SE1	I know I can do well in science.
	SE2	I think Science is challenging to learn.
Outcome Expectations (OE)	OE1	After I finish high school, I will use Science often.
	OE2	I believe that I can use Math and Science to solve problems in the future.
Interests (IN)	IN1	I enjoy Science activities.
	IN2	I enjoy solving Science and Math problems.
Choice Goals (CG)	CG1	Learning Science will help me get a good job.
	CG2	Knowing how to use Math and Science together will help me to invent useful things.
	CG3	Understanding engineering is not important for my career.
Choice Actions (CA)	CA	I try to get a good grade in science because I have an interest in science jobs.

At the start of the data collection, the students were asked to answer the SIQ. They then explored WHIMC, following the instructions in the teacher-prepared learning modules. After the exploration, the learners answered the knowledge assessment, the GEQ, and the SIQ again.

5. Analysis and Results

5.1 How do the behaviors of the learners who self-reported frustration and boredom differ from those who did not?

We started the analysis by first dividing the learners into two groups, the high-performing group and the low-performing group, based on the knowledge assessment. The high-performing learners were the learners who scored equal to or above the mean and the low-performing learners were the learners who scored below the mean score.

We then focused our study of in-game behavior on idle time, time exploring other worlds, and task completion. Idle time was defined as disengaging from the learning environment. It was operationalized as the time during which the learner stayed at a location point, with no change in distance traveled and no other activity, i.e. accessing science tools or making an observation.

We calculated the cumulative distance traveled by each learner in one-minute intervals. We also calculated the slopes between one-minute intervals and then calculated the average slope for each learner to determine how the learner explored the worlds. Lower values of the average slope mean there is a small change in the cumulative distance. Figure 2 shows that most of the learners with frustration have smaller slopes in the change of cumulative distance per minute and there are times that the slope is zero which indicates that the learner did not move from its location.

The GEQ used Likert scale thus self-reported frustration and boredom were treated as ordinal data and the Spearman rho assessment was used in the analysis. The average slope of change of distance per minute was negatively correlated with frustration for all learners and the result was marginally significant ($r = -0.136$, $p = 0.1$). This implies that frustrated students tend to move around the WHIMC world less. This tendency is most common among low-performing students as we found a significant positive correlation between these students' idle time and frustration ($r = 0.287$, $p = 0.04$). The average slope was not significantly correlated with frustration for both high- and low-performing learners. Idle time was also not significantly correlated with frustration for all learners and high-performing learners.

Both average slope and idle time were not significantly correlated with boredom for all learners, as well as for both high- and low-performing learners.

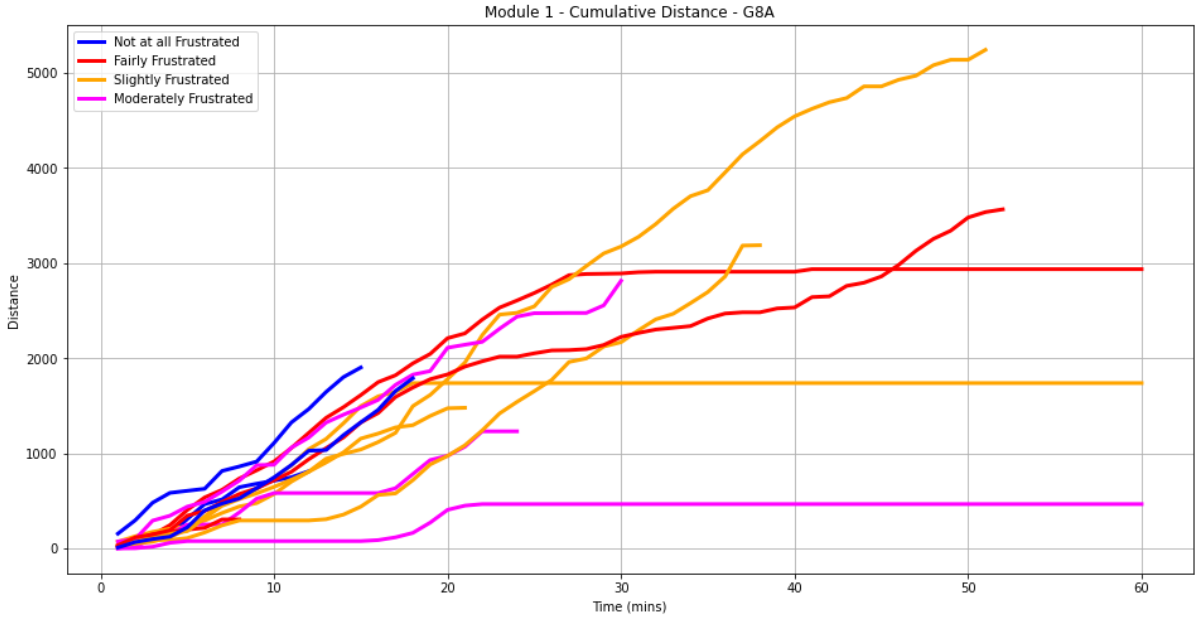


Figure 2. Cumulative Distance per minute interval of Learners and level of frustration in Module 1

Time exploring other worlds indicates the time that the learner visited worlds that were not part of the learning task. The time exploring other worlds was not significantly correlated with frustration or boredom for all learners, as well as for all high and low-performing learners.

Task completion refers to whether the missions per module were completed or not. Task completion of module 1 was marginally negatively correlated with frustration ($r = -0.185$, $p = 0.1$) among all learners and among high-performing learners ($r = -0.245$, $p = 0.1$). Task completion was not significantly correlated with boredom for all learners, as well as for all high- and low-performing learners.

5.2 What is the relationship between learners' self-reported frustration and boredom and their assessment outcomes?

We correlated the self-reported frustration and boredom with learners' assessment scores to determine the relationship between students' affective states and how they relate to their assessment outcomes. There was no significant correlation between frustration and assessment scores, but we found a significant negative correlation ($r = -0.17$, $p = 0.025$) between boredom and assessment scores. This suggests that bored students tend to learn less from playing with the game.

5.3 How did the STEM interest of the learners who self-reported frustration and boredom change after using WHIMC?

We then compared the pre-test and post-test results of the SIQ to determine how the STEM interest of the learners who self-reported frustration and boredom changed after they played with WHIMC. For self-reported frustration, we separated the learners into two groups: the first group was those learners who answered, "not at all" on the GEQ item "I felt frustrated" and the second group was those learners who have certain levels of frustration and answered "slightly", "moderately", "fairly", and "extremely". For self-reported boredom, another set of comparisons was also done. The learners were also separated into two groups: the first group was those learners who answered, "not at all" on the GEQ item "I felt bored" and the second group was those learners who have certain levels of boredom and answered "slightly", "moderately", "fairly", and "extremely".

Figure 3 shows that, on average, learners had a pretest SIQ of 3.56 and a post-test SIQ of 3.62. The difference was significant ($t(174)=2.13$, $p=0.03$). We also find significant differences in the Interest

category (t(174)=3.26, p<0.01) and Outcome Expectations category (t(174)=2.69, p=0.01). This implies that learners had increased interest in STEM after playing WHIMC.

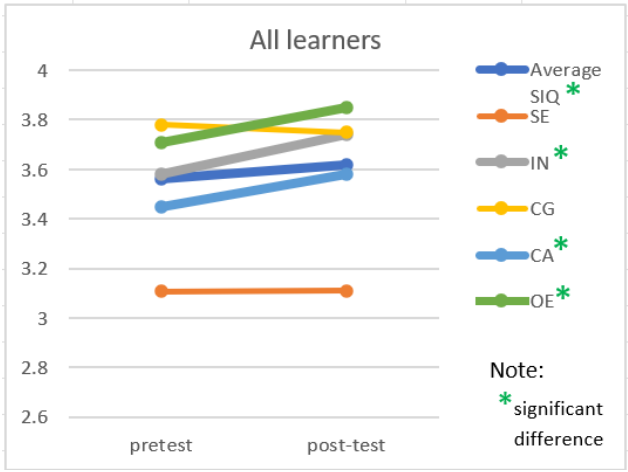
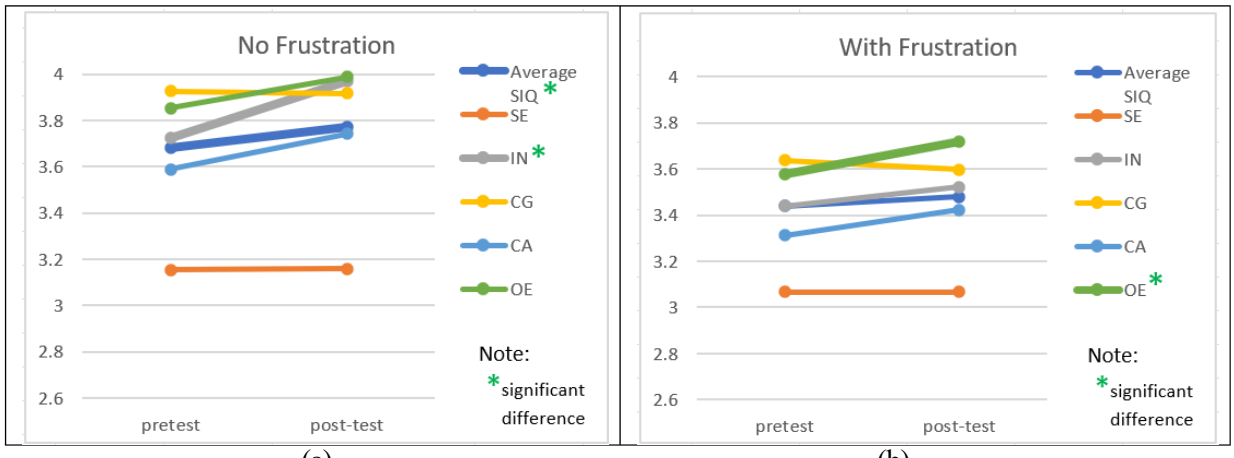


Figure 3. Comparison between pretest and post-test of SIQ of all learners

We now examine the differences in pretest and post-test SIQ among learners who expressed or did not express frustration. Figure 4a shows that, on average, learners who self-reported no frustration had a pretest SIQ of 3.68 and a post-test SIQ of 3.77. The difference was significant (t(84)=2.14, p=0.04). In the dimension of Interest in SIQ, the students had a pretest SIQ of 3.72 and a post-test SIQ of 3.97 and the difference between pretest and post-test was also significant (t(84)=3.92, p<0.01). This implies that students who self-reported no frustration had increased interest in STEM after playing WHIMC.

We compared the pretest and post-test SIQ scores of those who expressed frustration. Learners who self-reported frustration had a pretest SIQ of 3.44 and a post-test SIQ of 3.48 (see Figure 4b). We did not find a significant difference. However, when we examined student responses per category of SIQ, there was a significant increase in the category of Outcome Expectations (t(89)=2.07, p=0.04).

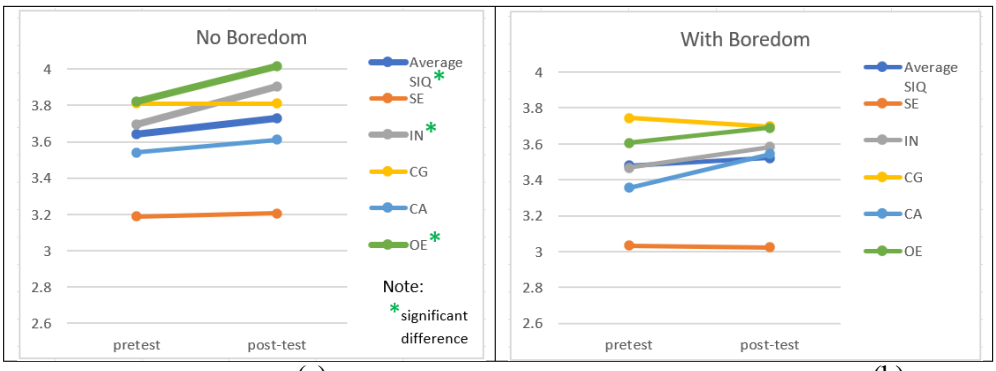


(a) (b)

Figure 4. Comparison between pretest and post-test SIQ of all learners who self-reported no frustration (a) and learners who self-reported frustration (b)

We also examined the differences in pretest and post-test SIQ among learners who expressed or did not express boredom. Learners who self-reported no boredom had a pretest SIQ of 3.64 and a post-test SIQ of 3.73 (see Figure 5a). The difference was significant (t(84)=2.6, p=0.01). We also find significant differences in the Interest category (t(84)=3.72, p<0.01) and Outcome Expectations category (t(84)=2.8, p=0.01). On the other hand, there was no significant difference in pretest and post-test SIQ

among learners who expressed boredom (see Figure 5b). This implies that learners who did not express boredom had increased interest in STEM after playing WHIMC but those who expressed boredom had no significant change in STEM interest.



(a) (b)

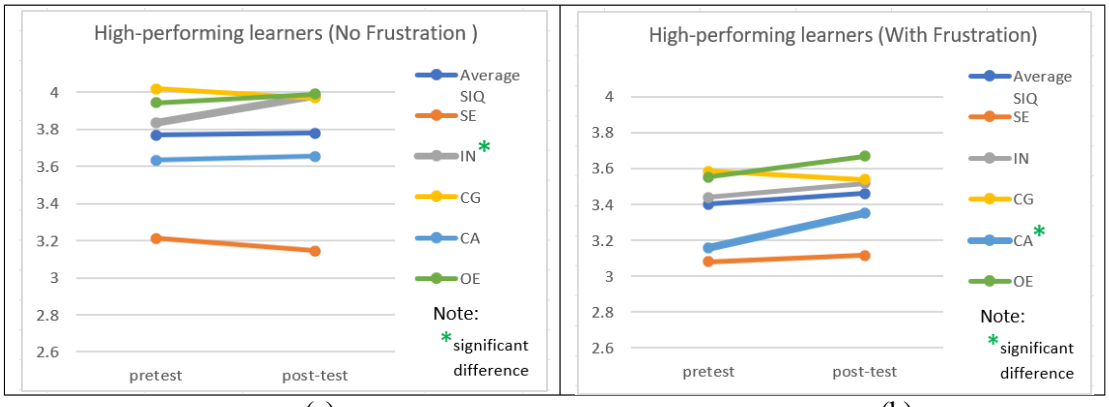
Figure 5. Comparison between pretest and post-test SIQ of all learners who self-reported no boredom (a) and learners who self-reported boredom (b)

5.4 How did the STEM interest of high- and low-performing learners who self-reported frustration and boredom change after using WHIMC?

Finally, we then tried to determine whether frustration affected high and low-performing learners differently. We found that high-performing learners who self-reported no frustration had a pretest SIQ of 3.77 and a post-test SIQ of 3.78 (see Figure 6a). The difference was not significant. However, there was a significant difference in the SIQ category of Interest ($t(51)=2.01$, $p=0.05$) with a pretest SIQ of 3.84 and a post-test SIQ of 3.98. On the other hand, the high-performing learners who self-reported frustration had a pretest SIQ of 3.4 and a posttest SIQ of 3.46 (see Figure 6b), in which the difference was not significant. However, there was a significant difference in the Choice Actions category ($t(56)=1.96$, $p=0.05$) with a pretest SIQ of 3.16 and a post-test SIQ of 3.35.

Figure 7a shows that low-performing learners who self-reported no frustration had a pretest SIQ of 3.55 and a post-test SIQ of 3.76 ($t(32)=2.8$, $p=0.01$). We also find significant differences in the Interest category ($t(32)=3.66$, $p<0.01$) which has a pretest SIQ of 3.55 and a post-test SIQ of 3.95, and in the Outcome Expectation category ($t(32)=2.06$, $p=0.05$) which has a pretest SIQ of 3.7 and a post-test SIQ of 3.99. On the other hand, low-performing learners who self-reported frustration had a pretest SIQ of 3.5 and a post-test SIQ of 3.53 (see Figure 7b), in which the difference was not significant.

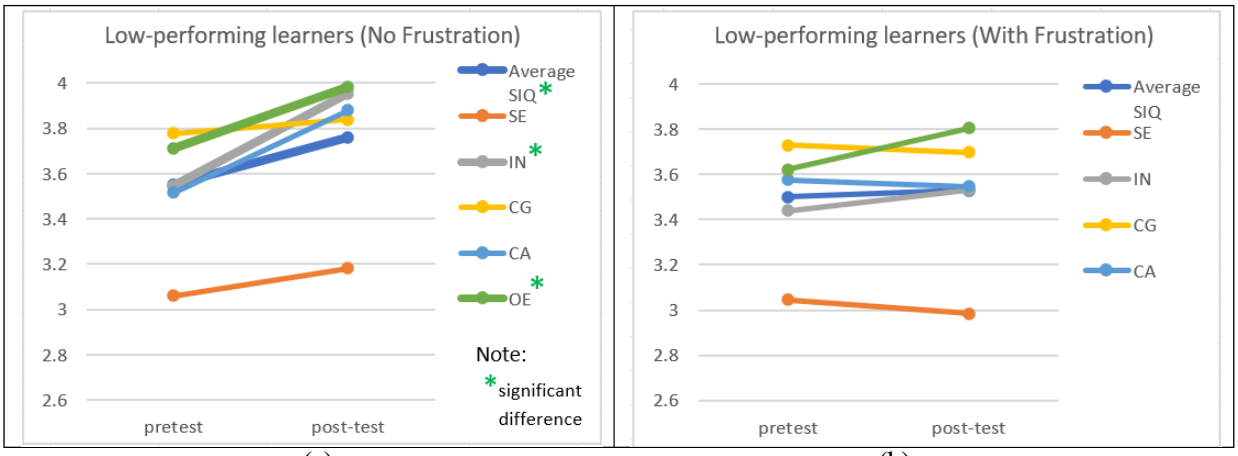
Based on the results, both the high-performing and low-performing learners who self-reported no frustration showed increases in some aspects of STEM interest.



(a) (b)

Figure 6. Comparison between pretest and post-test SIQ of high-performing learners

who self-reported no frustration (a) and those who self-reported frustration (b)



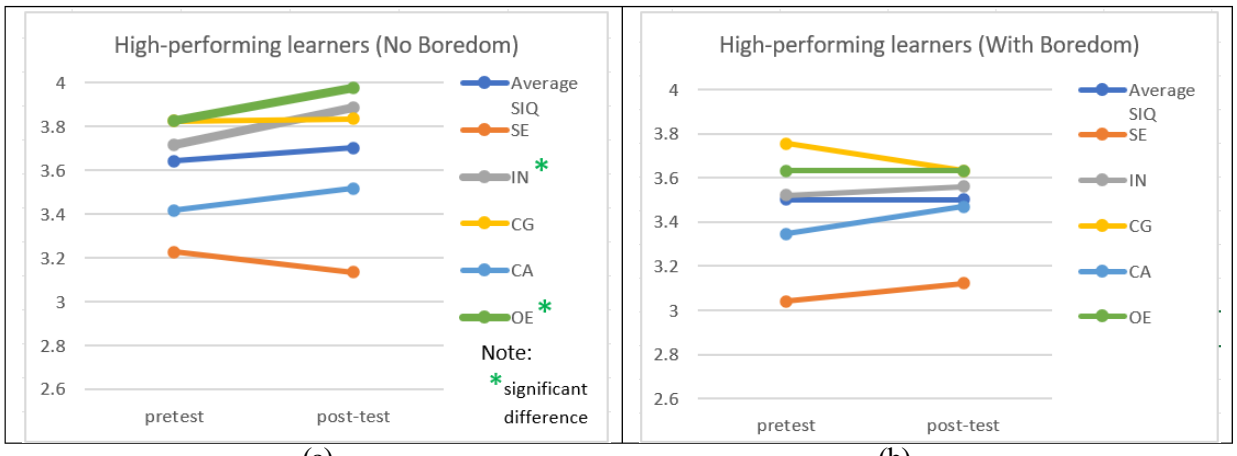
(a) (b)

Figure 7. Comparison between pretest and post-test SIQ of low-performing learners who self-reported no frustration (a) and those who self-reported frustration (b)

We also tried to determine whether boredom affected high and low-performing learners differently. The high-performing learners who self-reported no boredom had a pretest SIQ of 3.64 and a post-test SIQ of 3.7 (see Figure 8a). The difference was not significant. However, there was a significant difference in the Interest category ($t(59)=2.95$, $p=0.05$) with a pretest SIQ of 3.72 and a post-test SIQ of 3.88. On the other hand, the high-performing learners who expressed boredom had a pretest SIQ of 3.5 and a post-test SIQ of 3.5 (see Figure 8b), in which there was no significant difference. Neither did we find significant differences in the disaggregated SIQ categories.

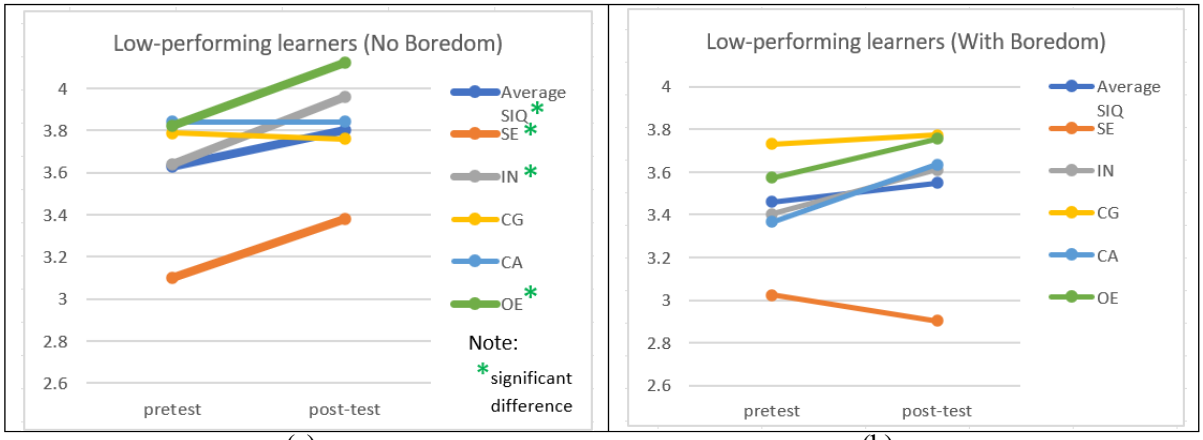
Low-performing learners who did not express boredom had a pretest SIQ of 3.63 and a post-test SIQ of 3.8 (see Figure 9a). The increase was significant ($t(24)=2.1$, $p=0.05$). We also found significant differences in the Interest category ($t(24)=2.32$, $p=0.03$) with a pretest SIQ of 3.64 and a post-test SIQ of 3.96, Self-Efficacy category ($t(24)=2.58$, $p=0.02$) with a pretest SIQ of 3.1 and a post-test SIQ 3.38, and Outcome Expectations category ($t(24)=2.08$, $p=0.05$). On the other hand, low-performing learners who expressed boredom had a pretest SIQ of 3.46 and a post-test SIQ of 3.55 (see Figure 9b), in which there was no significant difference.

Based on the results, high-performing learners who did not express boredom showed an increase in some aspects of STEM interest after playing WHIMC. Low-performing learners who did not express boredom had increased levels in the Self-Efficacy category as well as the aggregated SIQ. For those who self-reported boredom, there was no significant change in the STEM interest after using WHIMC.



(a) (b)

Figure 8. Comparison between pretest and post-test SIQ of high-performing learners who self-reported no boredom (a) and those who self-reported boredom (b)



(a) (b)

Figure 9. Comparison between pretest and post-test SIQ of low-performing learners who self-reported no boredom (a) and those who self-reported boredom (b)

6. Conclusion

In this paper, we presented an attempt to analyze the behavior of learners who self-reported frustration and boredom and determine the relationship between frustration and boredom to their in-game behaviors, knowledge assessment outcomes, and STEM interest.

We found that learners who are frustrated tend to move around the WHIMC world less. Frustrated learners also tended not to complete the missions or tasks. Low-performing frustrated learners tend to have more idle time, which is an indicator of off-task behavior.

When we examined the relationship between learners' frustration and boredom with their knowledge assessment outcomes, bored learners tended to have lower academic outcomes.

Learners who did not express frustration or boredom had increased interest in some categories of STEM interest after using WHIMC. Frustrated learners also had increased SIQ in the Outcome Expectations category while learners who expressed boredom had no significant change in SIQ. Frustrated learners tend to believe that after they finish high school, they will use Science often, and they also tend to believe that they can use Math and Science to solve problems in the future.

Both high- and low-performing learners who did not express frustration and boredom had increased interest in some categories of STEM. High-performing learners who expressed frustration had increased SIQ in the Choice Actions category. On the other hand, there was no significant difference in the pretest and post-test SIQ for both high- and low-performing bored learners. Note that low-performing learners who did not express frustration or boredom had increased SIQ scores, implying that the absence of these states was beneficial to this category of learners.

This study contributes to what is known about how affect relates to outcomes such as test scores, in-game behaviors, and attitudes regarding domains such as STEM. In general, students who expressed boredom or frustration tended to demonstrate minimal of engagement with STEM, as indicated by lower test scores and flat SIQ results. The study also contributes by raising the larger question about whether and how game-based interventions can be used to mitigate affective states such as frustration and boredom. How can educators best use games to pave the way towards greater student engagement in STEM? This is a research area that is open for further study.

Acknowledgements

The authors thank H Chad Lane and Jeff Ginger for their enthusiastic collaboration, Dominique Marie Antoinette B. Manahan, Jonathan D. Casano, and Ma. Rosario Madjos for their support, the Ateneo Laboratory for the Learning Sciences, the Ateneo de Manila University, the Department of Science and

Technology (DOST) as funding agency, and the Department of Science and Technology's Philippine Council for Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD) as monitoring agency for the grant entitled, "Nurturing Interest in STEM among Filipino learners using Minecraft."

References

All, A., Nuñez Castellar, E. P., & Van Looy, J. (2016). Assessing the effectiveness of digital game-based learning: Best practices. *Computers & Education*, *92-93*, 90–103. https://doi.org/10.1016/j.compedu.2015.10.007

Andres, J. M. L., Rodrigo, M. M. T., Baker, R. S., Paquette, L., Shute, V. J., & Ventura, M. (2015). Analyzing Student Action Sequences and Affect While Playing Physics Playground. *In EDM (Workshops)*.

Baker, R. S. J. (2007). Modeling and understanding students' off-task behavior in intelligent tutoring systems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Baker, R. S. J., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, *68*(4), 223–241. https://doi.org/10.1016/j.ijhcs.2009.12.003

Beserra, V., Nussbaum, M., Zeni, R., Rodriguez, W., & Wurman, G. (2014). Practising arithmetic using educational video games with an interpersonal computer. *Journal of Educational Technology & Society*, *17*(3), 343-358.

Bitner, J. (2021). What is Minecraft? Gaming. Digital Trends. Accessed from the Digital Trends website: https://www.digitaltrends.com/gaming/what-is-minecraft/

D'Mello, S., & Graesser, A. (2012). Dynamics of Affective States during complex learning. *Learning and Instruction*, *22*(2), 145–157. https://doi.org/10.1016/j.learninstruc.2011.10.001

D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition and Emotion*, *25*(7), 1299–1308. https://doi.org/10.1080/02699931.2011.613668

IJsselsteijn, W. A., De Kort, Y. A., & Poels, K. (2013). The game experience questionnaire.

Käser, T., & Schwartz, D. L. (2020). Modeling and analyzing inquiry strategies in open-ended learning environments. *International Journal of Artificial Intelligence in Education*, 30(3), 504–535.

Richey, J. E., Andres-Bray, J. M., Mogessie, M., Scruggs, R., Andres, J. M. A. L., Star, J. R., Baker, R. S., & McLaren, B. M. (2019). More confusion and frustration, better learning: The impact of erroneous examples. *Computers & Education*, 139, 173–190. https://doi.org/10.1016/j.compedu.2019.05.012

Liu, Z., Pataranutaporn, V., Ocumpaugh, J., & Baker, R.S.J.d. (2013). Sequences of Frustration and Confusion, and Learning. *Proceedings of the 6th International Conference on Educational Data Mining*, 114-120.

Pardos, Z. A., Baker, R. S. J. D., San Pedro, M., Gowda, S. M., & Gowda, S. M. (2014). Affective states and state tests: Investigating how affect and engagement during the school year predict end-of-year learning outcomes. *Journal of Learning Analytics*, *1*(1), 107–128. https://doi.org/10.18608/jla.2014.11.6

Rajendran, R., Iyer, S., Murthy, S., Wilson, C., & Sheard, J. (2013). A theory-driven approach to predict frustration in an ITS. *IEEE Transactions on Learning Technologies, 6*(4), 378-388.

Sabourin, J. L., & Lester, J. C. (2014). Affect and engagement in game-based learning environments. *IEEE Transactions on Affective Computing*, *5*(1), 45–56. https://doi.org/10.1109/t-affc.2013.27

Sabourin, J., Rowe, J. P., Mott, B. W., & Lester, J. C. (2011). When off-task is on-task: The affective role of off-task behavior in narrative-centered learning environments. *In International Conference on Artificial Intelligence in Education* (pp. 534-536). Springer, Berlin, Heidelberg.

Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2017). Enhancing student models in game-based learning with facial expression recognition. *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. https://doi.org/10.1145/3079628.3079686

Tze, V. M., Daniels, L. M., & Klassen, R. M. (2015). Evaluating the relationship between boredom and academic outcomes: A meta-analysis. *Educational Psychology Review*, *28*(1), 119–144.

Vogel-Walcutt, J. J., Fiorella, L., Carper, T., & Schatz, S. (2012). The definition, assessment, and mitigation of state boredom within educational settings: A comprehensive review. *Educational Psychology Review, 24*(1), 89-111.

Zhang, Y., Paquette, L., Baker, R. S., Ocumpaugh, J., Bosch, N., Biswas, G., & Munshi, A. (2021). CAN strategic behaviour facilitate confusion resolution? the interplay between confusion and metacognitive strategies in Betty's brain. *Journal of Learning Analytics*, *8*(3), 28–44. https://doi.org/10.18608/jla.2021.7161

Zhonggen, Y. (2019). A meta-analysis of use of serious games in education over a decade. *International Journal of Computer Games Technology*, 2019, 1–8. https://doi.org/10.1155/2019/4797032

WHIMC. (n.d.) What-If Hypothetical Implementations using Minecraft. Available online: https://whimcproject.web.illinois.edu/

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Two-way Knowledge Transfer Among University Academics, Young Entrepreneurs, NGOs and Students in STEM and IoT Metaverses: Conceptual Model, Research Agendas and Contextual Challenges

Percy Lai-Yin KWOK [a*], Joe Kwong-Ngai CHEN [b] & Jordan Tsz-Chun FUNG [c]
[a]*The Education University of Hong Kong*
[b] *ICT-In-Physical Education Foundation Ltd.*
[c] *Jordan Fung & Company (Hong Kong) Ltd.*
*lykwok@eduhk.hk

Abstract: This paper presents an innovative metaverse design through collaborations among university academics, industry entrepreneurs, NGOs and students through a two-way knowledge transfer (KT) model. Unlike previous one-way KT approaches, this paper endeavors to formulate an interactive two-way KT conceptual model with evaluation of creative technological products, entrepreneur coaching programs and social caring services through (Arts/Green/Sports) STEM and IoT Metaverses. In the top layer of the KT model, one gigantic integrative network of academics, scientists and creative technologists work as a big data pool to provide high-tech professional knowledge. In the lower layers, collaborations among university academics, technology developers, industry entrepreneurs, NGOs and students work out knowledge-building community-of-practices using youth entrepreneur leadership models to develop technological products and provide services. For sustainability of knowledge transfer, educational / social funds will be used to support student activities financially. Participating young entrepreneurs care the elderly and disadvantaged people and develop their moral conscience during some STEM activities in empathic societies. Noteworthy, this paper presents five key features of decentralization, immersive properties, digital currency ecosystems (using blockchain technology), social experiences and persistency in the construction of met connecting real and virtual worlds. It also shows valuable learning experiences of participating students in STEM, game-based learning and IoT metaverses to testify the feasibility of knowledge transfer. Future research agendas in knowledge transfer in metaverses and contextual challenges will be addressed through action project reflection in the paper.

Keywords: Knowledge transfer, metaverses, STEM, IoT, action project

1. Introduction

Metaverse is a current and future integrated totality of the physical world and virtual world connecting internet-of-things (IoT) and other aspects of human lives with a wide range of applications of (wired / wireless) computer networks, handheld mobile devices, augmented reality / virtual reality / substitutional reality / mixed reality / extended reality (AR/VR/SR/MR/XR), artificial intelligence (AI) and robotics. Currently, metaverses provide a range of human imaginations and potential functionalities that enable new modes of e-business, e-trades, e-digital currencies and e-learning, and lots of immersive social and virtual experiences in STEM education and game-based learning (GBL). IoT and STEM metaverses refer to focused uses of internet-of-things (IoT) and STEM applications respectively into daily lives connecting the physical world and virtual world in metaverses. Proponents of metaverse education often pay heed to its virtual world construction but fail to pay close attention to the connections between physical and virtual worlds especially when some new technology is being applied into our daily lives (Hirsh-Pasek, *et al.* 2022). In this paper, an innovative two-way knowledge transfer model is implemented with evaluation in some IoT and STEM metaverses.

Previous knowledge management (KM) models aim to turn implicit knowledge into explicit knowledge, partly originated from Polanyi (1962)'s distinction between tacit and explicit knowledge. New KM emerging concepts in 1995-2005 include knowledge economy, knowledge alliance, knowledge culture, knowledge organization, knowledge infrastructure, and knowledge equity (Baskerville & Dulipovici, 2006). However, it is very difficult for KM and educational researchers to find out those socio-cultural mechanisms for transforming implicit knowledge into explicit one. Facing such problems, some KM researchers in school education would like to start firstly with KT from individual students' tacit knowledge and sharing of learning experiences in apprenticeship or mentorship through **S**ocialization; secondly, transforming from tacit knowledge to explicit one in group activities through **E**xternalization; thirdly, a **C**ombination (socio-cognitive synthesis) of existing explicit knowledge into new explicit knowledge; and fourthly, students' putting what they have learned into practices from their explicit knowledge through individual **I**nternalization by adopting traditional KM (SECI) models of Nonaka & Konno (1998), Nonaka & Toyama (2003). The two processes for knowledge entering and leaving the KT system are *assimilation* and *dissemination* respectively in organizational knowledge. There would be a strong potential of using some ICT innovations in virtual worlds (AR/VR/SR/MR/XR) and real world to make connections with those socio-cognitive features in the four-step spiraling processes of SECI interactions between tacit knowledge and explicit knowledge through metaverse construction. And KT should not be limited to teacher-oriented knowledge transfer (TKT) (like Peng *et al.* 2021) and student-oriented KT should have strong communal connections with other members of communities-of-practices beyond school education.

Knowledge transfer (KT) is more than learning. KT conceptually differs from knowledge sharing or knowledge dissemination, which was merely used for knowledge exchange between sources and recipients of knowledge. KT takes certain forms of information science and software engineering as knowledge reuse (Markus, 2001) but extends beyond the KM organizations' absorptive capacity to identify, assimilate and exploit knowledge (Alavi & Leidner, 2001). Previous KT models in school and workplace education have only one-way direction from high-level technical knowledge in industry or academics to students and teachers or very limited two-way interactions between academic institutions and the larger communities without direct, practical applications to schools. Therefore, there would be another strong feasibility of constructions of AR/VR/SR/MR/XR to connect with student knowledge transfer (SKT) to go beyond the scope of prior KT research studies through STEM and IoT metaverses.

2. Interactive Two-way Knowledge Transfer Model (See Appendix 1)

2.1 Importance of big data pools for provision of high-tech knowledge

In the top layer of the KF model, one gigantic integrative network of academics, scientists and creative technologists as a big data pool to provide high-tech professional knowledge. Without such content knowledge, KT cannot be initiated or updated as students need to obtain new professional knowledge directly from those experts and indirectly from university academics. In Web 3.0 metaverses, student learners are engaged into a variety of NFT STEM programs in the action projects. There are five underlying characteristics of IoT metaverse depicted in table 1.

Table 1. *Key features of metaverses and socio-cognitive functions in student activities*

Key Features of Metaverses	Socio-cognitive Functions in Two-way Knowledge Transfer (KT) model	Students' exemplified activities in IoT and STEM metaverses
Decentralized	Communal ownership enhances socialization and externalization in KT	Students learn Ethereum blockchain to understand the importance of communal ownership for e-trade transactions, tokens, e-marketing prices & bitcoins
Immersive	4-D / 5-D visualization facilitates members' embeddedness into metaverses for cognitive assimilation and dissemination in KT	Students' valuable 'immersive' properties can be actualized by their usage of AR glasses and VR/MR headsets with some multiple-role perceptions of ultra-realistic digital worlds in metaverses experiences
Digitalization	Monetization in digital currencies to	In digital currency ecosystems using

	build value systems of buying goods and consuming services in combination (cognitive synthesis) in KT	blockchain technology, students use e-cash coupons in AI shops to buy / sell products and services in T metaverses
Social experience	Gamification and STEM / IoT experiences builds empathized knowledge and moral conscience through internalization, externalization, socialization and combination in KT	Students exercise young entrepreneurship skills to produce new knowledge products and provide caring services to the elderly and disadvantaged people
Persistency	Accessibility without spatiotemporal constraints consolidates communal membership and further enhances socialization and externalization in KT	Students share their STEM / IoT experiences by interacting with other members without spatiotemporal constraints in STEM / IoT metaverses

There is a strong potential for metaverse members to be fully engaged in private and sharable communal spaces without possible threats posed by the pandemic (Hady, 2022). Such privacy and health security further enhances two-way KT.

2.2 Key features of metaverses and new educational directions of metaverses

The following features of IoT metaverses help foster socio-cognitive development of students and other key members of the two-way KT model. Some shortcomings of traditional collaborative learning communities-of-practices would not be found in IoT metaverses. Various types of knowledge transfer (KT) have been fostered and sustained by the following features of IoT metaverses (e.g. What is Metaverse? Step-by-step guide into the Metaverse, 2022):

- Awareness of ethical and legal concerns in members' social and moral responsibilities, data privacy, patent and intellectual property rights of members and products / services in metaverses
- Flexibility and diversities in all types of individual and social seamless learning, formal and informal education, (serious or causal) game-based learning (GBL) and non-game-based learning, academic discussions and experiential sharing (in videoconferencing, cloud phones, and online events platforms) and knowledge-building discourses anytime and everywhere
- Freedom earned in selection of learning and teaching modes, choices in real world (RW), virtual world (VW) and RW-VW, inviting or recruiting particular members or groups of members for interactions or accepting others' invitation or recruitment (allowing self-learning times and modes for deep or extensive learning)
- Transferability of information and sharing in single-, cross- and multiple-platforms in all types of PCs, handheld devices, dataloggers, mobiles, other PCs in wired / wireless networks, IoTs and so forth

2.3 Roles of NGOs in knowledge transfer

Being a good strategic partner in two-way KT, non-government organizations (NGOs) have played significant roles of catering for ongoing and ever-changing needs of local communities in flexible mode, creating innovative collaboration projects for diverse needs of schools and social service centers, providing realistic startup and meaningful youngsters internship opportunities through project collaboration, maintaining regular business relationships with technology vendors and social worker partners and connecting learning, course materials and learning tools with real industry. In some circumstances, NGOs perform more flexibly and earn more freedom for collaborators to explore than formal education partners in the perspectives of industry entrepreneurs and students in the aforementioned action projects (Institute for Social Innovation, 2022; Zbuchea *et al.,* 2019).

3. Action Research Design and Deliverable Products / Services

3.1 Overview of three cycles of action projects

The action projects run in ceaseless spiral cycles of 'planning-monitoring-evaluating' in two successive stages. *Stage 1:* initiating individual-, group and class-oriented IoT and STEM projects with stepwise evaluations and preparing for future development of related metaverses at micro-and meso-levels (one-way KT). *Stage 2:* connecting current projects and developing big-scale ones after stepwise evaluation, adjustment and modifications in current development of related metaverses at micro-and meso-levels (two-way KT)

In Stage 1, a series of innovative STEM programs and products have been delivered by the second and third authors in the following activities in table 2.

Table 2. *Learning Outcomes in STEM programs related to KT*

Programs or Stories Behind	Learning outcomes (sharing links)	Aspects of KT
1. Projection Mapping Prince Fat Shing@ The Education University of Hong Kong:	True love to share the STEM education experience of climbing up from the bottom of life: https://youtu.be/3mbIgWe5xHw	Sharing student learning experiences for promoting educational services in STEM education
2. Cyberport Tai Chi Pass vs MoCap Events	Co-organized school summer activities held after students completed the course internship https://youtu.be/F42LxaNzZ34	Actualizing Sports STEM into martial arts Tai Chi using student internship
3. AIR – AI Retail (Community Retail):	Students' usage of e-cash coupons in purchasing products and consuming services https://www.ictinpe.org/ai-ministore	Sharing student learning experiences in digital currency ecosystems
4. IGB – IoT Green BMS (Building Management System)	Students' awareness in Green education https://www.ictinpe.org/iot-classroom	Raising awareness in Green education and revolution in empathic societies

In Stage 2, several educational, social and industrial funds are being drawn to support a series of Blockchain and NFT STEM program which will be delivered to some needy students. For instance, the two-way knowledge transfer will take place when one young entrepreneur with undergraduate major in Blockchain technology delivers some lectures and conduct some instep training tutorials for a group of participating secondary school students.

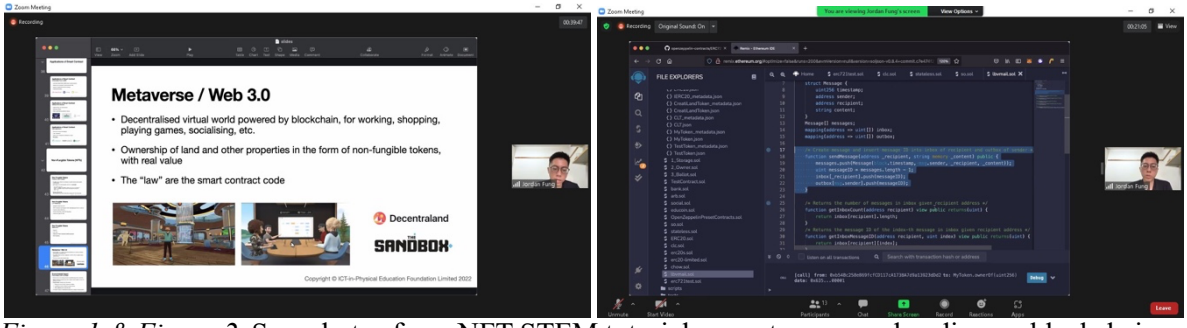


Figure 1 & Figure 2. Snapshots of one NFT STEM tutorial on metaverse and coding on blockchain

3.2 Nature of Collaborations Between Academics, Industry Entrepreneurs, NGOs and Students

In the two-way KT model, university academics provide professional research support, look for fund and co-organize coaching and training programs for participating students and young entrepreneurs. Industry entrepreneurs devise innovative technological products and services, launch coaching and training programs on new technology to needy students, and finetune or improve them after seeking valuable student feedback. Meantime, NGOs provide funding support, recruit young knowledge co-workers, provide career planning services, provide appropriate industrial resources (free space, connections with people and intern programs) for developing youth entrepreneurship models. Participating schools maximize students' other learning experiences (OLE) resources and applying for external grants to purchase necessary equipment and other learning resources.

3.3 Youth Entrepreneur Leadership Model in Metaverses

Past research literature on youth entrepreneurship focuses on social and cultural capital (Turner & Nguyen, 2005) and entrepreneurial resourcefulness (Quagrainie *et al.,* 2022). This paper presents another model using KT in metaverse education. During the implementation stages 1 and 2, proactive student leaders have played distinctive roles of creating innovative technological products in the business / industry sector, coordinating intern teams and supporting collaborations team works, nourishing moral conscience and positive / value education in entrepreneurship and providing caring

services to needy (elderly and disadvantaged) people. Such traits exhibit a new young entrepreneur leadership model in metaverses, which is an under-researched topic in ICT education (c.f. training programs offered by Summer Discovery, 2022). Youth entrepreneurs use an intrinsic motivation to explore interested learning topics and invent creative technology products, and they learn by mistakes.

4. Future Research Agendas

Since the applications of metaverses in school and workplace education are still in an immature stage. Currently, there has not been any detailed cognitive or socio-cognitive international or local studies on using IoT and STEM metaverses in school and workplace education. After the first and second stages of the action project, some heated research agendas need to be addressed such as quantitative and qualitative measurement of knowledge transfer (KT), socio-cognitive aspects of KT in Arts / Green / STEM and IoT metaverses, socio-cultural and sociological issues like gender inequalities, parental motives for buying paid products and services from schools and industry entrepreneurs and curricular development of relevant blockchain application course modules to those needy students in Arts / Green / STEM and IoT metaverses

Challenges ahead include sustainability in subsequent / parallel development of various types of metaverses, continuity of funding support, professional accreditation of organized products and deliverable services, lack of academic / know-how knowledge in blockchain applications in school education and uncertain collaboration channels and unclear virtual world-real world connections.

References

Alavi, M., & Leidner, D.E. (2001) Knowledge Management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Quarterly, 25*, 107-136. http://dx.doi.org/10.2307/3250961

Baskerville, R., & Dulipovici, A. (2006). The theoretical foundations of knowledge management. *Knowledge Management Research & Practice 4*, 83–105.

Hady, L. (2022). How the metaverse can help students and teachers adapt to a post-pandemic world https://www.scmp.com/comment/opinion/article/3160672/how-metaverse-can-help-students-and-teachers-adapt-post-pandemic

Hirsh-Pasek, K., Zosh, J. M., Hadani, H. S., Golinkoff, R. M., Clark, K., Donohue, C., & Wartella, E. (2022). *A whole new world: Education meets the metaverse:* https://www.brookings.edu/research/a-whole-new-world-education-meets-the-metaverse/

Institute for Social Innovation (2022). 4 key factors for improving knowledge management in NGOs. https://dobetter.esade.edu/en/knowledge-management-ngos

Markus, M. L. (2001). Toward a Theory of Knowledge Reuse: Types of Knowledge Reuse Situations and Factors in Reuse Success. *Journal of Management Information Systems, 18* (1):57-94.

Nonaka, I. and Konno, N. (1998). The Concept of Ba: Building a Foundation for Knowledge Creation. *California Management Review, 40,* 40-54. https://doi.org/10.2307/41165942

Nonaka, I. and Toyama, R. (2003). The Knowledge-Creating Theory Revisited: Knowledge Creation as a Synthesizing Process. *Knowledge Management Research & Practice, 1*, 2-10. http://dx.doi.org/10.1057/palgrave.kmrp.8500001

Peng, M. Y. P., Feng, Y.J., Zhao, X., & Chong, Y. L. (2021). Use of Knowledge Transfer Theory to Improve Learning Outcomes of Cognitive and Non-cognitive Skills of University Students: Evidence From Taiwan. *Front. Psychol. 12:*583722. doi: 10.3389/fpsyg.2021.583722

Polanyi, M. (1962). *Personal knowledge: Towards a post-critical philosophy.* Chicago, IL: University of Chicago.

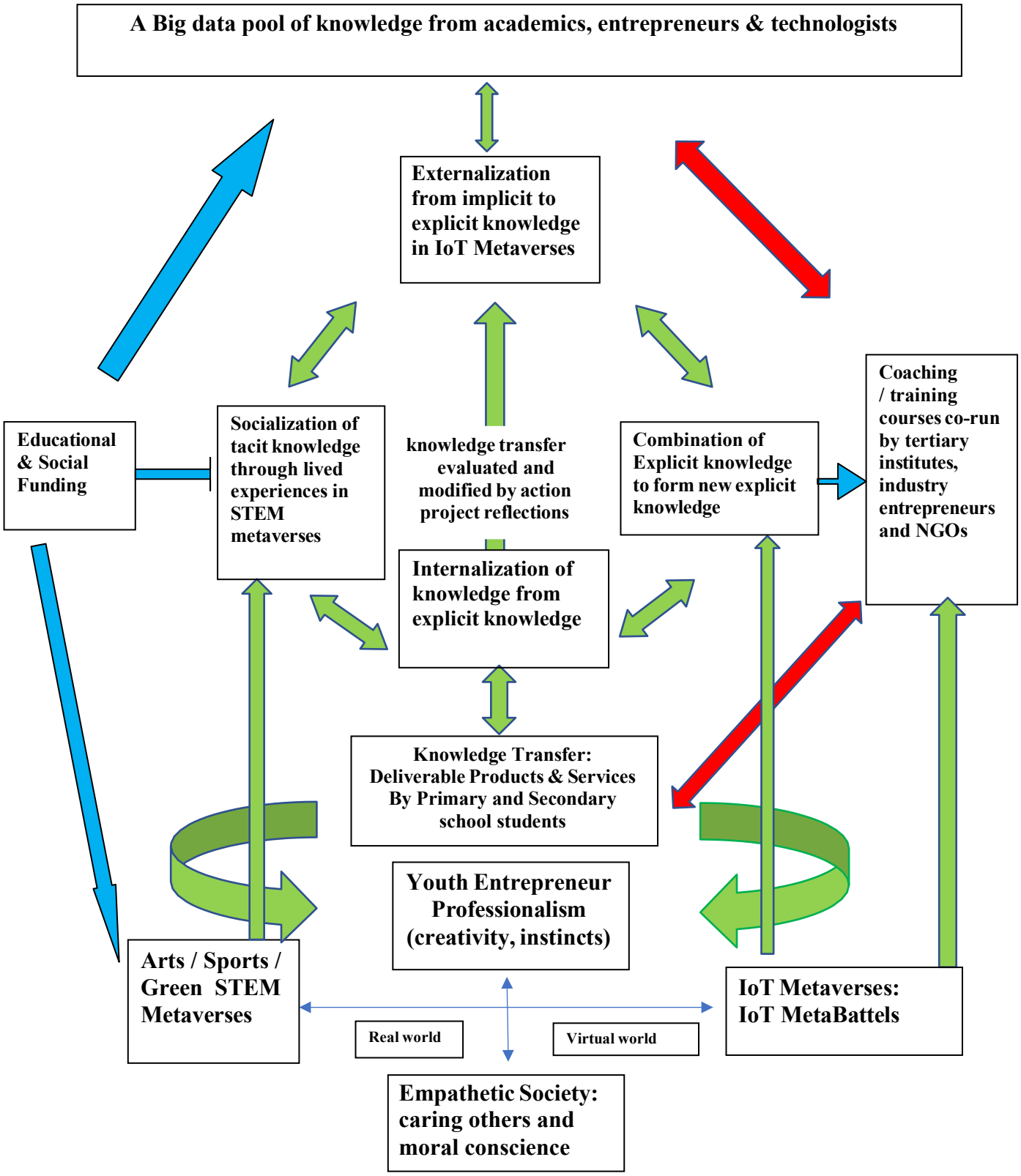
Quagrainie, F.A, Kabalan, A. A. M., Adams, S. & Dankwa, A. (2022). 3Ps model of entrepreneurial resourcefulness among micro-youth entrepreneurs. *Journal of Entrepreneurship in Emerging Economies* (IF), doi: 10.1108/jeee-07-2021-0276

Summer Discovery (2022). Pre-College Career Accelerators > Entrepreneurship for the Web 3.0 Metaverse. https://www.summerdiscovery.com/career-accelerators/entrepreneurship-for-the-web-30-metaverse

Turner, S., & Nguyen P. A. (2005). Young entrepreneurs, social capital and doi moi in Hanoi, Vietnam. *Urban Studies, 42 (10),* 1693– 1710.

What is Metaverse? Step-by-step guide into the Metaverse (2022). https://apptask.com/what-is-metaverse-step-by-step-guide-into-the-metaverse-%f0%9f%91%be/?gclid=CjwKCAjwx46TBhBhEiwArA_DjM8mLfE2XXD16N7XRP5S7Hez7uvuUAxvFuFS3qso-oOvqaxWlCY2AxoCFxsQAvD_BwE

Zbuchea, A., Ivan, L., Petropoulos, S., & Pinzaru, F. (2019). Knowledge sharing in NGOs: The importance of the human dimension. *Kybernetes*, *49* (1), 182-199. https://doi.org/10.1108/K-04-2019-0260



Green arrows denote knowledge transfer (KT)
Red arrows denote educational coaching / training;
Blue arrows denote financial support

Appendix 1. **A Multi-level Two-Way KT Model in STEM and IoT Metaverses**

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

What do Teachers Want to Know About Game-Based Learning Analytics: Cross-Case Study

Yoon Jeon KIM, Jennifer SCIANNA

University of Wisconsin-Madison, U.S.A
yj.kim@wisc.edu

Abstract: Game-based pedagogy requires innovation in learning analytics, so teachers can make sense of the complex data generated in the game in relation to evidence for learning. In this paper, we advocate for co-design as a process to elucidate the types of data that teachers want and need to effectively implement games in the classroom. We examine two cases to discuss affordances of co-design activities and present initial findings that built towards two prototypical dashboards.

Keywords: game-based learning analytics, teacher dashboard, co-design

1. Introduction

Educational games, a specific form of games designed to support and assess students' learning of educationally valuable knowledge and skills, provide a context to expand what learning means and what types of evidence educators can use to better understand what and how students are learning. Learning analytics (LA) and educational data mining (EDM) offer a unique opportunity to translate these insights, that are often hard to manually process, to the teachers who are using games as a learning tool in classrooms. There are many examples of the successful applications of LA and EDM techniques in game-based environments: sequence mining of problem solving processes (Gomez et al., 2021), prediction of how likely a player is to quit (Karumbaiah et al., 2018), prediction of prosocial behaviors (Alonso-Fernández et al., 2020), conceptual understanding of academic domains (e.g. Kim & Shute, 2015), identification of different patterns of engagement (Ruiperez-Valiente et al., 2020), exploration (Ruipérez-Valiente et al., 2019), and even affective states (Kai, 2015).

Despite these successful applications, little is known about how data visualizations of game analytics should be designed to fully actualize the unique pedagogical affordances of games (e.g., Plass et al., 2015). Particularly, the ecological validity of LA/EDM driven metrics in classrooms (i.e., how do teachers perceive and understand these measurements in relation to their own practices) is largely unknown. This paper presents preliminary findings from a cross-case study of two game-based learning projects where two independent research teams engaged classroom teachers as co-designers to create data visualizations for their respective, different educational games. The overarching research question that guided this cross-case study is: "*What did teachers want to know about students' learning, and how did the co-design activities elicit those insights for the design team*?" In this paper, we report (1) what co-design activities the teams used for what purposes and goals and (2) discuss findings related to what kinds/types of analytics that teachers requested for what goals.

2. Relevant Literature

Well-designed games can work as effective assessment (Shaffer & Gee, 2012). The researchers who advocate for game-based learning have recognized that the link between learning analytics for games and teachers' use of them has been largely disconnected (Kim & Ifenthaler, 2019). Because of this need, teacher dashboards have proliferated in the past decade, exploring a wide variety of approaches. For example, the MiGen project consisted of a suite of tools designed around student tracking, classroom dynamics, and goal achievements (Mavrikis, Gutierrez-Santos, & Poulovassilis, 2016). However, critics argue that many of the teacher dashboards for games still miss the mark on balancing complexity and simplicity in a way teachers can use (Duncan, Hoxie, & Miklasz, 2016). One potential reason for this challenge is that researchers' questions might not represent teachers' needs; thus, we might be developing tools that teachers do not really want, need, or understand how to utilize. Indeed, a

systematic literature review by Jivet et al. (2018) found that very few dashboards use pedagogical decisions as the input for their design.

In summary, to support teachers' use of analytics from educational games in the classroom, the first step should include research on understanding what kinds of analytics teachers need and for what pedagogical and classroom management purposes. This step is particularly of importance in game-based learning because students learn and demonstrate their learning differently than more traditional modes of learning in schools (e.g., complete a quiz). Thus, the design teams highlighted in this case study sought to address this gap in the research by working directly with teachers in a co-design process to identify their needs as relevant to the unique affordances and pedagogical utility of each game.

3. Cases

Team A: Shadowspect is an online puzzle game where players solve geometry puzzles by selecting, placing, and rotating 3D geometric primitives such as cubes, cylinders, spheres, etc. based on orthogonal views of a figure. In 2020, the research team implemented a 12-month co-design process with 8 practicing teachers (called design fellows) to develop additional analytics and data visualizations while iterating some existing models (since the game was initially developed as an assessment tool).

Team B: Lakeland is a web-based strategic, city-builder game intended for use in secondary science classrooms (grades 7-10) to teach students about nutrient flow through natural and human systems. The game has students manage a lakeside town by producing enough food and ensuring fields stay fertilized while avoiding algae blooms. The design team used early analytics as a starting point to engage the teachers through a 2-day co-design workshop in Spring 2020 where they collaborated to design "dream dashboards." The team then built functional prototypes for testing over several months.

4. Method

Data sources for this paper include co-design sessions transcriptions, artifacts generated by the teachers (e.g. definitions for metrics, hand-drawn diagrams for data visualizations), and research teams analytics memos. Two researchers, one from each team, met after the co-design workshops ended to conduct thematic analysis (Braun & Clarke, 2012) by reviewing all the co-design activities, memos, and transcriptions. The guiding question throughout the process was, "*What did the teachers want to know about students' learning, and how did the co-design activities elicit those insights for the design team*?"

5. Findings

5.1. Similarities and differences in co-design activities

Both teams had three goals for codesign activities that eventually led to data visualizations using game analytics: (1) to elicit what teachers value and wish to know from gameplay, (2) to operationalize these values, constructs, and design metrics/indicators and (3) to provide insights for data visualizations. We discuss two examples, one from each team, to illustrate how co-design activities were used.

Team A utilized a variety of generative design tools/methods adapted from general design tools/methods (e.g. User Personas) to elicit deep values and desires related to using games in classrooms (i.e. what kinds of skills and attitudes do teachers wish to better promote by using games?). For example, Team A used the scenarios and storytelling method (Kumar, 2012) to ask the design fellows to identify meaningful sequences from the gameplay data using a sequence mapping tool (Figure 1). Each icon (e.g., eyes, cubes) represents a game event (i.e. Eyes mean "the player used the perspective tool) that the team visualized in advance to be able to reconstruct game play. The label, "Take another look and another, and another," as well as the text below the sequence is generated by the teachers.

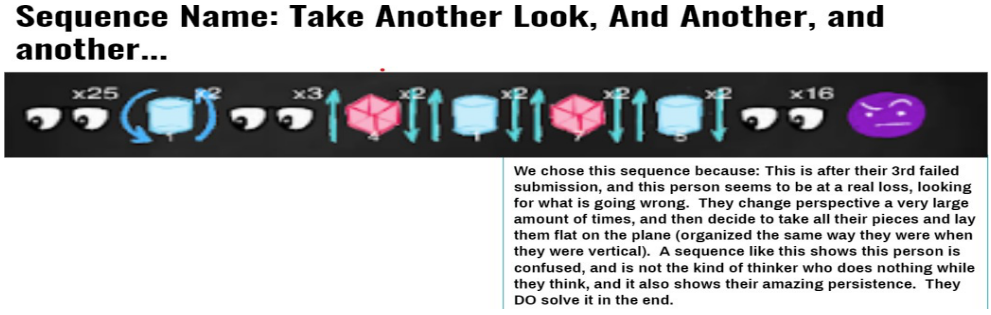


Figure 1. A sequence label created by teachers.

Team B began by having teachers observe one another playing Lakeland and recording moments of interest: any actions that evoked curiosity. These observations scaffolded teacher thinking about what they might want to "see" in their dashboards. Teachers were then provided with an activity (Figure 2) to expand on how the moments could become useful metrics for their dashboard. Each participant finally combined their metrics into a "Dream Dash" to explain the pedagogical utility of the elements.

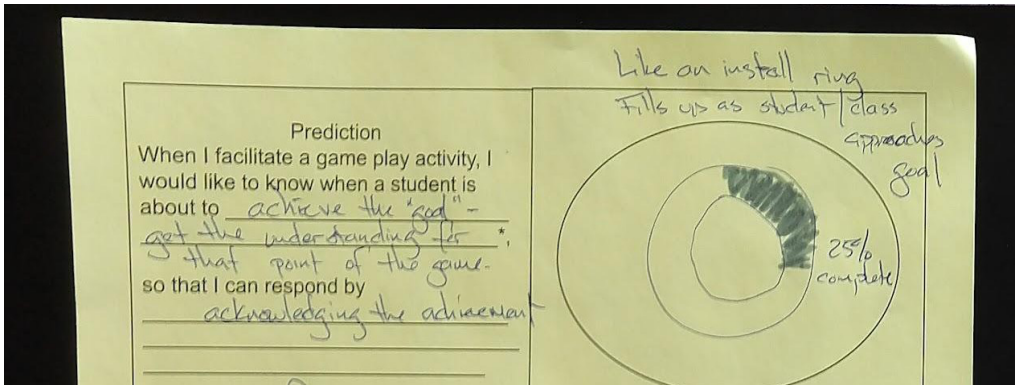


Figure 2. Scaffolding Sheet used by Team B.

5.2. Types of Teacher Needs

The following section describes four themes that emerged from comparing teachers' expressed needs across the two cases and the pedagogical implications associated with each theme.

Theme 1: How do I know how the student is progressing?

The design fellows for Team A discussed ways of thinking about progress. They expressed a strong need for diversifying achievements beyond completing levels: knowing what the overall pattern of progression is for different levels--classroom (e.g. on average, how many puzzles have the class worked on so far?), small groups based on filters (e.g. students who haven't logged into the game in last 2 days), and individual students (e.g. How many puzzles has Marisol tried and completed so far?). Transparency of these progress metrics to the students was also valued by the fellows to keep students motivated.

The fellows also recognized that there are many nuances about progress; therefore, they had a hard time defining what counts as evidence for progress. For example, fellows expressed concerns about students who are quickly solving the puzzles and "flying through" because this might mean that they are not learning anything new from the playing game. A design fellow said, *"So, I look at it from the perspective of, if they're just getting it right, and they're flying through, then they're not really learning, they already know it.* […] *That sweet spot of learning is where you struggle, learn from your mistake, and then now you've learned something, and now you can progress."*

Related to this notion that you can't really learn in the game unless you've made mistakes, the teachers expressed the need to know if/how students "revisit" problems even if they didn't solve them right away. A fellow from Team A said, "*Yeah, I almost wonder if, also, like ... So, say a student is on Puzzle B, they go on for five minutes, get stuck, you know, leave, but then go back into Puzzle B. And seeing if you compare the time and the actions of that first experience from the second and third. And so, like, if there's any progress."* They further elaborated a need to understand the characteristics of puzzles (e.g., difficulty) to make meaning with progress. For example, if a student is only solving easy puzzles, then the teachers can gently nudge them to try more challenging ones.

One of the challenges Lakeland presented for teachers is that student progression appeared to be nonlinear, allowing students to move towards any one of four defined (i.e. growing their town vs. growing their food production) and numerous undefined goals. One teacher reported that their student had kept their farm relatively small, so they could see how long it would sustain itself. Traditional progress metrics would likely flag this student as not progressing, but their ability to balance elements of their town demonstrated their mastery of the underlying system. Several teachers articulated a desire to capitalize on their own intuition, knowledge of their students, and ability to understand the game interface/actions by requesting a way to "see" what students had added to their town; visualizations such as a mini-map for each student or a linear counter to show buying activity allowed teachers make their own inferences about student growth.

In summary, both groups of teachers wanted to have analytics and data visualizations that help them to (1) flexibly define what progress means and what types of actions are being considered as evidence, (2) visualize how students are progressing, both linearly and nonlinearly, (3) visualize progress based on different types of achievement, so the students can see them.

Theme 2: What should we celebrate even if it wasn't successful?

Teachers who want to use games in the classroom are more likely to value and understand that there are multiple ways of solving problems. Teachers desire to better understand the intricacies of student thinking, so they can notice and celebrate it, even if the student is not "successful" (i.e., completed all tasks and puzzles). This need arises because games allow students multiple points of entry to demonstrate learning and as such motivate those who are not always academically successful, ultimately having a positive impact on learning (Karakoç et al., 2020).

One of Team A's design fellows considered the meaning behind achievements and success saying, *"Okay. So, let's see. So we picked ... So, if we just call this Achievements, just to give it, like, a measurement kind of thing, like measuring your achievements, like measuring your badges, I guess. Or we could call them Successes, like we talked about before, how it's not just your success of, like, answering the puzzle, but all these different things."* They were struggling with the tension of wanting to recognize more than one narrowly defined term could capture. Fellows also openly expressed the need to see how their less well-performing students shine through the game, *"...So I really don't necessarily care about kids who are doing well. I do like your use case, where if there is someone who has a really creative strategy, I would love for them to be able to share out with the class, as a great example of another way to approach it."*

Similarly, the teachers emphasized the need to recognize and celebrate "efforts" and "persistence": "*Revisiting something that you struggled with and trying to figure out what your misconceptions were, or where you were incorrect. And I think revisiting is definitely something that could be interesting in terms of a data collection, as well as just showing student or a teacher that this student bailed on this puzzle. But at the same time, two days later, they returned to it, and you know, that's good. We should give a shoutout or something to that.*" When it came to visualizing persistence, however, the teachers were less interested in knowing who's most persistent, instead valuing the description of different groups, i.e., persisting productively vs non-productively, so they can intervene.

Like Team A teachers, Team B teachers were quick to acknowledge that there were many creative ways to play the game. Teachers talked about "showcasing student solutions to the class" and "shouting out to the whole class" as actions that they would want to take in response to students achieving milestones, demonstrating proficiency, or even having a "lightbulb moment". In their Dream Dash, some teachers included displays for the SmartBoard to allow students to see others be recognized by the system. The integration of the dash into actionable information for both teacher and student indicates teachers thought of metrics as being useful to informing pedagogical decisions.

One question teachers came back to several times is "Who should be talking to who?" As Team B teachers considered whether metrics could detect and explain a student's strategy, they noticed potential to facilitate partnerships for collaborative discussion. If they could be alerted that two students were demonstrating opposing strategies, the teachers felt they could facilitate a unique instructional moment if those two students discussed progress. Similarly, when a student identifies a particularly effective strategy, potentially indicated by overall student success, other students who are struggling or frustrated may benefit from pairing temporarily with them to hear how they developed their strategies.

Furthermore, there is one moment that was important to teacher practice that moved away from the focus on individual students and instead centered around the entire class. Team A teachers wanted to know when the majority of students were confused, so they could stop instruction and reteach the directions or provide scaffolding. Similarly, one teacher from Team B wanted to add a "Whole Class" view to their Dream Dash, so they could have both the in-depth analysis of an individual student as well as the birds' eye view of the class.

Theme 3: How do I know if my student is actually learning something?

As teachers became engaged with thinking about metrics and dashboards, the most common request centered on progress as performance measures. Team B teachers wanted to know simple counts of how many things players had within their towns as a way of probing the complexity (or progress) of their system. Initial visualizations of these metrics included many different forms of progress bars, but one

teacher conceptualized progress rings as being moving targets for varying parts of the game. As the player fills up this ring, the teacher could acknowledge the achievement, and the ring would reset for the next challenge.

Fellows from Team A also had ongoing questions and concerns about how they can differentiate "learning math" from "progressing" in the game. In contrast to their attitudes about other metrics, the fellows were consistent about how students' performance related to math standards should be viewed using the mastery of learning framework, therefore, wanted to know "how close the student is mastering a standard." For example, a fellow even suggested having a tier system where, *"...And one thing that I found interesting with the idea of like mastery, is there are other games I've played where, when you complete a level, it kind of like does a quick grade of how well you did it, and it's like one to three stars, right? And it's just interesting that, for Shadowspect, it's basically, if you get it right, they're just like, 'Cool, you've finished.'"*

When it comes to standards, the fellows were also more comfortable with more traditional visualizations, such as a bar graph. A fellow expressed, *"...like, puzzles have certain aspects of standards. And so, like, just completing a puzzle, you're making one step closer to mastering that Common Core Standard. And so, it's like, that's kind of more of like a completion, but it is still growth, right? Like, as they're going through and completing levels, they are getting closer and closer to that mastery bar."*

Theme 4: How do I know my students' common errors and misconceptions?

The design fellows in Team A expressed the need to understand what common errors the students are making, and how these errors are indicative of their underlying mathematical misconceptions. First, the fellows suggested that they might benefit from having a video replay function embedded in the dashboard to filter out these students and review what errors they are making. However, with a prototype that included the video replay functionality, the fellows realized that they might not need that much information at individual student level. A fellow said, *"Because of this puzzle like that information to have quickly, at my fingertips would be good. But I don't know that I need the detail of how many students made [errors] for each individual type. I just need to know for my instruction, which types I should review as a class, if it seems that cross-section errors on that puzzle are an issue."*

Team B teachers voiced a desire for action-based metrics that might indicate misconceptions. For example, one teacher requested a monitoring metric for knowing when students had multiple algae blooms, so they could respond by, "*asking is they understand the relationship of what is causing… If they have an idea how to prevent."* Similarly, another teacher wanted to detect when students "*placed fertilizer after a washout,"* so they can respond by, "*discussing how they think that cause & effect rela[tionship] will affect their town."* Teachers wanted to know specific moment-to-moment actions instead of a broader label like "Nutrient Flow Misconception".

6. General Discussion

Using a co-design process, both teams sought to identify what teachers believe to be meaningful, pedagogically relevant features for a data dashboard. From the cross-case examination, we uncovered four themes that can assist in informing others wishing to undertake similar tasks. We acknowledge that there is often tension in the co-design process between the end user (teachers) and the design/development team; creating dashboards and supporting analytics is no exception due to its technical nature. Similar efforts in the community documented similar challenges regarding eliciting meaningful input from teachers to create analytics and dashboards (Prieto-Alvarez, Martinez-Maldonado, & Anderson, 2018). However, by undertaking co-design as a sense-making and skill-building activity, the design teams were able to elicit useful feedback and insights into teaching practice that they otherwise would not have been privy to.

While not fully investigated in this paper, we speculate that this is a challenge fundamentally related to teachers' assessment literacy. Playful learning environments offer rich and continuous interactions with the environments, and in turn, complex data about not just what students know, but how they do it (Dicerbo, Shute, & Kim, 2017). Many teachers do not have experience using these forms of assessment or considering the volume of data that can be generated from them. Thus, the work of a design team is closely linked to thoroughly grasping teachers' motivations for data (e.g. "I want to know who is stuck or falling behind to penalize them vs. "I want to know who is stuck, so I can provide useful help."). The co-design process must present metrics/models in the context of teaching practice and

should be closely tied to the actions that teachers want to take based on that information. Partnering with teachers to make meaning of their practice with respect to assessment more holistically, such as providing scaffolded prompts like Team 2's Fill in the Blank activity, can assist designers in better meeting the teacher's needs. Knowing what teachers want to know without knowing what they want to do with the information is not going to be helpful.

While preliminary, we could identify common themes, in terms of how different co-design activities are utilized by the teams, and what teachers need to know from analytics from educational games to fully amplify pedagogical affordances of games. In the future work, the team will continue to analyze the data to yield a set of generalizable design principles (Fu, Yang, & Wood, 2015) in terms of data visualization with specific attention to what actions and decisions that teachers should/need to make when they are implementing games in classrooms.

Acknowledgements

This material is based upon work supported by the National Science Foundation (STEM+C #1935450). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Alonso-Fernández, C., Martínez-Ortiz, I., Caballero, R., Freire, M., & Fernández-Manjón, B. (2020). Predicting students' knowledge after playing a serious game based on learning analytics data: A case study. *Journal of Computer Assisted Learning, 36*(3), 350-358.

DiCerbo, K., Shute, V., & Kim, Y. J. (2017). The future of assessment in technology rich environments: Psychometric considerations. Learning, design, and technology: An international compendium of theory, research, practice, and policy, 1-21.

Fu, K. K., Yang, M. C., & Wood, K. L. (2015, August). Design principles: The foundation of design. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Vol. 57175, p. V007T06A034). American Society of Mechanical Engineers.

Gomez, M. J., Ruipérez-Valiente, J. A., Martinez, P. A., & Kim, Y. J. (2020, October). Exploring the Affordances of Sequence Mining in Educational Games. In Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality (pp. 648-654).

Kai, S., Paquette, L., Baker, R. S., Bosch, N., D'Mello, S., Ocumpaugh, J., ... & Ventura, M. (2015). A Comparison of Video-Based and Interaction-Based Affect Detectors in Physics Playground. *International Educational Data Mining Society.*

Karakoç, B., Eryılmaz, K., Özpolat, E. T., & Yıldırım, İ. (2020). The effect of game-based learning on student achievement: A meta-analysis study. *Technology, Knowledge and Learning,* 1-16

Karumbaiah, S., Baker, R. S., Barany, A., & Shute, V. (2019, October). *Using epistemic networks with automated codes to understand why players quit levels in a learning game. In International Conference on Quantitative Ethnography* (pp. 106-116). Springer, Cham.

Kim, Y. J., & Ifenthaler, D. (2019). Game-based assessment: The past ten years and moving forward. In *Game-Based Assessment Revisited* (pp. 3-11). Springer, Cham.

Kumar, V. (2012). *101 design methods: A structured approach for driving innovation in your organization*. John Wiley & Sons.

Mavrikis, M., Gutierrez-Santos, S., & Poulovassilis, A. (2016, April). Design and evaluation of teacher assistance tools for exploratory learning environments. In Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 168-172).

Owen, V. E., & Baker, R. S. (2020). Fueling Prediction of Player Decisions: Foundations of Feature Engineering for Optimized Behavior Modeling in Serious Games. *Technology, Knowledge, and Learning, 25*(2), 225–250.

Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist, 50*(4), 258-283.

Prieto-Alvarez, C. G., Martinez-Maldonado, R., & Anderson, T. D. (2018). *Co-designing learning analytics tools with learners. In Learning Analytics in the Classroom* (pp. 93-110). Routledge.

Ruiperez-Valiente, J. A., Gaydos, M., Rosenheck, L., Kim, Y. J., & Klopfer, E. (2020). Patterns of engagement in an educational massively multiplayer online game: A multidimensional view. *IEEE Transactions on Learning Technologies, 13*(4), 648-661.

Shaffer, D. W., & Gee, J. P. (2012). The right kind of GATE: Computer games and the future of assessment. Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research, 211-228.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Changing Students' Perceptions of a History Exploration Game Using Different Scripts

Stefan SLATER [a*], Ryan S. BAKER [a], David GAGNON [b], Erik HARPSTEAD [c], Juliana Ma. Alexandra L. ANDRES [a], Luke SWANSON [b]

[a] *University of Pennsylvania, United States of America*
[b] *University of Wisconsin-Madison, United States of America*
* slater.research@gmail.com

Abstract: As digital games and simulations continue to see use in both formal and informal learning contexts, it becomes increasingly important to understand the goals, motivations, and interests of the learners using them. Recently, there has been increased interest in identifying the different engagement profiles of learners playing games, descriptions of groups of learners' tendencies and preferences when playing digital games. These engagement profiles represent a powerful tool for designers looking to create personalized and adaptive learning environments. In this work, we explore two aspects of these engagement profiles. First, we explore the different profiles of engagement in a large sample of elementary and middle school learners playing the history game *Jo Wilder and the Capitol Case* (2019) in a variety of contexts, using the data from a self-report survey. Second, we explore the effect that minor game dialogue changes had on learners' perceptions of different game elements, such as character likeability and humor, as well as personal preferences, such as enjoyment of history. These findings highlight opportunities for personalized and adaptive game design that leverage players' goals and motivations.

Keywords: Games, engagement profiles, typology, history education

1. Introduction

Games are powerful tools for learning and see ever-increasing use in many different educational contexts. Games and other interactive digital environments are used in both physical and digital classroom environments (Obedoza & Sison, 2020), in after-school programs (Pusey, Wong & Rappa, 2020), and in informal learning contexts such as museums (Kumar, Tissenbaum & Berland, 2017). Games have been shown to have positive learning outcomes in a wide variety of educational settings and domains, such as language learning (Culbertson et al., 2016), physics (Sun et al., 2022), and geometry (Ruiperez-Valiente & Kim, 2020). Part of the appeal of using games for learning lies in the way that they allow students to have experiences and do activities that are difficult to do in more traditional forms of media, such as assuming the role of urban planners (Metcalf, Clarke, & Dede, 2009; Foster et al., 2018) or participating in simulations of authentic science practices (Gaydos & Squire, 2010).

Players in educational games have a wide variety of strategies, goals, and motivations for play (Yee, 2006; Bonny & Castaneda, 2022; Kordyaka et al., 2019). These differences between players may influence both engagement with the *context* of the game, such as when and where players like to play games, as well as their engagement with the *content* of the game, such as whether a student playing a science game is interested in science as a subject. Examining and identifying students' game preferences provides a window into the experiences that they are having during gameplay. Recent research examining these engagement profiles suggests considerable variance in how different groups of students play the same game. For example, Ruiperez-Valiente et al. (2020) explored how students engaged with the STEM MMO *The Radix Endeavor*. Using data collected from students' interactions with the game itself, they constructed multiple metrics of players' gameplay engagement. By clustering learners' actions across these metrics, they identified four overall groups of players in the game: integrally engaged learners, who scored high across all recorded metrics; lone achievers, who scored low in social metrics but high in completion metrics and overall playtime; non-engaged learners, who scored low across all metrics; and social explorers, who scored high in social measures such as chatting, but scored low in quest completion and game progress.

Whether students' states (their responses to the game and the context in which it is being played) or traits (preferences of the player that persist across games and contexts) give rise to these engagement profiles is not currently well-understood, though there is evidence that patterns vary considerably between different groups of learners (Tabanao & Rodrigo, 2016). If engagement profiles were better understood, they could be an opportunity for implementing personalization in games, as the method of personalization should depend on the qualities of the learner during play.

In this paper, we explore how players' perceptions of a game, its characters, and its subject matter, relate to salient and easily modifiable features of the game environment. In doing so, we aim to explore the following research questions:

RQ1: What different types of student players exist in our game?
RQ2: Do changes to game dialogue change groups' perceptions of gameplay elements?

2. Methods

2.1 Game and Dataset

Data were collected from the point and click history adventure game *Jo Wilder and the Capitol Case*. In *Jo Wilder,* players control Jo, a pre-teenage girl who is attempting to help her historian grandfather determine the story of two artifacts that will be displayed in a new exhibit at the Capitol. She does so reluctantly, motivated primarily by her dislike of another historian, Wells, who is unkind to her grandfather and who she suspects kidnapped her pet badger. Players are free to move throughout the locations, character conversations and inspectable items freely, but the game's narrative structure moves forward linearly as clues are discovered.

2.2 Survey Development

We developed and piloted a 16 item Likert survey embedded within the game to more closely examine levels of student enjoyment of the game and attitudes towards history. This survey measured four aspects of play, including (1) how immersed are you in the game, adapted from the E-Game Flow Scale (Fu, Su, & Yu, 2009), (2) how much do you like the characters, adapted from the Reysen Likability Scale (Reysen, 2005), (3) how funny do you think the characters are, adapted from the Multidimensional Sense of Humor Scale (Thorson & Powell, 1993), and (4) what are your attitudes towards history, adapted from the Three-Dimensions of Student Attitude Towards Science (TDSAS; Zhang & Campbell, 2011). Students completed four-question item banks, drawing one item from each construct, at four points over the course of gameplay. At the beginning of gameplay students also answered two questions regarding their reading level, as we believed that the text-heavy nature of Jo Wilder could affect the enjoyment and accessibility of the game for strong readers versus weak readers.

2.3 Script Development

For this study, different versions of the game script were developed to make the lead character either more vocally reluctant and "snarky" (critical and sarcastic) or more enthusiastic and positive toward history, and to either introduce jokes and wordplays into the dialog between characters to make it funnier or be more serious, avoiding humor. These scripts were developed by first developing the snarky and funny version, then systematically removing each of those elements. Script changes altered between 20% - 40% of the total dialogue in the game, depending on the condition.

Table 1. *Matrix of different Jo Wilder scripts developed, showing presence of snark and/or humor*

	Snark	Humor
Initial Design: Snark and Humor	X	X
Script 1: No Snark		X
Script 2: No Humor	X	
Script 3: No Humor, No Snark		

2.4 Typology Development

We used Latent Class Analysis (LCA; McCutcheon, 2002; Muthen, 2004) to identify different groups of players based on their responses to the game feedback items. LCA is a method of statistical analysis that identifies one or more latent classes within a broader set of data. These latent classes represent subpopulations of individuals, and membership in a subclass is based on the patterns of variance among an individual's observed variables. Analyses were performed in the statistical software package *MPlus*. Following current recommendations from the literature, we began by fitting a k=2 model, hypothesizing that two subclasses existed within our dataset. We assessed the fit of this model, then increased the number of hypothesized subclasses, stopping when model fit no longer substantially improved (Zhou & Bowers, 2020; Nylund-Gibson, Grimm, & Masyn, 2019). We included the game script that students experienced as a covariate in the model, using the 3-step procedure in MPlus (Nylund-Gibson & Choi, 2018) to prevent the covariate from influencing the structure of the indicators on the model.

3. Results

There are several metrics that can be used to assess goodness of fit of an LCA model, and there is no one metric that is identified as the consensus best within the field. Instead, multiple metrics are used and compared in order to determine the correct model to interpret (Jung & Wickrama, 2008). In this work we followed guidelines from Samuelsen & Raczynski (2013) and Nylund-Gibson & Choi (2018) to select the model which we interpret. The fit statistics we considered are listed in Table 2; we chose to interpret the k=5 model due to it being the last model for which the Vuong-Lo-Mendell-Rubin statistic was significant (Lo, Mendell, & Rubin, 2001).

Table 2. *Summary of model results for fitted models. The model that we used is in* ***bold***.

Model	AIC	BIC	VLMR	BLRT	Entropy
K=2	102882	103166	< 0.01	< 0.01	0.847
K=3	100785	101167	< 0.01	< 0.01	0.881
K=4	98940	99421	< 0.01	< 0.01	0.883
***K*=5**	**97602**	**98181**	**0.048**	**< 0.01**	**0.861**
K=6	97166	97166	0.730	< 0.01	0.878

3.1. The k = 5 Model

A plot of the group means by variable for the k=5 model is shown in Figure 1. The model consists of five groups: (1) players who disliked most aspects of the game, but especially history (red, N = 184); (2) players who didn't like the characters (Jo in particular, blue, N = 546); (3) players who liked the game but didn't like history (green, N = 489); (4) players who liked both history and the game (pink, N = 665); and (5) players who liked the game, but didn't think it was very funny (brown, N = 533). Respectively, we refer to each of these groups as the "Didn't Like Anything" group (red), "Didn't Like Jo" group (blue), "Didn't Like History" group (green), "Liked It All" group (pink), and "Not Very Funny" group (brown).

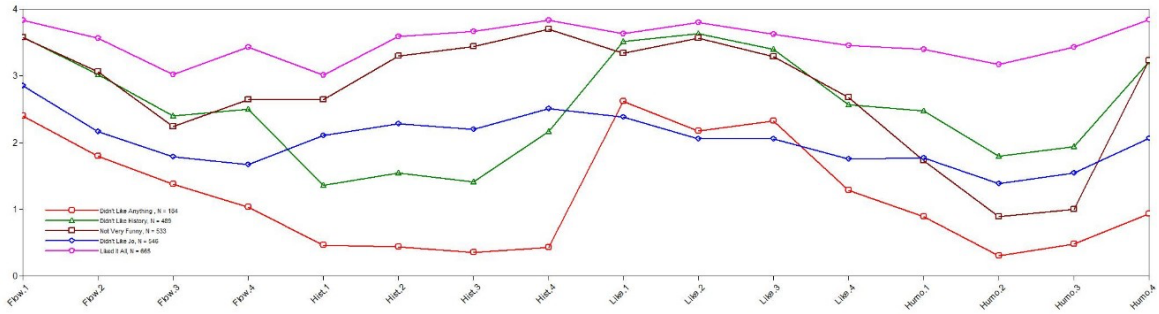


Figure 1. Plot of the sample means for survey items by group.

Once we established the structure of the latent class model, we were able to measure the influence of script type and reading level on students' latent class membership. We followed current

recommendations from the literature for the 3-step procedure (Nylund-Gibson & Choi 2018, MPlus Webnote 15 Appendix A) to examine the effect of our covariates, without allowing them to influence the structure of the model itself. A Benjamini & Hochberg post-hoc adjustment (Benjamini & Hochberg, 1995) was used to control for multiple comparisons – however, because this is an exploratory study, we also report differences that were initially statistically significant but did not remain significant after post-hoc adjustment.

3.2. Script Effects on Player Groups

The effects of different scripts on player evaluations are summarized in Table 3. We found no differences between the proportions of players placed into different groups when comparing the No Humor, No Snark script to all other scripts. We also found no differences between the proportions of players placed into different groups when comparing the Initial Design script to all other scripts.

For the Humor No Snark script, the only difference that we found was between the Liked It All and Didn't Like Jo groups of players. For this script, there was a trend suggesting that players were more likely to Like it All, and less likely to Not Like Jo, but this trend did not hold up after a post-hoc correction was applied.

Table 3. *Summary of differences in proportions between scripts.*

-- more likely to be in… --	-- less likely to be in… --	-- when playing… --	Odds ratio
Liked It All	Didn't Like Jo	Humor No Snark	1.519
Didn't Like Jo	Didn't Like History	Snark No Humor	1.660
Didn't Like Jo	Liked It All	Snark No Humor	1.820
Liked It All	Not Very Funny	Snark No Humor	1.506

For the Snark No Humor script, we found three differences between the proportions of groups of players in this script, compared to the other scripts. We found that players were more likely to be in the Didn't Like Jo group, and less likely to be in either the Didn't Like History group or the Liked It All group. We also found that there was a trend for players to be more likely to be in the Liked It All group than the Not Very Funny group compared to other scripts, though again this trend was not quite significant when a post-hoc correction was applied.

3.3. Reading Level Effects on Player Groups

We found strong differential effects of reading level on players' class, which are summarized in Table 4. As reading level increased, players were less likely to be in the "Didn't Like Anything" group, and more likely to be in any other group. Players with a higher reading level were also more likely to be in the "Liked It All" group.

Table 4. *Summary of odds ratios for all reading level effects.*

-- more likely to be in… --	-- less likely to be in… --	-- when a… --	Odds ratio
Liked It All	Didn't Like Any	Strong Reader	2.362
Didn't Like History	Didn't Like Any	Strong Reader	1.456
Didn't Like Jo	Didn't Like Any	Strong Reader	1.377
Not Very Funny	Didn't Like Any	Strong Reader	1.625
Liked It All	Didn't Like History	Strong Reader	1.622
Liked It All	Didn't Like Jo	Strong Reader	1.715
Liked It All	Not Very Funny	Strong Reader	1.454

4. Discussion

In this work, we found evidence for five different groups of players in the game Jo Wilder, based on their self-reported responses to a four-construct, 16-item survey bank. Two groups of players either "Liked It All" (n = 665) or "Didn't Like Any Of It" (n = 184), providing consistently high (or low) ratings of the game, its characters, and enjoyment of history, respectively. In between these two groups we found "Didn't Like History" (n = 489), "Didn't Like Jo" (the game's main character; n = 546), and

"Not Very Funny" (n = 533) groups of players, who generally provided average responses to the survey items, except for specific categories that they did not rate themselves as enjoying. Overall, however, students appeared to enjoy the game experience, with the largest overall group being "Liked It All" (28% of all players) and the smallest group being "Didn't Like Any" (7% of all players).

At present, there is open debate as to whether game typologies represent player *states* – the way that they feel about the activity in the present moment – or player *traits* – relatively fixed attitudes or predispositions towards these types of activities in general. In this work, we have found evidence for both of these hypotheses. Changing the presence of humor and snark in the game scripts did not change the proportions of players in the "Didn't Like Anything" group, suggesting that their attitudes towards Jo Wilder could be based on pre-existing traits related to educational games, or to history. These players may dislike history as a subject, or not enjoy narrative-focused games.

However, in the "Not Very Funny", "Didn't Like History", and "Didn't Like Jo" scripts, we found evidence that players' attitudes (and latent class membership) may be more malleable, and therefore more driven by learners' current states. By removing either humor or snark from the base game script, we found that students' evaluations of Jo as a character changed, as well as their evaluations of history learning and of the game. We believe that this may be because the 'sense of humor' in Jo Wilder comes at the expense of Jo's attitudes towards education. In the base game's script, Jo is not a model student and doesn't particularly care about school, and the humor and snark in the game scripts draw on this character trait. In the Initial Design script, where both humor and snark are present, Jo's character traits could have reinforced players' existing negative attitudes towards history learning. When either of these character traits were dropped from the script, players were more open to positive evaluations of their own history learning. However, these changes did come with a cost – our findings suggest that a "snarky" Jo may have come across to some players as being a mean character, and worsened player attitudes towards her as a character.

For reading level, players with higher self-reported reading level enjoyed the game more as an overall trend, observed across all groups. This finding underscores the importance of reading comprehension and understanding of game content for the employment of games in classroom environments. When designing curricula and pedagogy that utilizes games and digital environments, care needs to be taken to ensure that students will be able to read and understand the content of the game itself, and that its instructions and narratives are worded at an appropriate level for, and are culturally responsive to, the students with whom the game will be used.

Towards the development and refinement of adaptive learning systems, our findings suggest that some players' attitudes towards both game enjoyment and content topics can be substantially influenced by changing the dialogue and tone of the game world. In Jo Wilder, modifying game dialogue was a relatively easy change to make in the game's architecture. Interventions such as this one may then hold promise for improving student attitudes towards desired content topics, or promoting engagement in learning tasks based on players' characteristics. Salient qualities of a game, such as characters' appearances, mannerisms, or senses of humor, could be modified to suit the particular individual playing the game, towards optimizing their engagement with the overall learning activity. These potential changes represent powerful tools for improving players' capacity to engage with games as learning tools, and strengthen learning outcomes for diverse groups of learners.

Acknowledgements

We would like to acknowledge NSF #DRL-1907437 for funding this project, and the Field Day Lab team for the use of their data architecture and guidance in data processing.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289-300.

Bonny, J. W., & Castaneda, L. M. (2022). To Triumph or to Socialize? The Role of Gaming Motivations in Multiplayer Online Battle Arena Gameplay Preferences. *Simulation & Gaming*, *53*(2), 157-174.

Culbertson, G., Andersen, E., White, W., Zhang, D., & Jung, M. (2016). Crystallize: An immersive, collaborative game for second language learning. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 636-647.

Foster, A., Shah, M., Barany, A., Petrovich Jr., M.E., Cellitti, J., Duka, M., Swiecki, Z., Siebart-Evenstone, A., Kinley, H., Quigley, P., & Williamson Shaffer, D. (2018). Virtual learning environments for promoting self transformation: iterative design and implementation of Philadelphia land science. In *International Conference on Immersive Learning*, 3-22.

Fu, F. L., Su, R. C., & Yu, S. C. (2009). EGameFlow: A scale to measure learners' enjoyment of e-learning games. *Computers & Education*, *52*(1), 101-112.

Gaydos, M., & Squire, K. (2010). Citizen science: Designing a game for the 21st century. In *Interdisciplinary Models and Tools for Serious Sames: Emerging Concepts and Future Directions*, 289-305.

Jung, T., & Wickrama, K. A. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302-317.

Kordyaka, B., Jahn, K., Müller, M., & Niehaves, B. (2019). The Comparative Self: Understanding the Motivation to Play and the Subsequent Video Game Use. In *DiGRA Conference*.

Kumar, V., Tissenbaum, M., & Berland, M. (2017). What are visitors up to? Helping museum facilitators know what visitors are doing. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference,* 558-559.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767-778.

McCutcheon, A. L. (2002). Basic concepts and procedures in single-and multiple-group latent class analysis. *Applied Latent Class Analysis*, 56-88.

Metcalf, S. J., Clarke, J., & Dede, C. (2009). Virtual worlds for education: River City and EcoMUVE. In *MiT6 International Conference*, 1-6.

Muthén, B. (2004). Latent variable analysis. *The Sage Handbook of Quantitative Methodology for the Social Sciences*, *345*(368), 106-109.

Nylund-Gibson, K., Grimm, R. P., & Masyn, K. E. (2019). Prediction from latent classes: A demonstration of different approaches to include distal outcomes in mixture models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 967-985.

Nylund-Gibson, K., & Choi, A. Y. (2018). Ten frequently asked questions about latent class analysis. *Translational Issues in Psychological Science*, *4*(4), 440.

Obedoza, J.B., & Sison, R. (2020). Design of a Game-Based Intelligent Learning Environment to Remediate Fraction Addition/Subtraction Misconceptions. *Proceedings of the 28th International Conference on Computers in Education*.

Pusey, M., Wong, K.W., & Rappa, N. (2020). The effect of a more knowledgeable other on resilience while playing single-player puzzle video games. *Proceedings of the 28th International Conference on Computers in Education.*

Reysen, S. (2005). Construction of a new scale: The Reysen likability scale. *Social Behavior and Personality: An International Journal*, *33*(2), 201-208.

Ruipérez-Valiente, J. A., & Kim, Y. J. (2020). Effects of solo vs. collaborative play in a digital learning game on geometry: Results from a K12 experiment. *Computers & Education*, *159*.

Ruiperez-Valiente, J. A., Gaydos, M., Rosenheck, L., Kim, Y. J., & Klopfer, E. (2020). Patterns of engagement in an educational massively multiplayer online game: A multidimensional view. *IEEE Transactions on Learning Technologies*, *13*(4), 648-661.

Samuelsen, K., & Raczynski, K. (2013). Latent Class/Profile Analysis. *Applied Quantitative Analysis in Education and the Social Sciences*, 304.

Sun, C., Shute, V.J., Stewart, A.E.B., Beck-White, Q., Reinhardt, C.R., Zhou, G., Duran, N., & D'Mello, S. (2022). The relationship between collaborative problem solving behaviors and solution outcomes in a game-based learning environment. *Computers in Human Behavior*, *128*, 107-120

Tabanao, E., Rodrigo, M. M. T. (2016). A comparison of the experience of confusion among Filipino and American learners while using an educational game for physics. In *Proceedings of the International Conference on Computers in Education* (pp. 17-20).

Thorson, J. A., & Powell, F. C. (1993). Sense of humor and dimensions of personality. *Journal of Clinical Psychology*, *49*(6), 799-809.

Yee, N. (2006). Motivations for play in online games. *CyberPsychology & Behavior*, *9*(6), 772-775.

Zhang, D., & Campbell, T. (2011). The psychometric evaluation of a three-dimension elementary science attitude survey. *Journal of Science Teacher Education*, *22*(7), 595-612.

Zhou, X., & Bowers, A. (2020). A Typology of Parental Involvement in Student Experience. *The High School Journal*, *103*(2), 99-132.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Talk & Play: a VR Application to Improve Volume and Silences Use in Verbal Communication

Maria Esther RUIZ-CAPILLAS[*], Alejandro ROMERO-HERNANDEZ[b] & Borja MANERO[c]

[*,b,c], *Complutense University of Madrid, Spain*
*mruizcap@ucm.es, [b]alerom02@ucm.es, [c]bmanero@ucm.es

Abstract: Public speaking is present everywhere in our lives; therefore, it is important to know how to do it correctly. Most people have difficulties speaking in public due to their glossophobia (fear of speaking in public). Practicing is one of the best ways for treating this fear and improving the oratory skills.
There are three basic aspects in this discipline which are not always taken into consideration when trying to get better in public speaking: silences, speed and volume. This paper presents Talk&Play, an application for teaching how to use these aspects correctly in a speech and the importance of them.
For validating this application, an experiment for evaluating it is presented as well. Three main conclusions were reached: 1) Talk&Play increases user's self-security during public speaking, 2) it also raises awareness of the importance of a correct use of silences, volume and speed in the experiment participants, 3) User learns how to use volume and silences, but the application still needs to improve.

Keywords: public speaking, virtual reality, educational videogames

1. Introduction

Public speaking has a lot of presence in our daily lives, not only when a speech is done in a congress or in an award collecting, but also in class when asking questions to the teacher, meeting new people, doing a job interview, expressing an idea in a meeting, etc. (Pratiwi et al. 2018). Having a good oratory gives the opportunity of influencing people's behavior and attracting their attention more easily (Angraini 2016). Due to it, it is very important to firmly acquire this skill, but most people (75% of world's population) have difficulties because of their glossophobia (or fear of speaking in public), one of the most anxiety related phobias (Forsythe 2013) (Tse 2005). To overcome this situation, the best option is to practice and learn how to do it correctly. Also, some studies ensure Virtual Reality helps to reduce public speaking anxiety (Lim, Aryadoust, and Esposito 2022) (Noor Farah Wan Shamsuddin, Nur Adilin Mohd Anuardi, and Syafinaz Mohmad Rozee 2022) (North, North, and Coble 1998).

Speaking in public involves many aspects to take into consideration: hand gestures, body position, face gesticulation, eye contact with the public, volume of the voice, talking speed, speech content, etc. (Chance and Luck 1975). In the literature review, we found different methods and tools created for overcoming glossophobia and for oratory improving. Some of the virtual reality videogames are focused on improvisation and creativity (Fuertes Franco, Navarro Vaquero, and Montes Larrabaster 2020), measure sonority, emotions, hands movements and visual contact (Algaba Aranda et al. 2019), lets practice a speech in front of avatars with different moods (Jos, Franco, and Pais 2017), a virtual reality system which keeps tract of the heart rate (Khurpade et al. 2020) or a virtual reality public which is emotions-responsive (El-Yamri et al. 2019).

However, when practicing oratory, the focus is usually not on the voice volume used, talking speed or silences made during the speech but in the rest of the other aspects we have mentioned before. These three aspects are very important because they can make the different between letting the public understand the expressed information or not. Attracting public's attention is one of the main points in public speaking. If the public does not understand what it is being said, is not possible it keeps concentrated on the speech. Nevertheless, we could not find any tool focused on the correct use of voice volume and silences in a speech, despite their importance in oratory.

In this paper we present our tool Talk&Play, a virtual reality game designed and developed to improve these two aspects of public speaking with a preliminary evaluation of it. This paper is structured as follows. This section presents the importance of public speaking. Section 2 describes the application we have created. Section 3 expounds the conducted experiment for knowing if the application can teach the user about the correct use of silences and volume. Finally, Section 4 shows the results and future work.

2. Tool design

Talk&Play consists in a virtual reality videogame for teaching the correct use of voice volume, speed and silences in a speech. As well, it raises awareness on the importance of these three aspects.

We used gamification to take advantage of its motivational, attractiveness and learning related beneficial powers (Díaz and Troyano 2013). On the other hand, virtual reality environments are highly immersive, what allows the user to focus only on public speaking, the feedback received and trying to improve.

The Talk&Play experience starts putting on the headset. A start menu for selecting one of the three minigames appears. It is divided into three minigames because this allows the user to improve each aspect on its own and to practice them all together (See Figure 1). This way, the learning experience is as complete as possible.

The silences minigame teaches the importance, use and correct durations of them (indirectly including the talking speed). In the first level, the correct placement of the silences is indicated and an audio example is given. The second level only keeps the audio example as help. In the third level, the user must be able to do it without any help (See Figure 2). The number of silences needed and their durations is always given. During the game, the number of silences the user does during the speech are checked. They are counted as correct silences if their duration and placements are the desired ones. This is done with phrase recognition. The game penalizes if the user does spare silences or skips any. At the end, the user receives some points and feedback.

The volume minigame teaches the importance and correct use of the volume in different sizes scenarios. In the first level, the user must talk for a small scenario, in the second level for a medium scenario and in the third level for a big scenario. The size is indicated with the approximate number of rows the cinema would have. During the game, the volume the user is speaking at is detected. If it is the desired one, it is considered as correct. At the end, the user receives some feedback.

The practice minigame allows to practice silences and volume together. During the game, the game checks silences and volume and evaluates the user in the same way the third silences level and second volume level do.

During the game, the player can find itself constantly changing between four different phases. In the first phase, the user decides the minigame and level to play. The second phase consists in learning how to use the volume and silences correctly and being able to practice it in a speech by their own. In the third phase, the player can present the prepared speech being monitored by the game. Feedback is shown on the go so the user can rectify its errors before the end. Finally, in the fourth phase, the system gives the results and the feedback to the users.

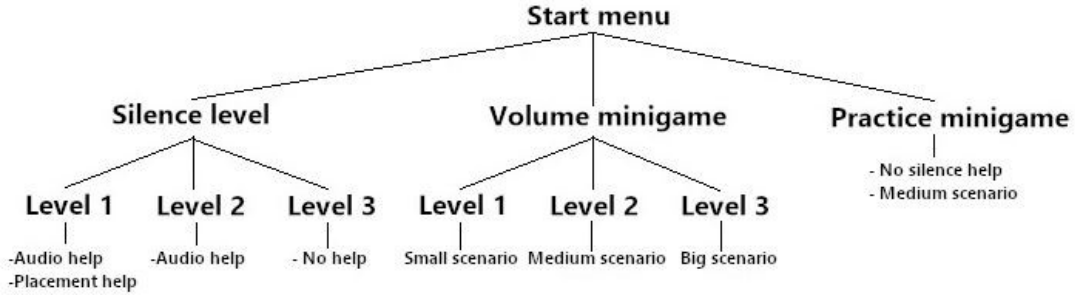


Figure 1. Talk&Play scene scheme.

The feedback received in phase three consists of the detected number of silences and if the used volume is between the asked thresholds, both in real time. The one in third phase, focuses on the whole presentation for stimulating the user to improve its final grade. Receiving two types of feedback keeps the user constantly informed about the course of the game. This helps to avoid frustration due to not understanding the game, (Díaz and Troyano 2013) and to teach the user how to do it better.

3. Experiment

The aim of this experiment was to test our application effectiveness in improving users 'verbal communication. Nevertheless, it is only a preliminary experiment.

3.1 Participants

Fourteen people between 20 and 30 years old participated in this experiment. 64.3% were men, 21.4% were women, 7.1% were gender fluid and 7.1% did not want to answer. All of them were studying or have finished university studies. Both, their initial self-security while talking in public and their videogames relation, covers the whole range (from very confident to not confident at all and from any relation to usual player). They were volunteers reached randomly between university community. Their role was to try the application and answer some questionaries about their public speaking and videogames related feelings and its perspective of the aid the application gave them.

3.2 Experimental design

The experiment consisted of making a first test with the application and collecting information about the participants and the application's behavior. For this, we asked fourteen people to participate, and we organized fourteen small intervals of 30 minutes each. Each interval was divided into 3 parts: part 1, part 2, and part 3 (See figure 3).

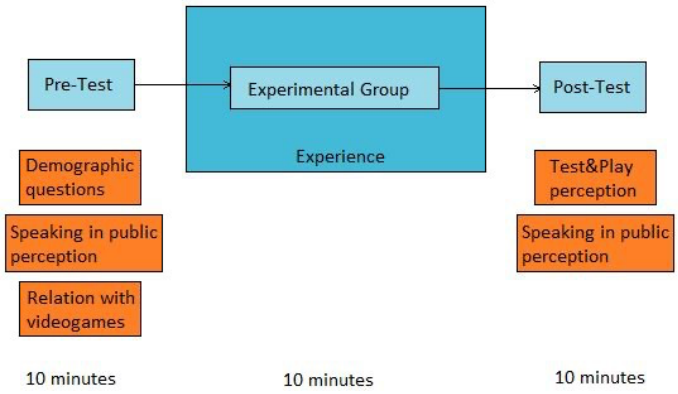


Figure 3. Experimental design scheme.

In part 1 of the experiment the participants answered two questionnaires prior to the use of the application. The first one identifies the type of game the participants are. The second one identifies if they consider themselves good or bad speakers and their self-security when speaking in public (See table 1).

Once the questionnaires were answered, part 2 started: a test with the application. The participants entered virtual reality and played Talk&Play for 10 minutes without external indications (See figure 2).

Finally, and after part 2, part 3 starts passing the final questionnaire. They answered a questionnaire to identify if after using the application they feel safer when speaking in public and if they think this application can help them improve their oratory.

Table 1. *Questionnaires summary*

Question	Type	Variable measured	How it is calculated
Self-confidence when talking in public before and after Talk&Play	7-point scale	Improvement of self-confidence	Summing up values before playing and after playing
Aspects of public speaking they consider important before and after Talk&Play	Multiple-choice question	Raise of awareness of the importance of volume and silences	Summing up the number of selected volume, speed or silences
Knowledge about the correct use of silences and volume after Talk&Play	7-point scale	Teaching properties of Talk&Play	Summing up all values

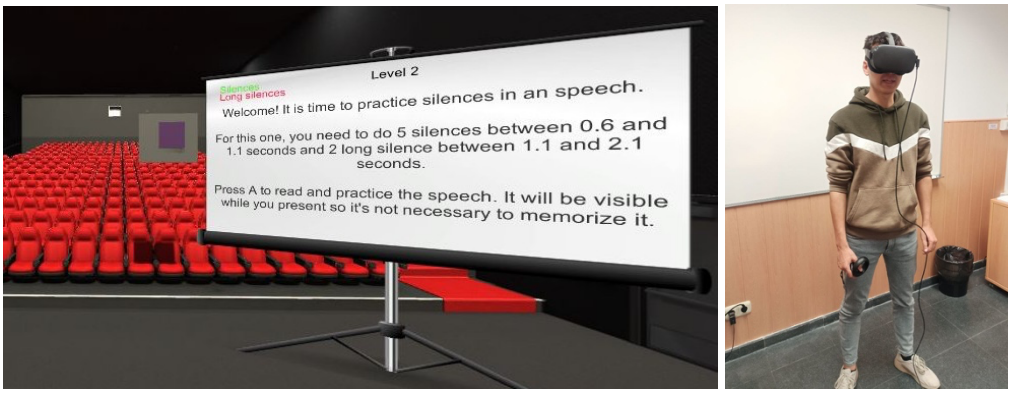


Figure 2. The silences minigame in phase 2 of level 2 and a participant in part 2 of the experiment.

3.3 Instruments and materials

For this experiment we needed hardware and software. The software part was the created application, a virtual reality videogame, developed in Unity 2020.3.33f1, which teaches about silences, speed and volume during a speech. The hardware part consisted of the Oculus Quest 1 virtual reality headset connected through Oculus link with a MSI laptop (MS-16p7) for executing the videogame.

Also, each participant had to answer three questionnaires. The first one was a validated questionnaire whose purpose is to know the type of gamer the participant is. The other two were questionnaires created specifically for this. The first one asked about demographic data of the participants, their vision of themselves related with oratory and their vision of their self-security when speaking in public, last two before playing the game. The second one asked about their perception of the teaching properties of the application and the aspects of the oratory they consider the most important, both before and after playing.

3.4 Results

At first, we did a T-student probe for related samples. In it, we compared the obtained results in user's talking in public self-confidence related questions. In this analysis we can observe (See table 2) a

statistically significant difference between the obtained results before and after playing Talk&Play (sig = 0.043*). The answers mean goes from 4.29 (out of 7) before playing to 5.07 (out of 7) after playing, having a total increase of 0.78 between both measures.

Secondly, we measured the raising awareness the videogame creates over the importance of volume, speed and silences in public speaking. For this, we summed up the number of elements the users think are important, before and after playing. Afterwards, we did a T-student probe for related samples between these two sums. The results this probe gives (See table 2) indicates a statistically significant difference (sig.=0.01*) between the elements the users think are important for public speaking. Measure goes from 1.43 (out of 3), before playing the videogame, to 2.43 (out of 3) after playing.

Table 2. *T-student test for related samples (Self-confidence and Awareness)*

	Mean	Std. Deviation	T value	Sig. (Bilateral)
Self-confidence	0.78	1.311	-2.242	0.043*
Awareness	1	0.877	-4.266	0.001*

$p<0.05$*

At last, evaluation questions were done for measuring how much capable of using pauses and volume correctly during a speech the users feel themselves. In relation to the first one, users answered a mean of 5.46 (out of 7). In relation to the second one, users answered a mean of 4.5 (out of 7).

3.5 Discussion

Using Talk&Play improves player's talking in public self-confidence and knowledge of the correct use of volume and silences. We think this is because the game gives information about the correct use of important aspects in public speaking and personalized feedback. When practicing with this information and feedback, the user can see its improvement and feel more secure doing a speech.

As well, it raises awareness of the importance of the treated aspects in a speech. We think it is because the game gives information about why these aspects are important and not only how to improve them. This lets the user figure out the difference between using them correctly or not in a speech.

Finally, we could see the demographic and videogames relation data collected does not modify the results. We think this is because this data in this experiment does not affect the learning process. Maybe, in a bigger experiment with a more varied population the results could change.

4. Conclusions and future work

Public speaking has a lot of presence in our daily lives, not only in not very common situations, but also in day-to-day activities. Most people have difficulties speaking in public because of their fears. Practicing and learning how to do it is the best way to overcome this situation. For speaking in public correctly many aspects must be taken into consideration, for example, voice volume, use of silences and speed. For improving these three aspects, we have created Talk&Play, a virtual reality videogame in which users can learn and practice the use of the volume silences and speed and their importance in a speech. In order to validate it, we have done an experiment evaluating the application in which the participants answered some questionnaires before and after playing with Talk&Play. Analyzing the received answers, we have drawn three conclusions:

1) Talk&Play increases user's speaking in public self-confidence. We have realized a large majority of the participants have increased their self-confidence during the experiment.
2) Talk&Play raises awareness of the importance of volume, speed, and silences in public speaking. The percentage of participants who thought volume, speed and silences are important in a speech before playing, increased significantly after playing.
3) Users learn with Talk&Play how to use volume and silences, but the application still needs to improve. The participants declared they have some confidence on knowing how to use silences correctly after playing, but not so much confidence on knowing how to use the volume correctly.

Nevertheless, the experiment presented in this paper only counted with fourteen people, making it a preliminary experiment and impossible to reach definitive conclusions.

There are many things to improve in the future. The main goal is to increase the learning factor. For achieving this, more visual changes for helping the user know which volume must be used should be added. Additionally, the volume real life feedback should be modified to make it easier to understand. Also, more explanation on how to decide which volume to use or where to do a silence during a speech will help on the learning process.

Furthermore, other minigames, such us fillers recognizer, could be implemented to teach the user it is all right to do a silence instead.

Finally, another experiment for evaluating the changes with a significant number of participants will allow to have definitive conclusions and find other aspects to improve.

Aknowledgements

This project has been funded by the Ministry of Science, Innovation and Universities of Spain (Didascalias, RTI2018-096401-A-I00).

References

Algaba Aranda, Jorge, Manuel Hernández Nájera-Alesón, Jorge Algaba Aranda, and Manuel Hernández Nájera-Alesón. 2019. "Desarrollo de Una Aplicación Para Analizar y Evaluar El Comportamiento de Oradores En Realidad Virtual."

Angraini, Yuli. 2016. "Rules of Three Analysis in Persuasive Public Speaking Presentation." *ANGLO-SAXON: Jurnal Ilmiah Program Studi Pendidikan Bahasa Inggris* 7(1):3. doi: 10.33373/anglo.v7i1.542.

Chance, Bonne, and Good Luck. 1975. "Public Speaking." 1–2.

Díaz, Jesús, and Yolanda Troyano. 2013. "El Potencial de La Gamificación Aplicado Al Ámbito Educativo." *III Jornadas de Innovación Docente. Innovación Educativa: Respuesta En Tiempos de Incertidumbre.*

El-Yamri, Meriem, Alejandro Romero-Hernandez, Manuel Gonzalez-Riojo, and Borja Manero. 2019. "Emotions-Responsive Audiences for VR Public Speaking Simulators Based on the Speakers' Voice." *Proceedings - IEEE 19th International Conference on Advanced Learning Technologies, ICALT 2019* 349–53. doi: 10.1109/ICALT.2019.00108.

Forsythe, Victoria. 2013. "The Effects of Graduated Exposure, Feedback, and Goal Setting on Public Speaking Anxiety." (May 2013).

Fuertes Franco, Enrique, Javier Navarro Vaquero, and Jagoba Montes Larrabaster. 2020. "VRetorik: Un Videojuego En Realidad Virtual Para Mejorar Las Habilidades de Hablar En Público."

Jos, Fernando, Teixeira Franco, and Gouveia Pais. 2017. "Interactive Avatars for Triggering Emotional Responses."

Khurpade, Jyoti Mante, Amar Ajay Gangawane, Gaurav Santosh Ostwal, Deep Manoj Lalwani, and Rushikesh Ganesh Vidye. 2020. "The Effect of Virtual Reality and Heart Rate Variability Using Deep Learning for Reducing Stage Fright- Glossophobia." *2020 International Conference on Industry 4.0 Technology, I4Tech 2020* 195–98. doi: 10.1109/I4TECH48345.2020.9102645.

Lim, Mei Hui, Vahid Aryadoust, and Gianluca Esposito. 2022. "A Meta-Analysis of the Effect of Virtual Reality on Reducing Public Speaking Anxiety." *Current Psychology* 1:3. doi: 10.1007/s12144-021-02684-6.

Noor Farah Wan Shamsuddin, Wan, Muhammad Nur Adilin Mohd Anuardi, and Intan Syafinaz Mohmad Rozee. 2022. "View of A Preliminary Study on the Potential of Virtual Reality Therapy in Reducing Public Speaking Anxiety." Retrieved May 26, 2022 (http://revistas.ucv.edu.pe/index.php/espergesia/article/view/2087/1800).

North, Max M., Sarah M. North, and Joseph K. Coble. 1998. "Virtual Reality Therapy: An Effective Treatment for Phobias." *Studies in Health Technology and Informatics* 58:112–19. doi: 10.3233/978-1-60750-902-8-112.

Pratiwi, Titin Indah, Bambang Suratman, Yatim Riyanto, and Pradita Arisgi Werhadiantiwi. 2018. "A Special Service of Guidance and Counseling Management to Reduce the Students' Glossophobia Level." 679–81. doi: 10.2991/ICEI-18.2018.153.

Tse, Andrew Yau Hau. 2005. "Glossophobia of University Students in Malaysia." *International Journal of Asian Social Science* 2(11):2061–73.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Impact of Gaming on the Mental Well-Being and Academic Performance among High School Students

Aaron James CAPINPIN, Roi Victor ROBERTO, Ramon Diego TAN & Judith AZCARRAGA*

De La Salle University, Philippines
*judith.azcarraga@dlsu.edu.ph

Abstract: This study aims to determine the impact of gaming on mental well-being and academic performance among teenage students. A survey was conducted among 34 senior high school students to gather data on whether gaming can affect their academic performance and mental well-being. The survey consists of questions regarding the students' studying habits, gaming habits, and overall academic performance. Results have shown that students who play video games frequently tend to score lower in terms of academic performance, particularly on academic competence, test competence, and strategic studying. However, they score significantly higher than moderate and low gamers when it comes to time management. In terms of mental well-being, results suggest that frequent gamers scored lower than moderate gamers in terms of engagement, perseverance, optimism, connectedness, and happiness. On the other hand, in social behavior, frequent gamers often obtain "normal" scores when compared to moderate gamers who obtained more "borderline" or "abnormal" scores.

Keywords: mental well-being, gaming, social behavior and psychological functioning, academic performance

1. Introduction

In the past two decades, video games have been gaining even more popularity due to various technological advancements in the field which allowed them to be used for more than just entertainment but as a means of relaxation and even to cope with mental stress due to the effects of the COVID-19 pandemic.

Some studies have shown some negative effects of gaming on children on their psychosocial wellbeing and even their quality of sleep. Frequent to excessive gamers may be at risk for behaving less prosocial (Lobel et al. 2017), poorer psychological well-being (Goh et al. , 2019), problematic behavior for the user such as increased physical aggression (Lemmens et al.,2011, Fischer et al., 2010), Holtz et al. 2014) and poor sleep quality and overall well-being (Capinpin et al., 2022).

However, some studies found negative correlation in examination marks with gaming frequency (Ip et al.,2008) and positive correlation of internet gaming with psychological distress, which is linked negatively with emotional intelligence and academic performance (Zahra et al., 2020). It is without doubt, that although gaming has benefits, it may still have numerous negative impacts on one's physical and mental health as well. This study attempts to determine the impact of gaming on mental well-being and academic performance among teenage students.

2. Methodology

A survey was conducted among 34 senior high school students from various places in the Philippines, who regularly play video games. Since this study involves minor participants, informed consent and assent forms were signed by the participants and their parents or guardians. The participants answered a set of questions on their demographics, study habits, gaming habits, and overall academic performance thru a Google form. The samples were clustered based on their answers in the demographic survey. Participants who stated that they play video games for more than 35 hours a week are clustered into the

heavy gamer category and those who stated that they played less than 35 hours a week are clustered into the *moderate* gamer category, following the Youth Risk Behavior Surveillance 2009 in Eaton et al. (2010) category for youth with health risk behavior who spend 3 hours or more per day in video and online games.

The survey questions are based on 3 instruments, namely, the Study Management and Academics Results Test (SMART) adapted in Kleijn et al. (1995), Strengths and Difficulties Questionnaire (SDQ) to measure social behavior adapted in Goodman (2001) and EPOCH instrument to measure the level of engagement, perseverance, optimism, connectedness, and happiness adapted in Kern et al. (2016). For the SMART instrument, four scales are measured, particularly, the (1) academic competence which is the capability of the student to handle the study loads; (2) test competence which is the capability of the student to handle exam stress and test-related works; (3) time management which is the ability of the student to balance study and leisure time; and (4) strategic studying which is the ability of the student to evaluate and adapt based on the given workloads in the school. The Strengths and Difficulties Questionnaire measures a range of mental health symptoms such as prosocial behavior which is the willingness to act for the benefit of other people, peer problems which is the behavior to his or her peers, hyperactivity which checks if a person is hyperactive or inattentive, conduct problems which is the ability of the to follow rules, or behave in a socially acceptable way, and emotional problems which measures the current mental or emotional state. The EPOCH Assessment, on the other hand, covers five scales such as engagement which refers to being absorbed, interested, and involved in an activity or the world itself, perseverance which refers to the continued effort to do or achieve something despite difficulties, failure, or opposition, optimism which refers to having a sense of hope and confidence about the future, connectedness which refers to feeling loved, supported, and valued by other and happiness which refers to the general feeling of happiness, cheer, and contentment with life.

3. Results and Discussions

Based on SMART results in Figure 1a, the heavy gamer group scored higher in test competence and significantly higher in time management than moderate gamers in the test but scored lower in both academic competence and strategic studying. Moderate gamers scored highest in academic competence while lowest in time management. This significant difference in time management of heavy gamers may suggest that this group tends to balance their academic time and their leisure time well. Since test competence is the capability of the student to handle exam stress, heavy gamers tend to allocate more time to their leisure time and are able to handle stress very well. Moderate gamers obtained higher scores in strategic studying and academic competence since they tend to dedicate more time to studying than leisure.

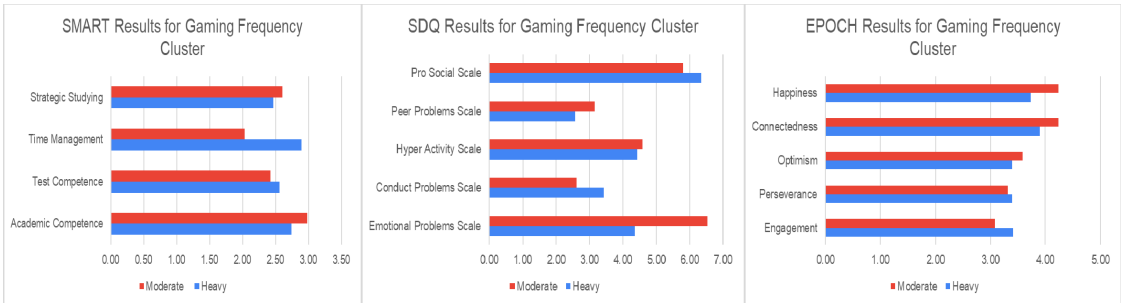


Figure 1. (a) Study Management and Academic Results Test; (b) Strengths and Difficulties Questionnaire Assessment; (c) EPOCH Questionnaire Assessment

For the SDQ scores in Figure 1b , lower scores imply a more positive behavior. *Heavy* gamers often obtain "normal" scores when compared to *moderate* gamers who obtained more "borderline" or "abnormal" scores. *Moderate* gamers got high scores when measuring the emotional problems scale and prosocial scale. EPOCH results in Figure 1c show that *heavy* gamers tend to score lower on happiness, connectedness, and optimism, while scoring higher levels of engagement and perseverance.

As such, this may suggest that heavy gamers tend to score higher on engagement and perseverance simply because they classify as those who engage in gaming for long periods of time and hence, they are more engaged and absorbed while playing video games.

4. Conclusion and Future Works

This study attempts to determine the impact of gaming on mental well-being and academic performance among teenage students during pandemic. A survey was conducted among 34 high school students to find out the impact that gaming had on various aspects of their academic performance and mental well-being. The results suggest that in terms of academic performance, gaming may help ease the stress being experienced by students from taking exams and may also help students manage their time more wisely. However, an excessive amount of gaming may affect the student's ability to handle study loads or even their ability to evaluate and adapt to the given workloads of the school. Moreover, it was found out that heavy gaming actually led to a healthier state of well-being due to the data obtained from the SDQ. However, heavy gamers exhibit lower scores of optimism, connectedness, and happiness. This may suggest that individuals who tend to engage in gaming for an unhealthy and extended period of time may feel generally less happy, loved/valued by others, and especially lose confidence about themselves or the future. Nevertheless, it is safe to assume that an unhealthy amount of gaming may be detrimental to the well-being of an individual.

Future works are recommended to use the same instruments with a larger sample size in order to obtain more accurate and well-represented results. Future study may explore other gaming habits of students such as which days they play, what moods and emotions have led them to play, and the genre of games they frequently play, and correlating how these have affected the overall mental health and academic performance of the students.

References

Capinpin, A. J. M., Roberto, R. V. P., & Tan, R. D. A. (2022). Analyzing the Impact of Gaming on the Mental Well-Being and Academic Performance among Senior High School Students. (*In press) SHS Research Congress*. De La Salle University Press.

Eaton, D. K., Kann, L., Kinchen, S., Shanklin, S., Ross, J., Hawkins, J., Harris, W., Lowry, R. et al. (2010). Youth Risk Behavior Surveillance – United States, 2009. *Morbidity and Mortality Weekly Report. Surveillance Summaries* 59(SS05), 1-142.

Fischer, P., Kastenmüller, A., & Greitemeyer, T. (2010). Media violence and the self: The impact of personalized gaming characters in aggressive video games on aggressive behavior. *Journal of Experimental Social Psychology*, *46*(1), 192-195. h

Goh, C., Jones, C., & Copello, A. (2019). A Further Test of the Impact of Online Gaming on Psychological Wellbeing and the Role of Play Motivations and Problematic Use. *Psychiatric Quarterly*, *90*(4), 747–760.

Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337-1345.

Holtz, P., & Appel, M. (2011). Internet use and video gaming predict problem behavior in early adolescence. *Journal of adolescence*, *34*(1), 49-58.

Ip, B., Jacobs, G., & Watkins, A. (2008). Gaming frequency and academic performance. *Australasian Journal of Educational Technology*, *24*(4).

Kern, M. L., Benson, L., Steinberg, E. A., & Steinberg, L. (2016). The EPOCH measure of adolescent well-being. Psychological assessment, 28(5), 586.

Kleijn, W. C., van der Ploeg, H. M., & Topman, R. M. (1994). Cognition, study habits, test anxiety, and academic performance. *Psychological Reports*, *75*(3), 1219-1226.

Lemmens, J. S., Valkenburg, P. M., & Peter, J. (2011). The effects of pathological gaming on aggressive behavior. *Journal of youth and adolescence*, *40*(1), 38-47.

Lobel, A., Engels, R. C., Stone, L. L., Burk, W. J., & Granic, I. (2017). Video gaming and children's psychosocial wellbeing: A longitudinal study. *Journal of youth and adolescence*, *46*(4), 884-897.

Zahra, S., Ahsan, S., Kiani, S., Shahbaz, K., & Andleeb, S. N. (2020). Internet Gaming, Emotional Intelligence, Psychological Distress, and Academic Performance Among University Students. *Pakistan Journal of Psychological Research*, *35*(2), 253-270.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Shelter GO: Multiplayer Location-based Game for Learning Evacuation

Hiroyuki MITSUHARA[a*] & Masami SHISHIBORI[a]

[a] *Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Japan*
*mituhara@is.tokushima-u.ac.jp

Abstract: We evacuate to a shelter in the event of a disaster. Furthermore, we abandon the evacuation or wander toward an unknown shelter if we do not remember our shelter location. For successful evacuation, residents should not only remember shelter locations in their local community but also shelter capacities and features. By visiting and observing shelters in advance, residents can remember not only shelter locations and capacities but also shelter features. However, it is difficult to maintain the motivation to visit shelters. Additionally, residents should remember shelters within and outside their local community and acquire several shelter-related knowledge from other residents. Thus, we developed a prototype of a multiplayer location-based game (a geofencing mobile application) that adopts a game element similar to Pokémon GO. The game prototype prompts players to visit different local communities by making them expect encounters from other players and enhancing knowledge acquisition from encountered players.

Keywords: Multiplayer location-based game, knowledge exchange, shelter, evacuation

1. Introduction

Severe natural and artificial disasters have caused damage and deaths worldwide (CRED, 2022). When a disaster strikes or is predicted, we evacuate to a shelter depending on the circumstances. Shelters are indispensable for protecting lives during disasters. However, the evacuation is usually difficult and the difficulty remarkably increases with unpredictable disasters. For example, there is time to decide which shelter to evacuate to and then calmly begin evacuation when there is an incoming typhoon. Contrarily, if an earthquake occurs, we must quickly decide which shelter to evacuate to. Such quick decisions are difficult if shelter locations are forgotten. Failure to remember these locations results in abandoning the evacuation or wandering toward an unknown shelter. In a worst-case scenario, lives may be lost during evacuation. Therefore, remembering the shelter location results in quick decisions and successful evacuation.

Shelters can also be marked on a local community map. Although when there is panic or limited perception (e.g., low visibility due to a blackout), residents may be unable to locate the map, as shelters can be marked, when an unpredictable disaster occurs, residents can begin evacuation by referring to the marked shelter locations on the map. Thus, it is important to remember shelter locations from the map in advance. However, simply looking at the map does not always help in remembering shelter locations because the map is not specialized for evacuation. A hazard map, which not only emphasizes shelters but also visualizes simulated damage (e.g., inundation depth by flood or tsunami), is not only used to remember shelter locations but also to make evacuation plans (e.g., selecting suitable shelter candidates). For example, looking at a flood hazard map can generate the idea that a suitable shelter in a flood is not the nearest shelter surrounded by inundated areas but a little-distant shelter unaffected by inundation. Recently, hazard maps have been digitized based on geographical information systems (GIS) to provide multimedia content and interactivity for better disaster risk reduction planning (Dransch et al., 2010) (McCallum et al., 2016). For example, a web-based hazard map system using social media GIS allows residents to upload and share local disaster information (e.g., potential disaster-induced dangers) on Google Maps (Yamamoto & Fujita, 2015). Another web-based hazard map system is used to educate high school students to identify flood risk areas (Song et al., 2022).

Compared with paper-printed hazard maps, digital hazard maps can be easily distributed among those who frequently use smartphones, i.e., modern people.

Under normal circumstances, residents who consider disasters as someone else's problem will never be motivated to look at a hazard map, regardless of whether it is paper-printed or digital. Thus, a new motivational method other than looking at a hazard map is required for such residents to remember shelter locations. Furthermore, to decide which shelter to evacuate to, residents must remember not only shelter locations but also shelter capacities and features. For example, if numerous residents rush into a small-capacity shelter, the elderly who cannot evacuate quickly might abandon the crowded shelter and evacuate to a less crowded one. Families with infants who do not remember shelter features might go to the nearest shelter, which does not have nursery facilities. A hazard map occasionally shows shelter capacities (and features in some cases) but does not always show information sufficient for various residents to decide on a suitable shelter. For example, as residents should not only study a hazard map but also visit and observe shelters to effectively remember shelter locations, capacities, and features, a hazard map not showing altitudes around shelters might make the elderly take a steep route to the nearest shelter in a high place. However, the new motivational method must prompt residents to visit shelters without feeling burdened, as even for residents who consider disasters a problem, it is difficult to be motivated to visit shelters because of a temporal and/or physical burden.

Introducing game elements is a promising approach for maintaining motivation. Particularly, educational games or game-based learning have attracted considerable attention and have proven high efficacy in various domains, such as disaster education. A narrative-driven location-based augmented reality (AR) mobile game was developed for players to study English by interacting with non-playable characters or to obtain items displayed on a digital map (Lee, 2022), where location-based games are frequently introduced when focusing on learning activities that require real world exploration. For example, a location-based game for local cultural learning displays context-aware learning content (e.g., quizzes) by detecting whether students are near historic monuments (Lin et al., 2018).

Based on this background, we developed a prototype of a single-player location-based game (a mobile application (app)) that motivates players (residents) to visit shelters by introducing a game element similar to that of Pokémon GO (Mitsuhara et al., 2022). The players visit a shelter to earn a point equal to the capacity of the shelter and visit an informative disaster management spot to collect a digital creature by paying a required point from their earned points. The digital creature collection can motivate players to visit and observe various shelters without feeling burdened; thus, players can remember several shelter locations, capacities, and features. Our preliminary experiment revealed that the game prototype helped players to remember more shelter locations and accurate shelter capacities compared with just looking at a hazard map. However, the game prototype focuses on a single-player mode and does not introduce a multiplayer game element, which is a characteristic of Pokémon GO. The single and multiplayer modes have several learning-related merits and demerits, and game/instructional designers should consider a better mode to maximize learning effects in educational games to be developed (Harteveld & Bekebrede, 2011). Because a multiplayer mode is effective in maintaining motivation and learning evacuation by encountering other players, we extended the game prototype to a multiplayer location-based game. The extended game prototype, Shelter GO, focuses on knowledge exchange (KE) among players.

2. Learning Process

Residents must remember shelters for a successful evacuation (i.e., location, capacity, and feature). Furthermore, they must acquire available shelter-related knowledge to decide which shelter to evacuate to (e.g., "This small shelter was crowded during a previous earthquake. Let's evacuate to a large shelter.") Thus, learning about evacuation corresponds to remembering shelters and acquiring shelter-related knowledge.

Residents can effectively learn evacuation by visiting and observing shelters. However, this type of learning forces residents to walk around the real world. Walking may entail a temporal and/or physical burden and consequently cause players to be unmotivated to continue learning. To remove the burden, we introduced a game element similar to that of Pokémon GO, in which players explore the real world to get a reward (i.e., collect imaginary creatures), in the game prototype.

2.1 Pokémon GO's Potential

Pokémon GO has gained worldwide popularity, and several studies have been conducted to investigate its influence on physical activity (Lee et al., 2021), mental health (Watanabe et al., 2017), and cognitive performance (Ruiz-Ariza, et al., 2018), among others. According to reports, factors driving the use and gratification of Pokémon GO are enjoyment, challenge, competition, socialization, etc. (Hamari, et al., 2019). Because the game system involves a real world exploration, Pokémon GO is available for health education (Finco, et al., 2018), language learning (Wu, 2021), etc. Using Pokémon GO for educational purposes has several advantages and disadvantages (Tran, 2018). For example, traffic accidents caused by the use must be avoided. However, adopting a location-based game is the most prospective approach for promoting an ideal learning process.

2.2 Single-Player Mode

The single-player mode of the game prototype followed the four-phase learning process described below. Although shelter-related knowledge is sharable via a web-based hazard map system, the game prototype does not rely on an external system and achieves knowledge acquisition within the learning process.

- Preview: In the introductory phase, a player (resident) previews a hazard map and gets a sense of shelter locations (e.g., distance and direction) and capacities. Then, the player decides the shelters to visit. This phase can be skipped if the player knows the shelter locations and capacities.
- Walk: This phase involves moving between shelters and acquiring shelter-related knowledge. During this phase, the player will recognize the conditions of routes to and/or surroundings of the shelter and acquire shelter-related knowledge (e.g., "Let's evacuate via a wider street if possible because a huge earthquake can make narrow paths to this shelter impassable.") Acquiring this type of shelter-related knowledge is difficult by only looking at the hazard map.
- Visit and observe (VO): This phase involves remembering shelters and acquiring shelter-related knowledge. The player visits and observes a shelter and clearly remembers the shelter. For example, if there is a remarkable difference between the official capacity shown on the hazard map and the player's presumed capacity via observation, the player will be surprised, thus making the player remember the official capacity. Furthermore, observing the shelter features (e.g., building material and apparent age) can help determine whether the shelter is suitable in the case of various scenarios, such as the occurrence of a mega earthquake. The player walks to another shelter after the observation.
- Visit and collect (VC): Informative spots are placed on disaster monuments and alert signs. The player visits an informative spot to collect a digital creature and obtain tips for successful evacuation. The digital creature collection, similar to Pokémon GO, will eliminate the walking-caused burden and keep players motivated; thus, the collected tips will result in acquiring a wide range of knowledge.

2.3 Multiplayer Mode

It is reasonable from an efficiency perspective that players (residents) want to only remember shelters in their local community (i.e., living area). The single-player mode covers the learning process, which presumes that individual players walk around their local community. Ideally, players must remember shelters within and outside their local community and acquire a variety of shelter-related knowledge from other players, as it is impossible to predict where and when a disaster will strike. Additionally, individual players' shelter-related knowledge may not vary depending on their interests, experience, and other characteristics.

The extended game prototype adopts a multiplayer mode that promotes players to visit shelters outside their local community and exchange knowledge with other players. Digital creatures unique to each local community are assigned regardless of the single or multiplayer mode. Players must visit shelters in different local communities to collect the unique digital creatures. Furthermore, the multiplayer mode enables players to meet face-to-face in the real world to exchange their collected digital creatures. The digital creature exchange, which is similar to that in Pokémon GO, prompts players to visit different local communities by making them expect to encounter other players there; the

encountered players are expected to have collected unique digital creatures for exchange. Furthermore, to enhance acquired knowledge with the encountered players, the multiplayer mode presents players with the following duties:

- Register knowledge: When collecting a digital creature from an informative spot, a player must register his/her knowledge to the creature. This type of knowledge registration is known as knowledge externalization, which helps stabilize the acquired knowledge. Both shelter-related and general knowledge for successful evacuation are acceptable as the registered knowledge.
- Transfer knowledge: The player can no longer see the registered knowledge on the extended game prototype as, when exchanging a digital creature with another player, the player must transfer the registered knowledge together with the creature. Therefore, the player must remember the registered knowledge while assuming that his/her collected digital creatures will leave.
-

Figure 1 illustrates the ideal learning process that combines single and multiplayer modes.

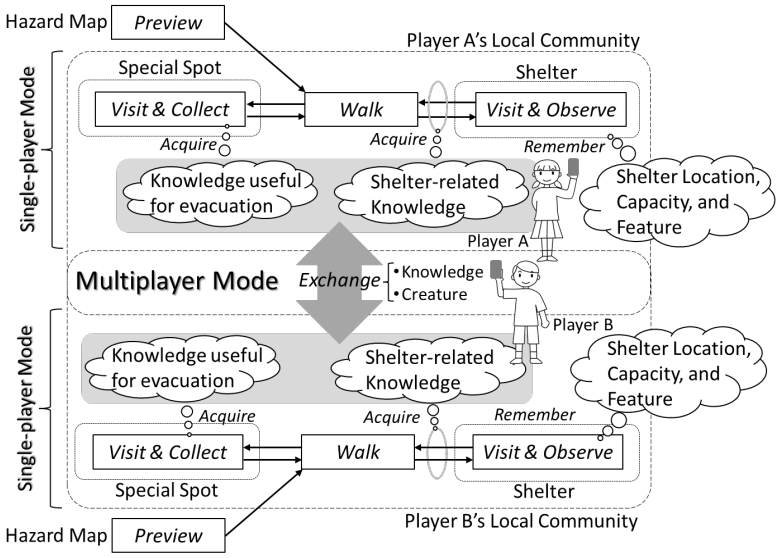


Figure 1. The ideal learning process for remembering shelters and acquiring knowledge.

3. Shelter GO

Shelter GO is a geofencing mobile app extended from our single-player location-based game prototype, which equips the multiplayer mode.

The mobile app, which was developed using Unity3D, works on a GPS-enabled smartphone (or tablet) and allows players to switch to the single-player (i.e., without an Internet connection) or the multiplayer modes (i.e., with an Internet connection). For the single-player mode, the app has basic modules (e.g., phase detection, phase transaction, and point management) and internal data storage (e.g., shelter, informative spot, and point). Shelter data consist of name, latitude, longitude, and capacity (i.e., point). Informative spot data consist of latitude, longitude, digital creature, and tips for successful evacuation. For the multiplayer mode, the app additionally has modules (e.g., knowledge registration and transfer) and external data storage required for KE. The external data storage, where a data interface module (implemented by PHP) makes the transaction, has a CSV file for each KE that records temporary data, such as player ID, player name, creature ID, and registered knowledge.

3.1 Single-player Mode

The app deals with three phases in the single-player mode, i.e., walk, VO, and VC.

1) Walk (Figure 2-a): The app takes a player to the walk phase where a digital map, "Gallery" button, and some information (e.g., earned points) are presented. The map shows the player's current location, shelters, and informative spots. When a shelter is reached (invisible rectangle area enclosing the shelter), the player can move to the VO phase corresponding to the shelter. When an informative spot is reached, the player can move to the VC phase, further, by tapping the button, the player is taken to the gallery where he/she can see the list of the collected digital creatures. At

random intervals, the app compulsorily takes the player to a digital creature encounter event, which is not associated with the location. Although not originally required, this compulsory transfer is prepared to avoid a situation where the player cannot efficiently visit informative spots.

2) VO (Figure 2-b): The app displays the name and capacity of the visited shelter, and the players receive a point equal to the capacity. They can backtrack to the walk phase by tapping the OK button. The app does not display a message encouraging the players to observe the shelter. However, they will be triggered to observe the shelter by comparing the displayed capacity and the shelter presented before them. They will visit another shelter to earn points, as the players cannot earn the point again at the same shelter until a certain time passes.
3) VC (Figure 2-c): At the arrived informative spot, the app first displays instructions (e.g., "Please look at the alert sign about a disaster around you."). Next, the app activates the rear camera of the smartphone and displays a digital creature in an AR mode, which superimposes the digital creature CG onto the real-time vision using a markerless AR framework. When the players tap on the creature on the screen, the app displays tips associated with the spot (e.g., "0.9 m above sea level here. If you stay here during a big earthquake, you will be engulfed by the tsunami."). The app provides general tips during the compulsory transfer (e.g., "School buildings are often registered as shelters.") After reading the tips, the players can acquire the creature by paying the required points.

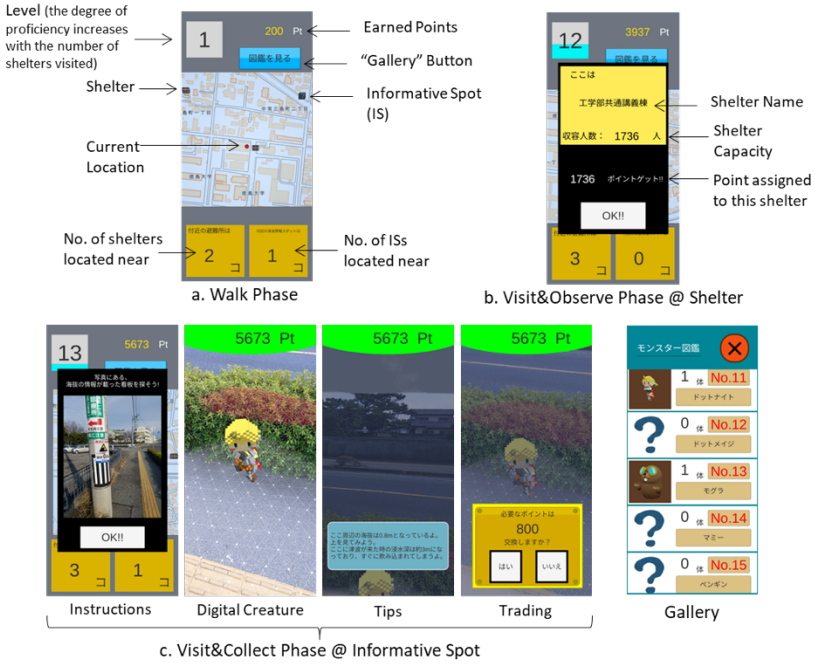


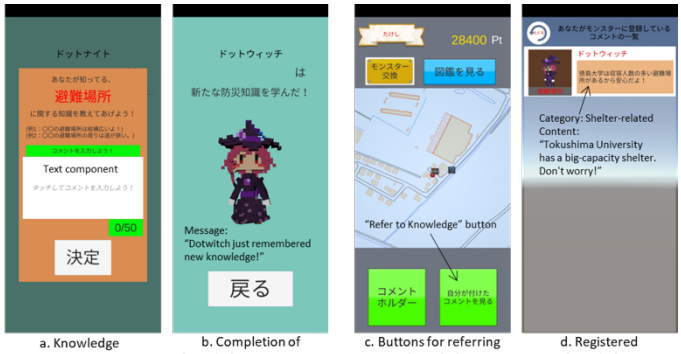
Figure 2. User interfaces in the single-player mode.

3.2 Multiplayer Mode

The app extended for the multiplayer mode forces a player to register his/her knowledge to digital creature and further adopts a client-server architecture required for KE (i.e., digital creature exchange). The server works as the relay point for KE. Currently, the app only allows KE between two players simultaneously. The KE phase begins when two players meet face-to-face in the walk phase and they agree to exchange knowledge.

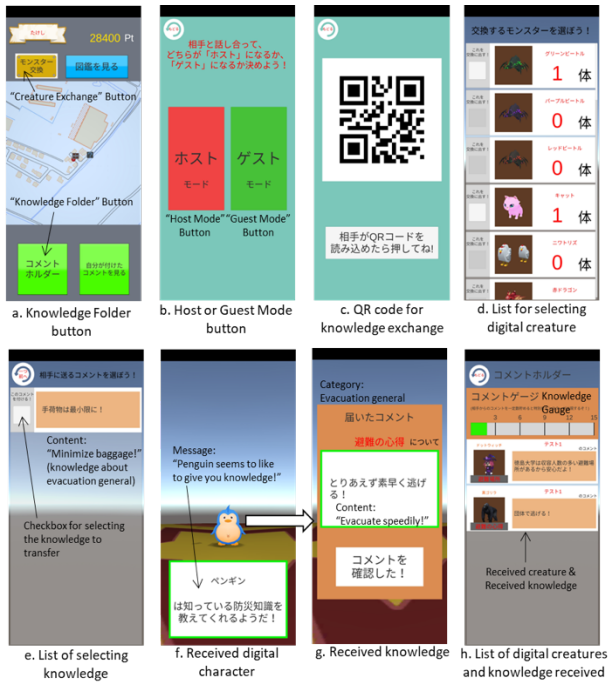
1) Register knowledge (Figure 3): When players acquire a digital creature in the VC phase, the app displays a text component for them to register their knowledge to the creature (Figure 3-a). They must input short text knowledge after selecting one of three categories: shelter-related, evacuation general, and disaster general. Furthermore, they can register some pieces of knowledge to the creature. After completing knowledge registration, the app displays the creature together with the message "[The creature's name] just remembered new knowledge" (Figure 3-b). By tapping the "Refer to Knowledge" button in the walk phase (Figure 3-c), the player can refer to their registered knowledge as the list of the digital creatures collected (Figure 3-d).
2) KE (Figure 4): A player who recognizes a nearby player in the walk phase can speak to the player and suggest KE. If the suggestion is accepted, both players tap the "Creature Exchange" button

(Figure 4-a) to display the menu (Figure 4-b). One player taps the “Host Mode” button and the app displays a QR code (Figure 4-c). Another player taps the “Guest Mode” button and then scans the QR code using the rear camera of a smartphone, which is activated using the app. After a successful scan, the app displays another QR code for the host player to scan. When both scans are successful, the app displays a list of digital creatures (Figure 4-d). The list includes unexchangeable digital creatures (i.e., not-yet-collected digital creatures) on each player’s app. Each player selects one from the exchangeable creatures as a creature to transfer and then selects one or more knowledge registered to the selected creature (Figure 4-e). Then, the selected knowledge is transferred together with the selected creature. The data of the transferred creature and knowledge are deleted from the app, where, when both players finish the selection, the app begins transacting the creature exchange (i.e., KE) and successfully displays the creature CG and the knowledge received via the server (Figures 4-f and g). Next, the players can continue creature exchange. In the walk phase, by tapping the “Knowledge Folder” button (Figure 4-a), the player can refer to the list of knowledge received from other players together with the corresponding received creatures (Figure 4-h). A gauge called “Knowledge Gauge” is displayed above the list. The value on the gauge indicates the amount of exchanged knowledge and expresses the player’s activeness of KE. Every time three pieces of knowledge are exchanged, a special digital creature not associated with an informative spot appears. The player acquires a special creature by paying his/her earned points. These special creatures will promote KE. Thus, the player will visit more shelters to earn points and acquire more knowledge from other players.



a. Knowledge registration component | b. Completion of knowledge registration | c. Buttons for referring to knowledge | d. Registered knowledge

Figure 3. User interfaces for knowledge registration in the multiplayer mode.



a. Knowledge Folder button | b. Host or Guest Mode button | c. QR code for knowledge exchange | d. List for selecting digital creature

e. List of selecting knowledge | f. Received digital character | g. Received knowledge | h. List of digital creatures and knowledge received

Figure 4. User interfaces for knowledge exchange in the multiplayer mode.

4. Preliminary Experiment

We conducted a preliminary comparative experiment to examine the difference between the single and multiplayer modes. Here, we focused on the tsunami evacuation because participants resided in a coastal area where a tsunami can cause devastating damage. Thus, the participants must originally learn about evacuation to tsunami shelters.

4.1 Settings

The participants were ten undergraduate students in a computer science course who were found to have low interest and knowledge about tsunami risk reduction through a pre-questionnaire. They were divided randomly into the following groups:

- Group A (N = 5): The participants simultaneously played an extended game prototype that excluded the KE function. They played the game from a different location and registered knowledge to a digital creature when acquiring the creature.
- Group B (N = 5): On a different day from Group A, they simultaneously played an extended game of the prototype (i.e., the multiplayer mode), which allowed them to exchange knowledge. They played from a different location and freely decided whether to exchange knowledge.

After they were instructed on how to use the game prototype, we asked them to learn about the evacuation to tsunami shelters using the app installed on prepared smartphones. We did not impose a time limit although they were informed that it would take approximately 1 h to visit the tsunami shelters. Following the gameplay, each participant completed a surprise test and a post-questionnaire. The surprise test was a recall test asking about the shelter locations and capacities. Showing whether they remembered the shelter locations, they marked the shelter locations on a blank map of an experimental area with no marked tsunami shelters. Furthermore, they wrote the shelter capacities on the map, showing whether they remembered the capacities. The questionnaire required them to itemize all their acquired knowledge from the gameplay, and they were tasked about the effects of the game they played.

The experimental area, covering a 700 m^2 area around the participant's university campus, had sixteen tsunami shelters (buildings that met safety standards) and eight informative spots (e.g., alert signs against tsunami inundation). The extended game prototype recorded participant behavior, such as visiting shelters, registering knowledge, and exchanging knowledge.

4.2 Results

The mean gameplay durations of Groups A and B were 43.4 and 52.6 min, respectively. Table 1 presents the results of the surprise test for all participants and the number of visited shelters and informative spots. We considered a recalled shelter location to be correctly recalled if a participant correctly marked the shelter–building correspondence on the blank map. For the accuracy rate for all shelter locations (ARASL), the mean values of Groups A and B were 43.7% (31.2%–68.7%) and 33.7% (18.7%–56.2%), respectively. For the accuracy rate for visited shelter locations (ARVSL), the mean values of Groups A and B were 66.5% (41.6%–100%) and 59.3% (33.3%–100%), respectively. For the accuracy of a recalled shelter capacity, we evaluated the mean absolute value of the difference between the correct and recalled capacities for each participant (MAVDC). The mean values of Groups A and B were 86.9 (36.0–217.5) and 70.9 (47.0–141.6), respectively.

For the number of registered knowledges or collected digital creatures (NRK), the mean values of Groups A and B were 6.4 (3–11) and 4.4 (3–6), respectively. For the number of acquired knowledges (NAK), i.e., the knowledge that the participants itemized on the questionnaire, the mean values of Groups A and B were 3.4 (2–6) and 2.6 (2–3), respectively. For the number of exchanged knowledges (NEK) in Group B, the mean value was 2.0 (1–3). Table 2 presents the mean values and standard deviations of the questionnaire with five-point Likert scale questions. The mean values of Q1–Q3 were 4.0 or more in both groups. However, for Q4, the mean value in Group B was 1.2 points higher than that in Group A. The mean values of Q5–Q7, only given to Group B, were more than 4.0.

Table 1. *Quantitative Results for Learning Evacuation Between Groups*

Group A						Group B					
P	NVS	NVIS	ARASL	ARVSL	MAVDC	P	NVS	NVIS	ARASL	ARVSL	MAVDC
1	10	6	68.7%	100.0%	44.0	6	13	3	43.7%	53.8%	55.8
2	9	5	31.2%	55.5%	101.0	7	9	4	18.7%	33.3%	58.3
3	7	4	43.7%	85.7%	36.0	8	6	8	25.0%	66.6%	47.0
4	12	7	31.2%	41.6%	36.0	9	6	6	56.2%	100.0%	141.6
5	10	5	43.7%	50.0%	217.5	10	9	7	25.0%	44.4%	51.7
M	9.6	5.4	43.7%	66.5%	86.9	M	9.2	5.6	33.7%	59.6%	70.9

P: participant, NVS: number of visited shelters, NVIS: number of visited informative spots, ARASL: accuracy rate for all shelter locations (i.e., sixteen shelter locations), ARVSL: accuracy rate for visited shelter locations, MAVDC: the mean absolute value of the difference between the correct and recalled capacities of visited shelters, and M: mean

Table 2. *The Mean Values (and Standard Deviations) of the Five-point Likert Scale Questions*

Questions	Group A	Group B
Q1. Do you agree that this game was enjoyable?	4.6 (0.74)	4.6 (0.80)
Q2. Do you agree that this game increased your interest in disaster management of tsunamis?	4.0 (0.48)	4.4 (0.63)
Q3. Do you agree that this game is more effective for remembering tsunami shelters than looking at a tsunami hazard map?	4.4 (0.80)	4.6 (0.63)
Q4. Do you agree that registering the knowledge to digital creatures contributes to learning evacuation?	3.2 (1.01)	4.4 (0.40)
Q5. Do you agree that exchanging digital creatures with other players contributes to acquiring new knowledge?	N/A	4.2 (0.74)
Q6. Do you agree that exchanging digital creatures is effective for learning evacuation?	N/A	4.2 (0.74)
Q7. Do you agree that you want to actively exchange digital creatures?	N/A	4.4 (0.80)

4.3 Discussion and Limitations

Because of the small number of participants, statistical tests were not applied. Group B participants learned approximately 10 min longer than those in Group A. However, participants in both groups visited a similar number of shelters (NVS) and informative spots (NVIS). Group B participants may have needed more time for KE. This means that Group A participants visited shelters and informative spots more efficiently than Group B participants. Additionally, Group A participants remembered more shelter locations (ARASL and ARVSL) and more accurate shelter capacities (MAVDC) than those in Group B. This means that they learned more successfully than the Group B participants. Although the reasons are unknown, these results may have been caused by participant characteristics, such as motivation, physical condition, and learning style. Thus, the characteristics of the participants may not have been normalized between Groups A and B. The results of the surprise test indicated that Group A was superior to Group B but the differences between the groups were unremarkable. Without or with the KE function, these results may indicate that KE takes time but does not remarkably prevent learning evacuation.

For NRK, the mean value of Group A was 2.0 points higher than that of Group B. However, there was no remarkable difference in NVS and NVIS between the groups. These results indicate that Group A participants collected digital creatures more eagerly than those in Group B when visiting informative spots. However, for the NAK, there was no remarkable difference between the groups. For both groups, and in results that may indicate that by focusing on the NRK and NAK, knowledge registration did not significantly affect knowledge acquisition, participants were expected to acquire shelter-related knowledge in the walk and VO phases and tips (knowledge) for the successful evacuation in the VC phase. For the NEK, under a certain conditions (i.e., five participants in a 700 m^2 area within an average of 52.6 min), it may be reasonable that meeting another player and exchanging knowledge was infrequent. We found one case where KE may have successfully affected knowledge

acquisition. On the pre-questionnaire, a participant itemized the two pieces of knowledge that had already been acquired before the experiment:

- *My apartment was assigned as a tsunami shelter. During the earthquake, I evacuated to a higher floor of my apartment.*
- *The tsunami may reach fast. So, I evacuated quickly.*

On the post-questionnaire, the participant itemized the four knowledge acquired through the gameplay:

- *A three-meter-high tsunami may come to the Suketobashi area.*
- *The Abestate building is a shelter.*
- *The Suketobashi area had fewer shelters.*
- *Tokushima University can accept 1,800 people per shelter.*

Before the gameplay, the participant did not know that Tokushima University was a tsunami shelter. However, the participant remembered that the university is a 1,800-person-capacity shelter after the gameplay. They received the following knowledge from another player through KE:

- *Tokushima University is a big-capacity shelter.*
- *I recommend Gym (in Tokushima University) because it is large (as a shelter).*

The participant accurately remembered the shelter location and capacity, as they visited and observed the university campus after receiving this knowledge. Although it is insufficient to clarify the effects of KE, the successful case shows that it should be extensively explored.

For the questionnaire results, the mean values were favorable except for Q4 in Group A. These results indicate that introducing a game element like Pokémon GO for learning evacuation was accepted. Group A participants actively registered their knowledge to the collected digital creatures, but they disagreed on the contribution of knowledge registration. This indicates that knowledge registration is a burden to players due to emphasizing learning rather than gaming. The balance between learning and gaming is a very crucial agenda in Shelter GO, as is often discussed in the field of game-based learning. For all questions, the mean values of Group B were favorable. Group B participants strongly evaluated the game element, knowledge registration, and KE (transfer). Based on the above results, we see that the single and multiplayer modes have merits and demerits, and Shelter GO can allow players to choose the mode based on their learning style and purpose, concluding that Shelter GO is effective in learning evacuation.

5. Conclusion

This study described a multiplayer location-based game (Shelter GO) for learning evacuation by not only remembering shelter locations, capacities, and features but also acquiring shelter-related and general knowledge for successful evacuation. The game prototype (a geofencing mobile app) adopts a game element similar to Pokémon GO and prompts players to visit and observe shelters by enabling them to collect digital creatures. In the multiplayer mode, players must register knowledge to the collected digital creatures and can exchange the registered knowledge (a corresponding creature) with the encountered players. We see that the small-scale experiment had some limitations and was insufficient to clarify the effect of KE. Therefore, we must conduct large-scale experiments involving a diverse range of participants. Additionally, we will refine the visual design of the app (e.g., more attractive digital creatures) to make it more appealing to the public. We concluded that, through the preliminary experiment with few participants, KE does not remarkably prevent learning evacuation, and it is effective in learning evacuation, as knowledge registration and exchange motivate players and enhance knowledge acquisition from the encountered players.

Acknowledgements

We thank Miyoshi, N. for his great effort in this study. This work was supported by the Japan Society for the Promotion of Science Grants-in-Aid for Scientific Research (Grant No. 18H01054).

References

Centre for Research on the Epidemiology of Disasters (CRED) (2022). Emergency Events Database (EM-DAT), https://www.emdat.be/ (Accessed on 2022.8.8)

Dransch, D., Rotzoll, H., & Poser, K. (2010). The contribution of maps to the challenges of risk communication to the public. *International Journal of Digital Earth*, *3*, 292-311.

Finco, M. D., Rocha, R. S., Fão, R. W., & Santos, F. (2018). Let's move!: The social and health contributions from Pokémon GO. *International Journal of Game-Based Learning*, *8*, 44-54.

Hamari, J., Malik, A., Koski, J., & Johri, A. (2019). Uses and gratifications of Pokémon Go: Why do people play mobile location-based augmented reality games? *International Journal of Human-Computer Interaction*, *35*, 804-819.

Harteveld, C. & Bekebrede, G. (2011) Learning in single-versus multiplayer games: The more the merrier? *Simulation & Gaming*, *42*(1), 43-63.

Lee, J. (2022). Problem-based gaming via an augmented reality mobile game and a printed game in foreign language education. *Education & Information Technologies*, *27*, 743-771.

Lee, J. E., Zeng, N., Oh, Y., Lee, D., & Gao, Z. (2021). Effects of Pokémon GO on physical activity and psychological and social outcomes: A systematic review. *Journal of Clinical Medicine*, *10*, 1860.

Lin, Y. T., Tseng, Y. M., Lee, Y. S., Wang, T. C., Tsai, S. I., & Yi, Y. J. (2018). Development of a SoLoMo game-based application for supporting local cultural learning in Taiwan. *Educational Technology & Society*, *21*(4), 115-128.

McCallum, I., Liu, W., See, L., Mechler, R., Keating, A., Hochrainer-Stigler, S., Mochizuki, J., Fritz, S., Dugar, S., Arestegui, M., Szoenyi, M., Bayas, J. L., Burek, P., French, A., & Moorthy, I. (2016). Technologies to support community flood disaster risk reduction. *International Journal of Disaster Risk Science*, *7*, 198-204.

Mitsuhara, H., Miyoshi, N., & Shishibori. M. (2022). Location-based game for remembering shelter locations, capacities and features, *Proceedings of ICTech-2022*, (in press).

Ruiz-Ariza, A., Casuso, R. A., Suarez-Manzano, S., & Martnez-Lpez, E.J. (2018). Effect of augmented reality game Pokémon GO on cognitive performance and emotional intelligence in adolescent young. *Computers & Education*, *116*, 49-63.

Song, J., Yamauchi, H., Oguchi, T., & Ogura, T. (2022). Application of web hazard maps to high school education for disaster risk reduction. *International Journal of Disaster Risk Reduction*, *72*, 102866.

Tran, K.M. (2018). Families, resources, and learning around Pokémon GO. *E-Learning & Digital Media*, *15*, 113-127.

Watanabe, K., Kawakami, N., Imamura, K., Inoue, A., Shimazu, A., Yoshikawa, T., Hiro, H., Asai, Y., Odagiri, Y., Yoshikawa, E., & Tsutsumi, A. (2017). Pokémon GO and psychological distress, physical complaints, and work performance among adult workers: A retrospective cohort study. *Scientific Reports*, *7*, 10758.

Wu, M.H. (2021). The applications and effects of learning English through augmented reality: A case study of Pokémon GO. *Computer Assisted Language Learning*, *34*, 778-812.

Yamamoto, K., & Fujta, S. (2015). Development of social media GIS to support information utilization from normal times to disaster outbreak times. *International Journal of Advanced Computer Science & Applications*, *6*(9), 1-14.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Evaluating the Performance of Chinese Multi-Label Grammatical Error Detection Using Deep Neural Networks

Tzu-Mi LIN[a], Chao-Yi CHEN[a], Lung-Hao LEE[a] & Yuen-Hsien TSENG[b*]

[a]*Department of Electrical Engineering, National Central University, Taiwan*[b]
Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taiwan
*samtseng@ntnu.edu.tw

Abstract: In this paper, we describe the process of building a benchmark data set for Chinese multi-label grammatical error detection tasks, comparing the performance of 10 representative neural network models. Experimental results reveal that no matter which deep learning model is used, the performance is still limited which confirms the difficulty of the multi-label detection task. Our constructed datasets and evaluation results will be publicly released on the GitHub repository (https://github.com/NCUEE-NLPLab/CMLGED) to promote further research to facilitate technology-enhanced Chinese learning.

Keywords: Grammatical error detection, multi-label classification, deep learning

1. Introduction

Chinese as foreign language learners make various kinds of grammatical errors during their language acquisition process. Coarse-grained grammatical error types, such as missing words, redundant words, incorrect word selection, or word ordering errors, originate from target modification differences from comparing sentence surface structure with the correct sentence. An effectively automated error detection system would facilitate Chinese learning. Previous Chinese grammatical error detection approaches were based on linguistic rules (Lee et al., 2013), machine learning classifiers (Liu et al., 2016), or their hybrid methods (Lee et al., 2014). Deep learning approaches had also been applied to detect Chinese grammatical errors (Lee et al, 2017; 2020; 2021). These studies assume that a learner's written sentence has only one grammatical error, when they may in fact contain multiple errors of different types. This motivates us to explore Chinese multi-label grammatical error detection.

This study describes our process of building a multi-label grammatical error dataset from the TOCFL learner corpus (Lee et al., 2018). Recent deep-learning models have achieved excellent results in many natural language processing tasks, so we evaluate the performance of 10 representative neural networks on this dataset. The dataset and results will be made publicly available to promote the research development for technology-enhanced language learning.

2. Benchmark Data Construction and Evaluation

The benchmark data was taken from the TOCFL learner corpus (Lee et al., 2018), including grammatical error annotation of 2,837 essays written by Chinese language learners with 46 different mother-tongue languages. In each essay, we divided each paragraph into individual sentences, splitting based on the newline symbol and punctuations including an exclamation marks, question marks, or full stops. We selected sentences containing at least one of four error labels: **M**issing words (denoted as M), **R**edundant words (R), incorrect word **S**election (S), and **W**ord ordering error (W). This filtering process left us with 19,546 sentences with a total of 27,189 labels. Each sentence contains average 27.9 characters and 1.39 labels. About two-third all sentences (13,265/67.8%) were annotated with only one label, followed by two labels (5,024/25.7%), three labels (1,152/5.9%), and four labels (105/0.005%).

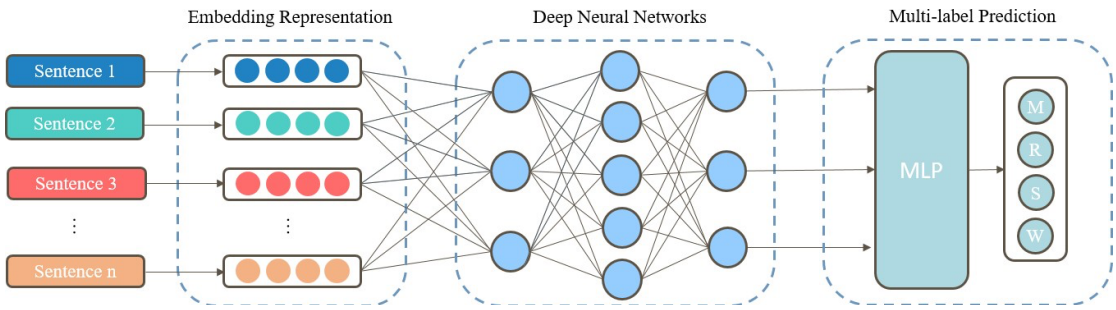


Figure 1. Performance evaluation workflow.

Figure 1 shows our experimental architecture for Chinese multi-label grammatical error detection. For word embedding representation, we dumped Chinese Wikipedia (from March 4, 2021) and automatically segmented words. Words occurring in Wikipedia at least 5 times were retained for embedding training. We pretrained embedding vectors using the GloVe technique (Pennington et al. 2014), obtaining 1,126,163 distinct word vectors, where the dimension of vectors is 300. These word vectors in sentences were then regarded as initial representations to feed into deep neural networks.

Our experimental neural networks can be further divided into three types: 1) Stacked-based: conventional neural network architectures, including traditional BiLSTM (Graves et al., 2013), CNN (Kim, 2014), and fastText (Bojanowski et al., 2017); 2) Graph-based: a class of neural networks used for processing data represented by graph data structures. We used the Graph-CNN (Defferrard et al., 2016), TextGCN (Yao et al., 2019), and HyperGAT (Ding et al., 2020); 3) Transformer-based: a type of pre-trained model that uses the encoder and self-attention mechanism. We compared the BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019) and ELECTRA (Clark et al., 2020). All the sentences with their labeled classes are used to train our deep neural networks to automatically learn all the corresponding parameters. To predict the error classes of a sentence during the testing phase, the sentence unseen in the training phase goes through the neural network architecture to yield a probability value corresponding to each class (the activation is a sigmoid function). The class with a probability exceeding 0.5 will be returned as the prediction results.

3. Experimental Results

Table 1. *Evaluation on Chinese Grammatical Error Detection*

Neural Network Models		Micro F1	Macro F1	Weighted F1	Subset Accuracy
Stack-based	BiLSTM	0.4827	0.2993	0.4261	0.2258
	CNN	0.5035	0.3021	0.4338	0.2313
	fastText	0.436	0.2355	0.3457	0.1844
	Ave.	*0.4741*	*0.2790*	*0.4019*	*0.2138*
Graph-based	Graph-CNN	0.4810	0.3972	0.4801	0.1260
	TextGCN	0.4242	0.3436	0.4228	0.1767
	HyperGAT	0.5078	0.2972	0.4246	0.2461
	Ave.	*0.4710*	*0.3460*	*0.4425*	*0.1829*
Transformer-based	BERT	0.5698	0.4973	0.5645	0.3374
	RoBERTa	0.5670	0.4954	0.5612	0.3287
	XLNet	0.5362	0.4495	0.5278	0.3030
	ELECTRA	**0.6147**	**0.5059**	**0.5978**	**0.3739**
	Ave.	*0.5719*	*0.4870*	*0.5628*	*0.3358*

Table 1 shows the experimental results of 5-fold cross-validation. Similar to other natural language processing findings, transformer-based models significantly outperformed graph-based or

stack-based neural networks, regardless of the evaluation metrics used. Although the ELECTRA model achieved the best performance among our 10 evaluated neural networks, final results still need considerable improvement.

4. Conclusions

This study describes the process of building a benchmark data set for the Chinese multi-label grammatical error detection task, comparing the performance of 10 representative neural network models. Experimental results indicate considerable room for improvement regardless of which model is used. Our constructed datasets and evaluation results are publicly available at https://github.com/NCUEE-NLPLab/CMLGED to promote further research for technology-enhanced Chinese language learning.

In addition to developing data-intensive computing models such as neural networks, future work will focus on extracting discriminatively linguistic features for the detection task.

Acknowledgements

This study was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3, MOST 109-2410-H-003-123-MY3, MOST 111-2628-E-008-005-MY3.

References

Bojanowski, P., Grave, E., Joulin, A., & Mikolov T. (2017). Enriching word vectors with subword information. *Transactions of the Association of Computational Linguistics*, 5, 135-146.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: pre-training text encoders as discriminators rather than generators. *Proceedings of ICLR'20*, https://arxiv.org/abs/2003.10555

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Proceedings of NIPS'16* (pp. 3844-3852), Barcelona, Spain.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint, https://arxiv.org/abs/1810.04805

Ding, K., Wang, J., Li, J., Li, D., & Liu, H. (2020). Be more with less: hypergraph attention networks for inductive text classification. *Proceedings of EMNLP'20* (pp. 4927-4936), Online: ACL Anthology.

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *Proceedings of ICASSP'13* (pp. 6645-6649), Vancouver, Canada: IEEE Digital Library.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of EMNLP'14* (pp. 1746-1751), Dohar, Qatar: ACL Anthology.

Lee, L.-H., Chang, L.-P., Lee, K.-C., Tseng, Y.-H. & Chen, H.-H. (2013). Linguistic rules based Chinese error detection for second language learning. *Proceedings of ICCE'13* (pp. 27-29), Bail, Indonesia: APSCE.

Lee, L.-H., Hung, M.-C., Chen, C.-Y., Chen, R.-A., & Tseng, Y.-H. (2021). Chinese grammatical error detecting using adversarial ELECTRA transformers. *Proceedings of ICCE'21* (pp. 111-113), Online: APSCE.

Lee, L.-H., Lin, B.-L., Yu, L.-C., & Tseng, Y.-H. (2017). Chinese grammatical error detection using a CNN-LSTM model. *Proceedings of ICCE'17* (pp. 919-921), Christchurch, New Zealand: APSCE.

Lee, L.-H., Tseng, Y.-H., & Chang, L.-P. (2018). Building a TOCFL learner corpus for Chinese grammatical error diagnosis. *Proceedings of LREC'18* (pp. 2298-2304), Miyazaki, Japan: ACL Anthology.

Lee, L.-H., Wang, Y.-S., Lin, P.-C., Hung, C.-T., & Tseng, Y.-H. (2020). Multi-channel CNN-BiLSTM for Chinese grammatical error detection. *Proceedings of ICCE'20* (pp. 558-560), Online: APSCE.

Lee, L.-H., Yu, L.-C., Lee, K.-C., Tseng, Y.-H., Chang, L.-P., & Chen, H.-H. (2014). A sentence judgment system for grammatical error detection. *Proceedings of COLING'14* (pp. 67-70), Dublin, Ireland: ACL Anthology.

Liu, Y., Han, Y., Zhou, L., & Zan, H. (2016). Automatic grammatical error detection for Chinese based on conditional random field. *Proceedings of NLPTEA'16* (pp. 57-62), Osaka, Japan: ACL Anthology.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V. (2019). RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint, https://arxiv.org/abs/1907.11692

Pennington, J., Socher, R., & Manning C. (2014). GloVe: global vectors for word representation. *Proceedings of EMNLP'14* (pp. 1532-1543), Doha, Qatar: ACL Anthology.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: generalized autoregressive pretraining for language understanding. arXiv preprint, https://arxiv.org/abs/1906.08237

Yao, L., Mao, C., & Luo, Y. (2019). Graph convolutional networks for text classification. *Proceedings of the AAAI'19* (pp. 7370-7377), Honolulu, Hawaii, USA.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

An Exploration of a Novice Kindergarten Teacher's Enactment of Multiliteracies Pedagogy during the Pandemic: A Case Study of a Virtual Kindergarten Classroom

Fengchao YU, Lian CHANG & Mi Song KIM*

Curriculum Studies, Faculty of Education, Western University, Canada
*mkim574@uwo.ca

Abstract: We aim to gain insight into technology-enhanced literacy learning for kindergarten students during the COVID-19 pandemic by exploring a novice kindergarten teacher's practice of multiliteracies pedagogy in his virtual kindergarten classroom. This qualitative case study collected data from multiple sources such as virtual interviews and classroom observations, the Kindergarten Program (KP) document, teacher's reflective notes, lesson plans, students' artefacts, and researchers' observational notes and reflective journals. This study found that although the novice kindergarten teacher provided various multimodal learning opportunities for students, his literacy practice emphasized phonological awareness, phonemic awareness, and letter-sound correspondence. Also, he faced numerous challenges due to inadequate teacher preparation and professional development, inconsistency of the quality and utility of technology, constraints of virtual learning for young learners, varying degrees of parental support, and challenges of implementing multiliteracies pedagogy with young children virtually. This study contributes to the existing literature on online learning for kindergarten students and expands the burgeoning multiliteracies research from physical to virtual learning environments. Also, this study demonstrates how virtual learning opens up opportunities to advance the multiliteracies pedagogy and highlights the importance of strengthening teacher education programs and providing continuous professional development for teachers.

Keywords: Technology-enhanced literacy learning, multiliteracies, multimodality, virtual learning, kindergarten

1. Introduction

Literacy education has gone through tremendous changes over the past few decades (Bull & Anstey, 2018; Cope & Kalantzis, 2009). The term literacy was first defined as the ability to read and write print-based texts in the late 1800s, and literacy education emphasized teaching knowledge and skills such as "grammar, spelling and punctuation and comprehension" (Anstey & Bull, 2018, p. 5). However, as a response to the increasingly diversified modes of communication and multiple forms of culture and language, the New London Group (NLG) (1996) coined the term *multiliteracies*, referring to the ability to communicate with and make meaning of the multimodal texts, and proposed multiliteracies pedagogy to reconceptualize literacy teaching and learning in a globalized environment. The multiliteracies pedagogy expands literacy education to embrace the diversity that learners bring from their social, cultural, and linguistic backgrounds and the multimodal forms (e.g., visual, audio, linguistic, spatial, gestural) that learners adopt for meaning-making and representation (Cope & Kalantzis, 2009).

The multiliteracies pedagogy comprises four components: (1) situated practice; (2) overt instruction; (3) critical framing; and (4) transformed practice (NLG, 1996). Specifically, situated practice underscores the importance of providing opportunities and meaningful experiences for students to draw on their prior knowledge and out-of-school experience to make meaning in new contexts. Overt instruction refers to active interventions that scaffold students' learning by making

explicit the patterns of meaning to help students gain a conscious understanding of their learning. Critical framing encourages students to interpret the designs of meaning from a critical perspective and interrogate its situated social-cultural contexts and purpose. Transformed practice occurs when learners transfer what they have learned in one context to another and shift their roles from consumers of the knowledge to designers and meaning makers (NLG, 1996). It is worth noting that these pedagogical orientations are not in singular forms or linear sequences but are a knowledge process of "weaving between different knowledge processes in an explicit and purposeful way" (Cope & Kalantzis, 2009, p. 187).

Aside from the four fundamental components, central to the multiliteracies pedagogy is the notion of design (Cope & Kalantzis, 2009). As the forms with which people interact and make meaning in their everyday life have become increasingly multimodal (NLG, 1996), meaning-making and representation can be conceived as "a form of design" that allows learners to move freely between different modes (Cope & Kalantzis, 2009, p. 166). Also, since design consists of "teachers' pedagogic designs of learning processes and students' designed constructions of meaning," it is very useful for "theoriz[ing] the relationships between modes, pedagogy, and context," and "understand[ing] the changed dispositions toward information and knowledge" (Jewitt, 2008, p. 252).

Literacy education has been challenged unprecedentedly with the outbreak of the COVID-19 pandemic. The school closure has forced learning at various levels to shift online, including kindergartens. This move has triggered heated discussions over the impacts of online learning on students' academic performance (Dore et al., 2021; Hu & Lu, 2020), particularly literacy loss (Bao et al., 2020; Chamberlain et al., 2020). Therefore, this study aims to gain insight into technology-enhanced literacy learning for kindergarten students during the pandemic by exploring a novice kindergarten teacher, Michael's (pseudonym) practice of multiliteracies pedagogy in his virtual kindergarten classroom and the challenges he faced. Specifically, we asked the following research questions: 1) How did the novice kindergarten teacher enact multiliteracies pedagogy in his virtual kindergarten classroom? 2) What challenges did the novice kindergarten teacher face when implementing multiliteracies pedagogy in his virtual kindergarten classroom?

2. Literature Review

2.1 Multiliteracies Practice in Kindergarten

Over the years, scholars all over the world have drawn upon the multiliteracies framework to examine literacy practices in diverse socio-cultural contexts (e.g., Anstey & Bull, 2018; Baguley, Pullen, & Short, 2010; Jewitt, 2008). For instance, in a kindergarten to Grade 5 setting in an Innu community, Lavoie et al. (2012) illustrated how the conscious introduction of multiliteracies pedagogy and its enactment allowed teachers to acknowledge and respect the identity formation, meaning-making, and expression of culturally and linguistically diverse students by drawing on their lifeworld experiences and the available multimodal and multilingual resources in their local community. Taylor and Leung (2020) also described the experiences of teachers who promoted multimodal learning experiences reflective of students' social and cultural backgrounds. As our previous study (Author, 2021) indicated, through multiliteracies as a culturally responsive pedagogy, teachers need to intentionally create spaces and learning experiences that encourage the development of multimodal literacy and further develop socialization skills. However, as Zhang et al.'s (2019) systematic review on multiliteracies practice showed, teachers faced various challenges when enacting multiliteracies pedagogy, such as the lack of "training in recruiting multimodal resources and designing learning environments," "insufficient, inappropriate, or missing materials," and contradiction between multiliteracies pedagogy and the prescriptive literacy curriculum (p. 48). Also, of the 66 reviewed studies, only six were conducted in kindergarten settings. They thus called for more multiliteracies research "in various contexts of the globalized world, including virtual spaces" (p. 41). Since most of the studies on multiliteracies practice were conducted in physical classrooms before the pandemic, it is critical and time-sensitive to explore the implementation of multiliteracies pedagogy in a virtual kindergarten classroom.

2.2 Online Learning for Young Children

Online learning that uses various educational platforms such as email for asynchronized learning and

Google Classroom, Google Teams, Zoom, and SeeSaw to support students' synchronized learning (Hu & Lu, 2020) is a valuable alternative to teaching and learning during school closures (Abernathy & Thornburg, 2021; Dore et al., 2021). However, since online learning was uncommon for young children before the pandemic, teachers faced various challenges. For instance, teachers generally lacked sufficient training and preparation and struggled with engaging students in an online learning context (Ewing & Cooper, 2021). Moreover, the availability of technological tools and the accessibility and stability of the internet posed challenges for effective online learning (Fitri & Latif, 2021). Also, as young children were not equipped with relevant knowledge and/or skills to handle online learning, parents' guidance and support were expected. However, parents often provided either too much help or inadequate support for their children due to their beliefs and attitudes about online learning and digital literacy (Azizah & Eliza, 2021; Firmanto et al., 2020).

3. Methods

As part of a larger research project hosted by the third author, we aim to explore a novice kindergarten teacher, Michael's (pseudonym) practice of multiliteracies pedagogy in his virtual kindergarten classroom in Ontario, Canada, where kindergarten programs (KP) were guided by the Ontario KP document. Teacher participant Michael was an Ontario Certified Teacher, teaching kindergarten to Grade 6, and a Master student in Education. Michael was in his first Long-Term Occasional full-year contract position when the study took place. He partnered with two part-time early childhood educators (ECE) to teach 23 students, among whom five parents granted us consent. We employed a qualitative case study to capture the complexity of the implementation of multiliteracies pedagogy as it allows "a rich portrayal of insights and understandings interpreted in the particular context" (Simons, 2014, p. 466). Michael learned about multiliteracies pedagogy in his Master's program, which made him realize the importance of diversifying literacy practices in an increasingly globalized society and eventually motivated him to adopt multiliteracies pedagogy in an online environment with his kindergarten students. Interviews and classroom observations were conducted virtually, and data collection lasted for four months until data saturation occurred (Saunders *et al*., 2018). To help "reduce bias and increase confidence in the robustness of the research results" (Kipping et al., 2013, p. 312), we also analyzed data from multiple sources such as the KP document, lesson plans, student artefacts, Michael's reflective notes in computer-supported collaborative learning platform (i.e., Knowledge Forum, KF), and researchers' observational notes and reflective journals. Drawing on the multiliteracies pedagogy, we employed deductive and inductive thematic analysis to go through each of these phases: familiarizing with our data, (2) generating initial codes, (3) searching for themes, (4) reviewing themes, and (5) defining and naming themes (Braun & Clarke, 2006).

4. Findings

4.1 The Practice of Multiliteracies Pedagogy in a Virtual Kindergarten Classroom

Deductive data analysis found that situated practice, overt instruction, and transformed practice were evident in Michael's multiliteracies practice. Also, inductive data analysis revealed three emerging themes: multimodality and design, collaboration with teaching partners and parents, and the influence of the KP document on his literacy practice.

4.1.1 Situated Practice

In multiliteracies practice, the situated practice encourages the teacher to integrate students' interests, funds of knowledge and out-of-school experiences into teaching, so the learning experiences are authentic and meaningful for students (NLG, 1996). Michael asked his students what they wanted to learn at the beginning of the semester, organized their inquiry interests into learning activities (e.g., living things/animals, superheroes/dress up, dance/ballet, art/craft, building/construction), and provided relevant learning experiences for students. Also, he embraced students' lifeworld experiences in his teaching and took their culture and family situations into consideration in curriculum design. For instance, Mothers' Day or Fathers' Day could be a special moment between a child and their grown-ups, but as Michael explained in the post-interview, sometimes students did not

have a father or a mother in their lives. So, he proposed a Special Person Day to honour anyone like grandmas or neighbours (May 28, 2021) (Figure 1).



Figure 1. Mother's Day Artefacts and Special Person Day Template

Situated practice also "require[s] the skillful scaffolding or temporary support structures of experts to enable learners to draw upon the cultural resources for meaning making necessary to transfer knowledge to new, multimodal designs" (Mills, 2006, p. 28). For example, the Writer's Workshop in Michael's class allowed his students to use multimodal forms and different semiotic tools to express their ideas. At the beginning of the semester, students were asked to draw a picture and write something about it with the teacher's help, such as their names, dates, letters, or words to label the pictures. As the semester went on, students were asked to write kindergarten sentences, meaning that when the students heard words phonetically, they tried to sound out these words and spell them out phonetically how they heard them. Also, educators explicitly talked about finger spacing, capital letters, punctuation, etc.

4.1.2 Overt Instruction

Overt instruction refers to the explicit instruction of the patterns of meaning by identifying students' specific needs (NLG, 1996). In the post-interview, Michael noted the importance of "explicit instruction that is evidence-based in terms of phonemic, phonological awareness, phonemic awareness, and letter-sound correspondence, in a scope and sequence manner" (May 28, 2021). For sight words instruction, he also highlighted the importance of meaning-focused activities by providing "an example of a sentence for that comprehension" and explaining "why do we use that word and what does it mean" (May 28, 2021). Further, overt instruction comes with new assessment techniques. Specifically, evaluation "should be developmental" and serves as "a guide to further thought and action" (NLG, 1996, p. 86). Michael also noted that "the direct instruction is reflective of the assessments" and could help identify "where the child excels in," "where the child's strengths are," "where do they still require more," and "(where) requires next step to continue with" (Post-interview, May 28, 2021). Additionally, the assessment allowed Michael and his team to "see areas that require more in-depth review" and assist them in curriculum planning (KF, January 18, 2021). Therefore, he carefully observed and documented his students' learning progress in a Writer's Workshop.

4.1.3 Transformed Practice

Transformed practice occurs when learners transfer what they have learned to new contexts and shift their roles from consumers of the knowledge to designers and meaning makers (NLG, 1996). Michael provided opportunities for students to apply what they learned in class to new learning situations. For example, he asked students to get an old magazine and look for sight words or write down all the letters they know and teach their stuffies the letters and sounds. In free reading time, Michael encouraged students to share pictures, interesting facts, ideas, or literacy concepts learned in class. He

also encouraged students to take their learning outdoors by going on nature walks and recording their observations. According to Cope and Kalantzis (2009), situated practice and transformed practice were sometimes intertwined, which could be seen in Michael's enactment of a multiliteracies pedagogy. Since "students usually engage in the transformed practice stage by designing multimodal texts" (Rajendram, 2015, p. 2), the Writer's Workshop that exemplified Michael's situated practice was also a good example of the transformed practice.

4.1.4 Multimodality and Design

Multiliteracies pedagogy stresses the importance of utilizing multiple modes (e.g., visual, audio, linguistic, spatial, gestural, etc.) to make meaning and communicate as "much of our everyday representation experience is intrinsically multimodal" (Cope & Kalantzis, 2009, p. 179). As Michael's class was primarily conducted in the synchronized online form, he heavily relies on technological tools such as Smart Notebooks, Google Slides, songs, YouTube videos, dance, and movement in his teaching practice. He also integrated multimodal forms using audio, video, and art in his teaching. For example, reading a physical book was difficult for students to see, so Michael used the video-based read-aloud on YouTube instead. To promote students' engagement when practicing sight words, he integrated cartoon GIFs (i.e., someone jumping, hacking, squatting, push-ups) into his teaching. Specifically, after doing a sight word, they exercised for 20 seconds and repeated the cycle. He also tried different strategies, such as saying the sight words in a funny voice like Mario or a monster.

As design encompasses 'teachers' pedagogical designs of learning processes and students' designed constructions of meaning' (Jewitt, 2008, p. 252), Michael used "a lot of resources from everywhere," including the self-purchased resources, the kindergarten document, and the electronic resources provided by the school board" (Pre-interview, April 9, 2021). Also, Michael and ECEs carefully designed Choice Board activities in which they explicitly asked parents to choose activities that align with their child(ren)'s interests and can be easily implemented with materials available at home (Figure 2). As Michael stated in the pre-interview, "we try to be cognizant of… Obviously, we don't want our parents to spend enormous amounts of money. We just try basics that can be used in all different aspects. So, we try our best, to provide parents with opportunities, either with recycled materials, or things around the home" (April 9, 2021).

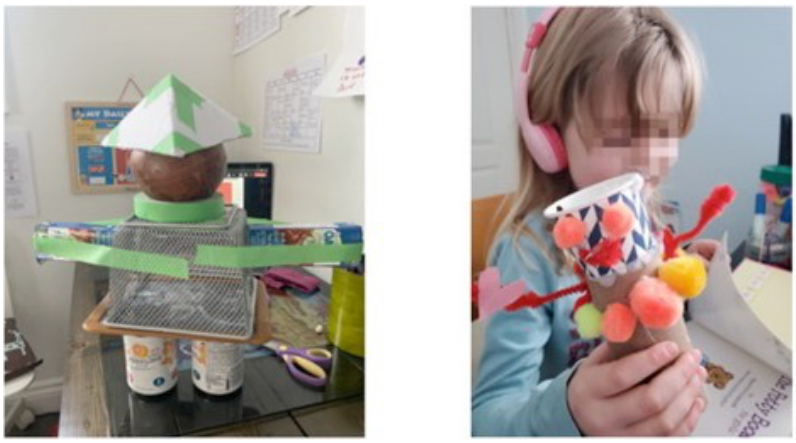


Figure 2. Robots Using 3D Shapes and Recyclables at Home

4.1.5 Collaboration with Teaching Partners and Parents

In Michael's virtual classroom, ECEs played an indispensable role. In addition to leading instructions when needed, ECEs were also responsible for conducting pedagogical documentation. Moreover, Michael and ECEs worked as a team to assess student learning by using Early Literacy Assessment (ELA) and other assessment tools. Once Michael completed the report drafts, the ECEs would read the report, point out any typos or errors, and provide other recommendations. As Michael said, "these progress reports would have been much more challenging if the ECEs did not have the well-done daily pedagogical documentation" (Post-interview, May 28, 2021). Further, since kindergarten

students need guidance from their caregivers (Azizah & Eliza, 2021; Firmanto et al., 2020), parents played a critical role in Michael's virtual classroom. Based on the KP document, Michael provided parents with a monthly checklist and examples of what could be evidence of their child(ren)'s learning, so Michael and ECEs could "assess the child(ren)'s learning and growth" (Pre-interview, April 9, 2021).

4.1.6 The Influence of the KP Document on the Teacher's Literacy Practice

The Kindergarten Program (KP) document greatly influenced Michael's virtual classroom. The KP document, issued by the Ontario Ministry of Education (2016), suggests "thinking about literacy in the broadest possible way," viewing children's communication "through gestures, physical movements, words, symbols, and representations, [and] the use of a variety of materials," and seeing children's literacy behaviours as "evident in the various ways they use language, images, and materials to express and think critically about ideas and emotions" (p. 64). The definition of literacy in the Ontario KP document is consistent with the key components of multiliteracies pedagogy, namely design and multimodality. As such, the KP document recommends the literacy practices that encourage students' expressions and practices of skills needed for 21st century learning. The KP document stressed the importance of providing rich and engaging learning experiences that build on children's strengths, prior knowledge, and experiences and working collaboratively with families. Also, the KP document listed examples of critical literacy behaviours of kindergarten students, such as "matching spoken words with written words; using familiar or high-frequency words; using approximate spellings of words based on their ability to hear, identify, and manipulate sounds (phonological and phonemic awareness) and on their knowledge of letter-sound correspondence (phonics)" (p. 67). In our pre-interview, Michael stressed several times that he has used the KP document as a "guide" and has "looked at all the learning goals to see what has missed, if he missed anything." For Michael, phonological awareness, phonemic awareness, and letter-sound correspondence were the "basic fundamental skills" that he "want(ed) to make sure that the students know and have" (April 9, 2021). Regarding the expected learning outcomes for students, Michael stated that "the only one was I wanted to make sure that we, the students, were increasing in terms of their phonological awareness and phonemic awareness and that we were progressing in letter-sound correspondences" (Post-interview, May 28, 2021). As described earlier, Michael's virtual class encompasses various activities to foster students' phonological awareness, practice sight words, recognize word families, and learn rhyming words. He also provided electronic resources with all these blends for parents so their child(ren) could continue to practice these skills. As Michael reflected, "[w]e as a team and our parents have noticed tremendous gains in our student phonological awareness, phonemic awareness, and letter-sound correspondences" (KF, April 2021).

4.2. Challenges the Novice Kindergarten Teacher Faced

Overall, Michael felt that the virtual literacy teaching was "a very rewarding experience" (Pre-interview, April 9, 2021). However, due to the complexity of enacting multiliteracies pedagogy with young children in a virtual learning environment, it is interesting to note that the challenges Michael encountered in implementing multiliteracies pedagogy in his virtual classroom was largely in line with the research findings from the literature on online learning for young children in general. Specifically, these challenges were: insufficient preparation in the teacher education program, inadequate professional development from the current work, varying degrees of parental support, and constraints of online learning for young children. The research also revealed the unique challenges of implementing multiliteracies pedagogy with young children virtually.

4.2.1 Insufficient Teacher Preparation and Professional Development

Consistent with the literature, the lack of training and preparation for online teaching was one of the challenges Michael encountered (Ewing & Cooper, 2021). In the post-interview, he noted that his teaching certification did not do its justice in preparing him to be a kindergarten teacher. The Bachelor of Education program he took was ideologically driven, and no evidence-based research to support it. As a result, he has "learned very little," and he "never knew about phonological awareness, phonemic

awareness" when he was a student. So, he had to look it up and construct it on his own because "obviously the board didn't provide it, or we didn't learn it in our teacher education program." When asked if he has learned anything about the multiliteracies pedagogy in the teacher education program, Michael responded that "I don't think so. Maybe it was very shortly talked about, but it was nothing in-depth. Maybe it was brushed over or something like that" (May 28, 2021). It needs to be pointed out that Michael was teaching the kindergarten and doing his master's degree in curriculum studies when the research was conducted, which provided a great opportunity for Michael to apply what he had learnt in his master's program into his teaching practice.

Michael also encountered inadequate professional development from the current work. When asked in the pre-interview if the school board provided any professional development opportunities for things related to literacy, multiliteracies, or even vocabulary, Michael expressed that "nothing from the PD days." Even though the school board has provided many electronic resources like Headsprouts and Epic, teachers were "just given those resources," but "there's no training for it," no professional development or anything like that in terms of interactive devices or electronic resources. Michael recalled his last PD Day and said that "we had the coordinator provide that phonological, phonemic awareness, letter-sound correspondences, and the pilot for Haggerty, but that's about it." Instead of doing a "one and done" type of professional development, Michael mentioned that the school board "would need to train all the current teachers now and then each year" (April 9, 2021). Also, Michael believed that "educators require more workshops… and many would come in to learn how to incorporate meaningful and purposeful technology to augment learning" (KF, December 2020). However, due to the pandemic, of the seven professional development days allocated for teachers, three were used for health and safety training.

Although multiliteracies pedagogy is recommended, it is not mandatory for teachers in his school board to implement it. However, as a passionate, self-driven and reflective teacher, Michael took his time to look for research and evidence-based instructions from the school library where he was doing his master's degree, Google Scholar, and other resources. He believed that the teaching profession should "be highly regarded as doctors and lawyers, who are in their own time keeping up with the literature or those case studies and stuff like that" (Post-interview, May 28, 2021). Also, Michael constantly reflected on his teaching practice, asking himself what else could have been done differently, and collaborated with his ECE partners to make changes along the way and refine his teaching skills.

4.2.2 *Varying Degrees of Parental Support*

Like any other online learning with young children, the implementation of multiliteracies pedagogy requires the support of caregivers. For instance, Michael intended to embed active learning opportunities such as singing, dancing or hands-on experiences into his multiliteracies pedagogy, which needed the collaboration of parents to encourage their children to implement these activities at home. However, as Michael said, "it's just difficult virtually" (April 9, 2021). Similar challenges were found when requesting parents to submit children's work for assessment. For example, Michael asked parents to describe their child(ren)'s work when submitting them as it would allow educators to see the students' thinking processes. However, some parents submitted students' work based on the checklist educators provided and added detailed descriptions to the work. In contrast, others submitted a few works with little or no descriptions or did not submit at all. Also, when the weather was getting better and families started enjoying the warmer weather, Michael noticed the changes in student attendance. Michael has expressed concerns because they were doing ABC Bootcamp that introduced students to all 26 letters and sounds in 26 days. "Student attendance is necessary" for the ABC Bootcamp to be effective. However, if parents took a vacation or the child was sick in extenuating circumstances, the child would miss one letter a day. It would be hard to combat that problem later as "this is extremely crucial for the child's education, and it will only compound after" (Post-interview, May 28, 2021).

4.2.3 *Constraints of Online Learning*

As Michael reflected in the KF, one challenge his team faced during online teaching was "the quality of the video and sound transmission" and the internet connection of teachers and/or students. The

picture was very blurry and choppy during class lessons when everyone's camera was on. Michael believed that high-quality internet connections and technological devices are required to provide students with high-standard education (December 2021). Further, Michael found that he could not see the students well while he was teaching and sharing his screen, let alone see every child's work when the educators wanted the students to do some extended activities. Also, it was difficult to pause and ask questions or draw students' attention to some unknown words during read-aloud in a virtual classroom, whether reading a physical book or playing a read-aloud video. As Michael said, "read alouds are important, it's just difficult virtually" (Post-interview, May 28, 2021). Similar challenges could be found when teaching the sound of the letter to students. Michael reflected that "with distance learning, students must keep their microphones muted unless called upon" because of the background noises in their homes (KF, December 2021). However, "it's just by the time, you ask a question and then you pick a child and then the child unmutes themselves and then they respond and then they mute themselves, it just takes a long long time, right?" (Post-interview, May 28, 2021). Additionally, the implementation of interactive software was challenging in his kindergarten since "only about ⅓ of my students would be able to participate in an interactive virtual activity" and "there are still a few students where their parents sit in our class and unmute/mute when the student is called upon" (KF, December 2021).

4.2.4 *Challenges of Implementing Multiliteracies Pedagogy with Young Children Virtually*

In addition to the above-mentioned challenges posed by online learning for young children in general, Michael also encountered some unique challenges in implementing multiliteracies pedagogy with young children. Two important components of multiliteracies include "increasing multiplicity and integration of significant modes of meaning-making" and "increasing local diversity and global connectedness" (NLG, 1996, p. 64). Although Michael took advantages of the affordances of digital technologies and provided various multimodal learning opportunities for students' meaning-making and representation, there was a lack of autonomy and social interaction for children to explore and manipulate technology and discuss their learning with peers. Michael stated this level of independence and peer collaboration was due to the age of his students and the learning curve of virtual technology. In addition, as one of the key components of multiliteracies, critical framing was not utilized by Michael in this study. Under critical framing, teachers create learning experiences and encourage students to view design in its situated social cultural context (NLG, 1996). This level of higher order thinking can be challenging for children to grasp as it is crucial for them to learn the basics before moving towards this step.

5. Discussion & Implications

Our research findings show that Michael's multiliteracies practice highlighted situated practice, multimodality, and design which are essential aspects of multiliteracies practice (Cope & Kalantzis, 2009). For example, he integrated students' interests into his curriculum plan, embraced students' lifeworld experiences and family traditions, and provided opportunities for students to express their ideas and cultural identities through multimodal forms and different semiotic systems. Moreover, he took advantage of the affordance of online learning and incorporated a wide variety of modes and learning forms in his literacy practice. Particularly, he encouraged students to use available resources like recycled materials at home to make meaning and transfer learning to new situations. Also, his multiliteracies practice was significantly influenced by the KP document that emphasizes phonological awareness, phonemic awareness, and letter-sound correspondence.

The findings also revealed the complexity of implementing multiliteracies pedagogy in a virtual kindergarten classroom. Michael faced similar challenges as those who enacted multiliteracies pedagogy in physical classrooms, such as lack of appropriate training in multiliteracies pedagogy and the contradiction between prescriptive curriculum and multiliteracies pedagogy (Zhang et al., 2019). As a novice teacher, neither Michael's Bachelor of Education program nor his current work prepared him for enacting multiliteracies pedagogy or conducting online teaching. Michael's literacy practice centred around the expectations of the KP document. Meanwhile, he also faced similar challenges

identified from the literature on online learning for young children. For example, the lack of training and preparation for online teaching (Ewing & Cooper, 2021), varying degrees of parental support (Firmanto et al., 2020), and the constraints of online learning for young children (Fitri & Latif, 2021). This study also highlighted the importance of collaborating with teaching partners in the virtual learning environment, which was not evident from the existing literature.

Thus, our findings enrich the extant literature on online learning for young children during the pandemic. This study also contributes to the burgeoning research on multiliteracies practice while expanding it to virtual learning environments in kindergarten settings. Our study shows how virtual learning opens up opportunities to advance multiliteracies pedagogy. Through the exploration of the lived experience of a novice kindergarten teacher, our qualitative case study may broaden teachers' conceptions of literacy activities and persuade them to diversify literacy activities and embrace technological tools in physical classrooms in meaningful ways. Further, this study reveals the challenges the novice kindergarten teacher faced in enacting multiliteracies pedagogy in his virtual kindergarten classroom and calls for the actions from relevant stakeholders to strengthen teacher education programs and provide continuous training and development for teachers. Since communication and meaning-making are getting increasingly multimodal and learners come from diversified social, cultural, and linguistic backgrounds, it is critical for teachers to have multiliteracies pedagogy in their repertoire. Also, as the pandemic is still going on and the future is uncertain, it is imperative to integrate online teaching into the existing teacher education curricula and professional development plan to better prepare teachers for the future.

This study is limited to one novice kindergarten teacher and five students whose parents granted consent for data collection. Although data from multiple sources were collected for triangulation and the trustworthiness of the study, the research findings may be only applicable to other similar research settings. Therefore, working with diverse participants such as novice and experienced teachers and students from various cultural and linguistic backgrounds may enable more generalizable research findings. This study only looked at the implementation of multiliteracies pedagogy in a virtual kindergarten classroom from a novice teacher's perspective. Future research may examine other relevant stakeholders' perspectives like students, parents, the school board, etc. Also, future research could explore the actualization of multiliteracies pedagogy in virtual kindergarten classrooms with children from culturally and linguistically diverse backgrounds. Another limitation of the study is its short data collection period. This study may miss the possible changes in Michael's multiliteracies practice as he accumulates more experiences and becomes more comfortable with online teaching. Thus, a follow-up study may reveal the changes that the current study may have missed.

Acknowledgements

We are grateful for the support from the Social Sciences and Humanities Research Council of Canada and Grants from the University of Western Ontario.

References

Abernathy, D. F., & Thornburg, A. W. (2021). The K-12 teacher's perspective on teaching and learning during the COVID-19 pandemic. In Thornburg A. W., Ceglie, R. J., & Abernathy, D. F., *Handbook of research on lessons learning from transitioning to virtual classrooms during a pandemic* (pp. 1-15). IGI Global.

Anstey, M., & Bull, G. (2018). *Foundations of multiliteracies: Reading, writing and talking in the 21st century*. Abingdon, Oxon: Routledge.

Azizah, A., & Eliza, D. (2021). The role of parents in stimulating the development of early children's literation in pandemic time (COVID-19). *International Journal of Progressive Sciences and Technologies*, *24*(2), 154-159.

Baguley, M., Pullen, D. L., & Short, M. (2010). Multiliteracies and the new world order. In Pullen, D. L. & Cole, D. R. *Multiliteracies and technology enhanced education: Social practice and the global classroom.* (pp. 1-17). IGI Global.

Bao, X., Qu, H., Zhang, R., & Hogan, T. P. (2020). Literacy loss in kindergarten children during COVID-19 school closures. *International Journal of Environmental Research and Public Health*, *29*(5), 1-8.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Bull, G., & Anstey, M. (2018). *Elaborating multiliteracies through multimodal texts: Changing classroom practices and developing teacher pedagogies*. Boca Raton, FL: Routledge.

Chamberlain, L., Lacina, J., Bintz, W. P., Jimerson, J. B., Payne, K., & Zingale, R. (2020). Literacy in lockdown: Learning and teaching during COVID-19 school closures. *Reading Teacher*, *74*(3), 243-253.

Cope, B., & Kalantzis, M. (2009). "Multiliteracies": New literacies, new learning. *Pedagogies: An International Journal*, *4*(3), 164-195.

Dore, R., Justice, L., Mills, A. K., Narui, M., & Welch, K. (2021). Virtual kindergarten readiness programming for preschool-aged children: Feasibility, social validity, and preliminary impacts. *Early Education and Development*, *32*(6), 903-922.

Ewing, L. A., & Cooper, H. B. (2021). Technology-enabled remote learning during COVID-19: Perspectives of Australian teachers, students and parents. *Technology, Pedagogy and Education*, *30*(1), 41-57.

Firmanto, A., Sumarsono, P., & Nur, F. (2020). A family-school partnership-based learning: An effect to organize early childhood education during pandemic. *Advances in Social Science, Education and Humanities Research*, *477*, 100-103.

Fitri, M., & Latif, M. A. (2021). Adaptive learning for early childhood education during the COVID-19 pandemic in Aceh Jaya District: Online vs. offline. *Al-Athfal: Jurnal Pendidikan Anak*, *7*(1), 27-38.

Hu, Z., & Lu, W. (2020). Pre-kindergarten and kindergarten virtual school programs under COVID-19: A two-case comparative study. *Proceedings of the Association for Information Science and Technology*, *57*(e338), 2-5.

Jewitt, C. (2008). Multimodality and literacy in school classrooms. *Review of Research in Education*, *32*(1), 241-267.

Kipping, M., Wadhwani, D. R., & Bucheli, M. (2013). Analyzing and interpreting historical sources: A basic methodology. In M. Bucheli, & R. D. Wadhwani (Eds.), *Organizations in time: History, theory, methods,* (pp. 305-329). Oxford: Oxford University Press.

Lavoie, C., Sarkar, M., Mark, M. P., & Jenniss, B. (2012). Multiliteracies pedagogy in language teaching: An example from an Innu community in Quebec. *Canadian Journal of Native Education*, *35*(1), 194-210.

Mills, K. A. (2006). "We've been wastin' a whole million watchin' her doin' her shoes" situated practice within a pedagogy of multiliteracies." *The Australian Educational Researcher*, *33*(3), 13-32.

Ontario Ministry of Education. (2016). The kindergarten program. Retrieved from https://files.ontario.ca/books/edu_the_kindergarten_program_english_aoda_web_july21.pdf

Rajendram, S. (2015). Potentials of the multiliteracies pedagogy for teaching English language learners (ELLs): A review of the literature. *Critical Intersections in Education*, *3*, 1-18.

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity, 52*(4), 1893-1907.

Simons, H. (2014). Case study research: In-depth understanding in context. In P. Leavy (Ed.), *The Oxford handbook of qualitative research* (2nd ed., pp. 455-470). Oxford: Oxford University Press.

Taylor, S. V., & Leung, C. B. (2020). Multimodal literacy and social interaction: Young children's literacy learning. *Early Childhood Education Journal*, *48*(1), 1-10. https://doi.org/10.1007/s10643-019-00974-0

The New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*, *66*(1), 60-92.

Zhang, Z., Nagle, J., McKishnie, B., Lin, Z., & Li, W. (2019). Scientific strengths and reported effectiveness: A systematic review of multiliteracies studies. *Pedagogies*, *14*(1), 33-61.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Using Group Awareness Tools to Enhance Students' Behavioral and Cognitive Engagement with Peer Feedback in Online Collaborative Essay Writing

Jing REN[a], You SU[a*], Xiaohui SONG[a] & Haizhen GUO[a]
[a]*Beijing University of Posts and Telecommunications, China*
*suyou@bupt.edu.cn

Abstract: Group awareness tools have the potential to increase student engagement in computer supported collaborative learning. However, few studies have investigated the influence of group awareness tools on student engagement with peer feedback in online collaborative language learning settings. This study used a quasi-experiment method to explore the effects of group awareness tools on students' behavioral and cognitive engagement with peer feedback in online collaborative essay writing. The experimental group (n = 21) was required to complete 3-round collaborative writing and peer assessment of an argumentative essay supported by group awareness tools and the control group (n = 21) was asked to complete the same tasks without group awareness tools. The results showed that the students with group awareness tools demonstrated higher behavioral engagement with peer feedback. They provided significantly more comments across all the three rounds of peer feedback. Additionally, the students provided with group awareness tools showed significantly higher cognitive engagement in providing both surface-level and meaning-level feedback. The students in the experimental class engaged in significantly more meaning-level feedback in the last round of peer assessment. This paper also provides suggestions for future studies on applying group awareness tools in collaborative writing.

Keywords: Online collaborative writing, behavioral engagement, cognitive engagement, group awareness tools, peer feedback

1. Introduction

Collaborative writing has drawn much attention in second language learning in recent years (Calzada & García Mayo, 2021). Collaborative writing is defined as a task in which students work in small groups to collaboratively discuss, negotiate, and construct a shared written text (Li, 2018; Storch, 2013). Collaborative writing can develop writing skills (Awada, Burston & Ghannage, 2019), improve writing performance (Rahimi & Fathi, 2021) and increase writing efficiency (Chao & Lo, 2011). With the advent of web 2.0 tools, research in collaborative writing is now turning its interest towards online collaborative writing. Without the limitation of time and space, online collaborative writing provides more convenience and has the potential to make collaboration more effective. However, online collaborative writing may also result in low student engagement (Peng, Li, Su, Chen, & Jiang, 2022) and increase free rider (Michinov & Primois, 2005).

Group awareness is generally defined as some types of perception or understanding of the collaborating group or learning partners (Bodemer, Janssen, & Schnaubert, 2018). Previous research has indicated that group awareness tools can improve students' engagement (Liu, Liu, & Liu, 2018) and enhance social presence (Kimmerle & Cress, 2008) in online collaborative learning environment. Group awareness tools have been mainly adopted in domains of educational psychology, electronic commerce, and computer science (e.g., Li, Li, Zhang, & Li, 2021; Lin, Szu & Lai, 2016). However, the effects of group awareness tools have not been

adequately addressed in the field of language education, particularly in the context of peer evaluations in collaborative learning. Therefore, this study aims to explore whether group awareness tools will enhance students' behavioral and cognitive engagement with peer feedback in online collaborative essay writing. Specifically, we sought to address the following two questions:

(1) Did group awareness tools enhance students' behavioral engagement with peer feedback in online collaborative essay writing?

(2) Did group awareness tools enhance students' cognitive engagement with peer feedback in online collaborative essay writing?

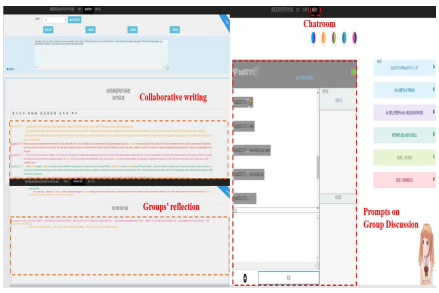
2. Methodology

2.1 Participants

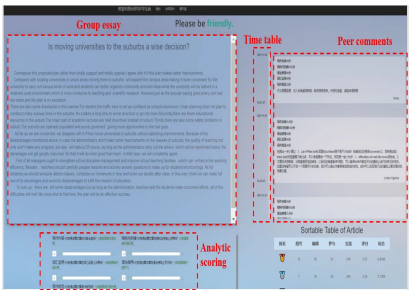
The participants were 42 first-year undergraduate students majoring in telecommunication at a university in northern China. They were enrolled in a 16-week College English course which was a compulsory degree course for the non-English majors at the university. Both classes were randomly divided into the experimental group and control group. Before the experiment, we checked the differences between the two groups' English proficiency using their English scores of entrance examination, and the results indicated that there was no significant difference in the two groups' English language proficiency ($p=0.23>0.05$, $t=1.23$).

2.2 System description

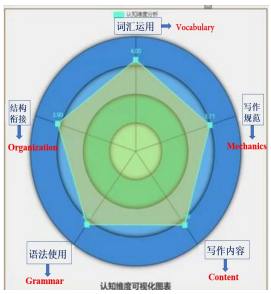
The system used in this study is *CollaWrite* which is a web-based writing platform that allows users to participate in group discussion, synchronous editing, and peer assessment of essays. Figure 1 showed the functions of *CollaWrite*. During collaborative writing, students can engage in real-time discussions in groups' chatroom and collaboratively edit, revise, and reflect on their writing. After writing, they can give comments and marks for another group's writing. This platform records the files of students' real-time discussion, essays, and peer feedback. *CollaWrite* can also provide cognitive and behavioral group awareness tools to support students' writing. Cognitive awareness tool displays students' writing performance on five dimensions of content, organization, vocabulary, grammar, and mechanics while behavioral awareness tool mainly shows group members' participation status like the number of discussion messages.



Screenshot of *CollaWrite*



Peer feedback



Cognitive group awareness tool



Behavioral group awareness tool

Figure 1. The screenshot of *CollaWrite* platform

2.3 Procedures

Students in both the experimental and control group were required to work in small groups to

jointly write an argumentative essay. The task took the process writing approach and it included three rounds of peer feedback activities. In each round of peer feedback, students rated other groups' writing from five dimensions: content, organization, vocabulary, grammar, and mechanics. More importantly, they provided qualitative feedback and suggestions on other's writing. Being different from the control group, students in the experimental group were provided with group awareness tools. The collaborative writing task was carried out on *CollaWrite* and the whole task lasted six weeks.

2.4 Measures

Behavioral engagement with peer feedback was measured by the number of comments and the word count of comments. The collaborative writing system collects data regarding the number of comments and the word count of comments.

Cognitive engagement with peer feedback was coded by the researchers using a code scheme adapted from Tian and Zhou (2020). As Table 1 shown, cognitive engagement is classified into surface-level feedback and meaning-related feedback. Surface-level feedback indicates few changes of meaning, whereas meaning-level feedback involves meaning changes. In addition, student engagement with feedback were further categorized into sub-codes including lexical, sentence, and paragraph levels, grammar feedback (i.e., tense, single plural forms and the third-person singular), and mechanic feedback (i.e., capitalization, punctuation marks and academic writing conventions).

The coding was completed by two coders. After training, both coders coded 20% of peer feedback and discussed discrepancies. Inter-rater reliability using Cohen's Kappa achieved 0.72, indicating a relatively high level of consistency. Then both coders finished the rest of the coding work.

Table 1. *Coding scheme of cognitive engagement with peer feedback*

Categories			Examples
Surface-level feedback	Meaning-preserving	Lexical	*"Changing 'lower' to 'decline' is better."*
		Sentence	*"There are many long sentences. You can combine long sentences and short sentences."*
		Paragraph	*"Good consistency between paragraphs."*
	Grammar		*"In the sentence 'people have they own idea', 'they' should be changed into 'their'."*
	Mechanics		*"There are more cases of capital confusion in the article."*
Meaning-level feedback	Meaning-related	Lexical	*"Aboriginals means original inhabitants, and it cannot be used to describe people living in suburbs."*
		Sentence	*"The last sentence seems to be a bit off topic."*
		Paragraph	*"Reference to our school is useless content in my opinion. Because it doesn't support the argument."*

3. Results

3.1 Comparison of behavioral engagement with peer feedback

Behavioral engagement with peer feedback involves the number of comments and the word count of comments. As shown in Figure 2, the number of comments in the experimental group is higher than those in control group in round 1, round 2, round 3, and the sum of the total three rounds as well. Figure 3 indicated that the experimental group was higher than control group on word count of comments in round 1, round 2, round 3, and the total of three rounds.

To further explore whether group awareness tools enhance students' behavioral engagement with peer feedback in online collaborative essay writing, independent samples t-test was used to analyze the differences in the number of comments and the word count of comments between experiment group and control group in the three rounds of peer assessment. As can be seen in Table 2, there was significant difference between experimental group and control group in the total number of comments. Additionally, significant difference also existed on the word count of comments between experimental group and control group in round 1, round 2, round 3, and the total of the three phases. The results showed that group awareness tools had a positive effect on students' behavioral engagement with peer feedback in online collaborative essay writing.

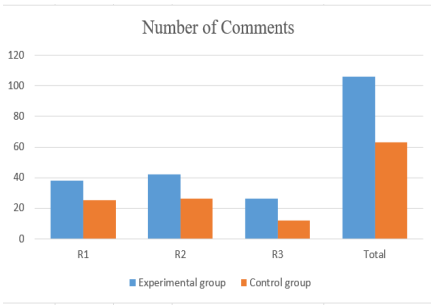


Figure 2. Number of Comments

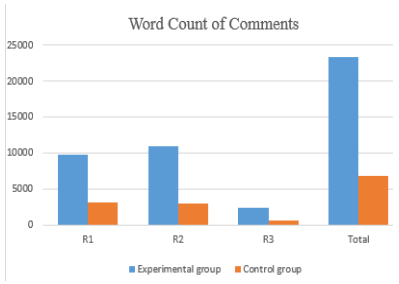


Figure 3. Word Count of Comments

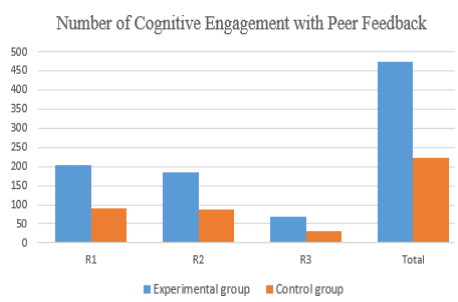


Figure 4. Number of Cognitive Engagement with Peer Feedback

Table 2. *Independent samples t-test for measures of behavioral engagement with peer feedback*

		Experimental Group		Control Group			
Round	Categories	Mean	SD	Mean	SD	t	*p*
Round 1	Number of comments	7.60	1.52	5.00	2.45	2.02	0.08
	Word count of comments	1970.60	672.64	627.00	505.05	3.57	0.01
Round 2	Number of comments	8.40	2.41	5.20	3.90	1.56	0.16
	Word count of comments	2205.20	907.56	606.20	368.62	3.65	0.01
Round 3	Number of comments	5.20	2.78	2.40	1.52	1.98	0.08
	Word count of comments	491.00	248.05	125.40	78.34	3.14	0.01
Total	Number of comments	21.20	4.09	12.60	3.51	3.57	0.01
	Word count of comments	4666.80	1121.80	1358.60	645.46	5.72	0.00

Table 3. *Independent samples t-test for measures of cognitive engagement with peer feedback*

Round	Categories	Experimental Group			Control Group				
		N	Mean	SD	N	Mean	SD	t	*p*
Round 1	Surface-level feedback	122	24.40	2.19	82	16.40	7.83	2.20	0.08
	Meaning-level feedback	55	11.00	3.00	37	7.40	2.30	2.13	0.07
Round 2	Surface-level feedback	113	22.60	7.23	73	14.60	8.79	1.57	0.16
	Meaning-level feedback	56	11.20	4.60	33	6.60	5.27	1.47	0.18
Round 3	Surface-level feedback	53	10.60	4.28	16	3.20	4.15	2.78	0.02
	Meaning-level feedback	23	4.60	2.51	8	1.60	1.14	2.43	0.04
Total	Surface-level feedback	288	57.60	7.96	187	37.40	9.71	3.60	0.01
	Meaning-level feedback	134	26.80	6.38	89	17.80	5.26	2.43	0.04

3.2 Comparison of cognitive engagement with peer feedback

Cognitive engagement with peer feedback includes surface-level feedback and meaning-level feedback. Figure 4 showed students' cognitive engagement with peer feedback in the experimental and control class. The results showed that the students with group awareness tools demonstrated higher cognitive engagement with peer feedback across all the three rounds of writing. To further explore whether group awareness tools enhance students' cognitive

engagement with peer feedback in online collaborative essay writing, independent samples t-test was conducted. As shown in Table 3, regarding the total of the three rounds of peer feedback, the results indicated that the students in the experiment group showed significant higher engagement with both types of feedback. Analysis for each round of peer review indicated that the difference in both level of feedback between the experimental group and control group was only significant in round 3.

4. Discussion and conclusion

This study conducted an empirical study to explore whether group awareness tools will enhance students' behavioral and cognitive engagement with peer feedback in online collaborative essay writing. In comparison with students in the control group, the experimental group had a greater number of comments and word count of comments. The results revealed that the students in the experimental class outperformed those in the control class regarding behavioral engagement with peer feedback. This result coincides with Pifarré, Cobos and Argelagós (2014) in that group awareness tools could contribute to promote students' behavioral involvement with peer feedback. The result also confirms Lin and Tsai (2016) in that collaborating students provided with group awareness tools engaged in more peer interactions and made more personal contributions compared to those without group awareness support.

In addition, the students provided with group awareness tools also showed significantly higher cognitive engagement in providing both surface-level and meaning-level feedback. The result indicated that group awareness tools contributed to the improvement of students' cognitive engagement with peer feedback. By using group awareness tools, the students in the experimental class may establish a better understanding of the group's knowledge and learning status, which may implicitly triggered socio-cognitive strategies that could be beneficial to the learning process (Sangin, Molinari, Nüssli, & Dillenbourg, 2011).

Though this study sheds light on the effect of group awareness tools on language learners' engagement in collaborative learning, it has two main limitations. First, this is only a pilot study and involves a relatively small sample size. Future studies need to invite more students to participate in the writing activity to increase the sample size. Second, in this study students only wrote one genre of writing (i.e., argumentative essay). In the future, we will further explore whether group awareness tools have different influence on writing of different genres in online collaborative learning settings.

Acknowledgements

This paper was supported by the Humanities and Social Sciences Fund of Chinese Ministry of Education (18YJC7 40084) and the Teaching Reform Project of Beijing University of Posts and Telecommunications (2021JXYJ19).

References

Awada, G., Burston, J., & Ghannage, R. (2019). Effect of student team achievement division through WebQuest on EFL students' argumentative writing skills and their instructors' perceptions. *Computer Assisted Language Learning, 33*(3), 275300.

Bodemer, D., Janssen, J., & Schnaubert, L. (2018). *Group awareness tools for computer-supported collaborative learning.* In F. Fischer, C. E. Hmelo-Silver, S. R. Goldman, & P. Reimann (Eds.), International handbook of the learning sciences, 351–358. Routledge.

Calzada, A., & García Mayo, M. D. P. (2021). Child learners' reflections about EFL grammar in a collaborative writing task: When form is not at odds with communication. *Language Awareness, 30* (1), 1–16.

Chao, Y., & Lo, H. (2011). Students' perceptions of Wiki-based collaborative writing for learners of English as a foreign language. *Interactive Learning Environments, 19*(4), 395–411.

Kimmerle, J., & Cress, U. (2008). Group awareness and self-presentation in computer-supported information exchange. *International Journal of Computer-Supported Collaborative Learning, 3*(1), 85–97.

Li, M. (2018). Computer-mediated collaborative writing in L2 contexts: an analysis of empirical research. *Computer Assisted Language Learning, 31*(8), 882–904

Li, Y., Li, X., Zhang, Y., & Li, X. (2021). The effects of a group awareness tool on knowledge construction in computer-supported collaborative learning. *British Journal of Educational Technology, 52*(3), 1178–1196.

Lin, J., Lai, Y., Lai, Y., & Chang, L. (2016). Fostering self-regulated learning in a blended environment using group awareness and peer assistance as external scaffolds. *Journal of Computer-Supported Collaborative Writing, 32*(1),77–93

Lin, J., Szu, Y., & Lai, C. (2016) Effects of Group Awareness and Self-Regulation Level on Online Learning Behaviors. *International Review of Research in Open and Distributed Learning, 17*(4), 224–242.

Lin, J., & Tsai, C. (2016). The impact of an online project-based learning environment with group awareness support on students with different self-regulation levels: An extended-period experiment. *Computers in Education, 99*, 28–38.

Liu, M., Liu, L., & Liu, L. (2018). Group awareness increases student engagement in online collaborative writing. *The Internet and Higher Education, 38*, 1–8.

Rahimi & Fathi (2021): Exploring the impact of wiki-mediated collaborative writing on EFL students' writing performance, writing self-regulation, and writing self-efficacy: a mixed methods study. *Computer Assisted Language Learning,* 1–48.

Michinov, N., & Primois, C. (2005). Improving productivity and creativity in online groups through social comparison process: new evidence for asynchronous electronic brainstorming. *Computers in Human Behavior, 21(1)*, 11–28.

Peng, Y., Li, Y., Su, Y., Chen, K., & Jiang, S. (2022). Effects of group awareness tools on students' engagement, performance, and perceptions in online collaborative writing: Intergroup information matters. *The Internet and Higher Education, 53*, 100845.

Pifarré, M., Cobos, R. and Argelagós, E., 2014. Incidence of group awareness information on students' collaborative learning processes. *Journal of Computer Assisted Learning, 30*(4), 300–317.

Sangin, M., Molinari, G., Nüssli, M. A., & Dillenbourg P. (2011). Facilitating peer knowledge modeling: Effects of a knowledge awareness tool on collaborative learning outcomes and processes. *Computers in Human Behavior, 27*(3), 1059–1067.

Storch, N. (2013). *Collaborative writing in L2 classrooms*. Bristol, UK: Multilingual Matters.

Tian, L., & Zhou, Y. (2020). Learner engagement with automated feedback, peer feedback and teacher feedback in an online EFL writing context. *System, 91*, 102247.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Research on activity design and application of AR gamification resources to promote children's second language learning

Min-Sheng FAN, Zhen-Hua LIU* & Xian-Ping JIN

School of Educational Technology, Northwest Normal University, China
xmanfms@126.com, 2020212889@nwnu.edu.cn, 364905079@qq.com
*Zhen-Hua LIU@corresponding.author

Abstract: The application of gamified learning resources based on AR technology in the field of early childhood education can support children's learning in a way that is more in line with the laws of children's cognitive development. Using AR learning resources to design children's learning activities and carry out teaching activities can effectively improve children's language level and learning ability. Compared with traditional teaching methods, the use of AR learning resources for children's activity teaching has a more significant effect on language learning.

Keywords: Augmented Reality, Activity Design, Early Childhood Education, Language Learning

1. Introduction

In January 2022, the Ministry of Education and other three departments issued the "Implementation Plan of the National Common Language Popularization and Improvement Project and the Promotion of Rural Revitalization Program". The "Implementation Plan" accurately analyzes the outstanding problems of unbalanced and insufficient popularization of the national common language and script, adheres to the unity of objectives, methods and effects, and implements the "three major actions". Action 1: Focus on provinces where the penetration rate of Putonghua has reached 85% and urban areas with good foundations, and carry out high-quality popularization of the national common language. Arrange the tasks of improving the education and teaching of the national common language and words, the standardization of language and words in the social field, the empowerment of language technology, and the service capacity of language and words. Augmented reality (AR) technology is based on virtual reality to virtualize physical things, and then perform physical simulation, that is, a technology that combines virtual images with the real world. AR technology has the characteristics of combination of virtual and real and real-time interaction. Users can increase the sense of presence and real experience through human-machine sensing equipment (Wang, 2021). China is a multi-ethnic country. For Tibetan children whose mother tongue is Tibetan, Chinese is their second language. This study takes Tibetan children as the research object to conduct a research on children's second language learning activities based on AR resources, in order to verify the learning resources and the effect of activities.

2. DESIGN OF CHILDREN'S LEARNING ACTIVITIES BASED ON AR LEARNING RESOURCES

2.1 Theoretical Research Basis

2.1.1 Flow theory

In 1957, Dr. Mihaly Csikszentimihalyi of the United States first proposed the flow theory, that is, users carry out activities independently and spontaneously in a certain environment. When people engage in an activity, they are highly focused, fully engaged in the situation, and completely filter out all irrelevant feelings (Li & Ma, 2022). AR technology builds this kind of "realistic" experience

environment for users, so that the user's immersive experience is similar or identical to the real physical environment. This kind of "realistic" experience allows users to build a "real" environment, which can more easily integrate into the virtual and real scene, and better complete the corresponding learning tasks.

2.1.2 Embodied cognition theory

Plato believed that mind and body are inseparable, and the study of embodied cognition involves perception, thinking, emotion, etc. (Chen, et al, 2014) Embodied cognition theory holds that the way and steps of the cognitive process are actually determined by the physical properties of the body; the content of cognition is provided by the body; cognition is embodied, and the body is embedded in the environment. (Ye,2010). The embodied environment includes the physical environment that provides the place of embodied perception, the resource support environment containing teaching resources, cognitive tools, and teacher-guided resources, and the social and cultural environment composed of interpersonal relationships, learning atmosphere, students' thinking patterns and behavior habits (Li, et al, 2018). Through embodied cognition, students can deeply construct cognition in immersive experiential learning, so as to improve the learning effect.

2.2 Gamification Resources Based on AR Technology

2.2.1 AR technology features and educational advantages

AR technology can realize direct interaction between users and the environment through computer generation of multi-mode virtual environment integrating sight, hearing, force and touch, that is, using technology to present the virtual world in the real environment and enhance people's perception of things (Weng, 2022). Therefore, AR technology has the characteristics of integration of real world and virtual information, real-time interaction, and positioning of virtual objects in three-dimensional scale space.

With the rapid development of new information technology, the integration of AR technology and education is getting closer and closer. Compared with traditional education, integrating AR technology into education can show the real things to learners in a three-dimensional and intuitive way. The Tibetan children in this study live in remote pastoral areas and have regional restrictions on the utilization of various social resources. Therefore, the use of AR technology can make abstract things more concrete and help improve children's cognition and understanding of social resources. In addition, AR technology integrated education has the advantages of strong interaction and participation, which is in line with children's hands-on and playful personality. AR gamification resources can allow children to participate in game interaction and improve children's hands-on ability and learning efficiency. In early childhood education activities, game activities are the most common way of learning. Using the unique advantages of AR technology, three-dimensional 3D motion pictures can be designed to attract children's interest. Realize that children learn through play, find fun in learning, and achieve effective teaching (Shen & Zhang, 2021). It is also the most effective way for children to subtly learn life skills.

2.2.2 Resource design

The "Learning and Development Guide for Children Aged 3-6" proposes that language is a tool for communication and thinking. Early childhood is an important period for language development, especially oral language development. At the same time, make educational suggestions: give young children more opportunities to listen and talk. For example: often talk to the young child about topics he is interested in, or read books and tell stories together. The design and development of AR resources follows the principle of language listening and conforms to the laws of cognitive development of children in Tibetan areas. Make local social entity cards, and embed voice prompts for nouns such as "cow", "goat", "bus", and "watermelon" for children to follow and learn.



Figure 1. AR Card Example.

On the one hand, AR cards can meet the needs of children's language and reading learning, and on the other hand, using AR devices to scan the cards will show the three-dimensional shape of things, so that children can more intuitively recognize the specific shapes and characteristics of things, and improve children's cognitive ability. At the same time, interesting human-computer interaction methods can attract children's attention and improve learning efficiency.

2.3 Learning Activity Design

The design of children's language learning activities based on AR resources should follow the principles of gamification and development.

Because resource design is playful, the basic form of activity design and organization should be game activities. Children's language learning is a process of actively accepting experience, and children can gain experience through practical operations (Chen, 2016). Therefore, the design of activities needs to follow the principles of children's own gamification practice operation, so that children can actively participate in activities, so that children can get the greatest activity experience in activities and achieve the purpose of activities.

The principle of development is that kindergarten education design activities are based on the comprehensive development of children (Wang, 2017). The design of children's activities needs to flexibly guide the development of children's language, society, art, health, science and other aspects. Development is a step-by-step process. Therefore, the design of activities for Tibetan children should follow the actual development needs of the local area, and use known knowledge to lead to unknown areas, making it easier for children to accept.

Based on the above principles, this study designed corresponding learning activities for young children.

3. IMPLEMENTATION OF LEARNING ACTIVITIES

3.1 Research objects, methods and implementation

This research is based on Gannan X Kindergarten, Gansu Province. There are 15 children in both the experimental class and the control version. The experimental class uses learning activities based on AR gamification resources for teaching, and the control version is taught according to the conventional teaching mode. Before the experiment, the experimental group and the control group were subjected to pre-test analysis, and after 6 weeks and 18 teaching activities, the post-test data were collected and analyzed.

This research uses the quasi-experimental method to test children before and after learning, and conducts interviews with children and teachers through the interview method to understand the real experience of teachers and children on the use of resources, and to better optimize AR learning resources.



Figure 2. Teaching Implementation Example.

According to the interviews with children and teachers in the early stage of the experiment, the main reason for the insufficient language learning of children in Tibetan areas is the lack of an environment that supports children's language learning in family life; The learning method based on recitation and memory leads to insufficient understanding of children's language; The group teaching-based teaching practice leads to the lack of communication opportunities and hands-on ability for young children.

3.2 Assessment and Data Collection

In the early stage of teaching implementation, the "Children's General Language Proficiency Test Questions" are used to pre-test children's language proficiency, and the pre-test data are collected. After the teaching activities are completed, post-processing is performed to collect the final data. And in the process of teaching implementation, real-time recording and observation of children's state is carried out.

Test example: Guide children to say the names and numbers of the following animals, and tell us which animals are most abundant and which are least abundant. (1 point for each correct answer to the name and number, 3 points for the relationship between the quantity.)

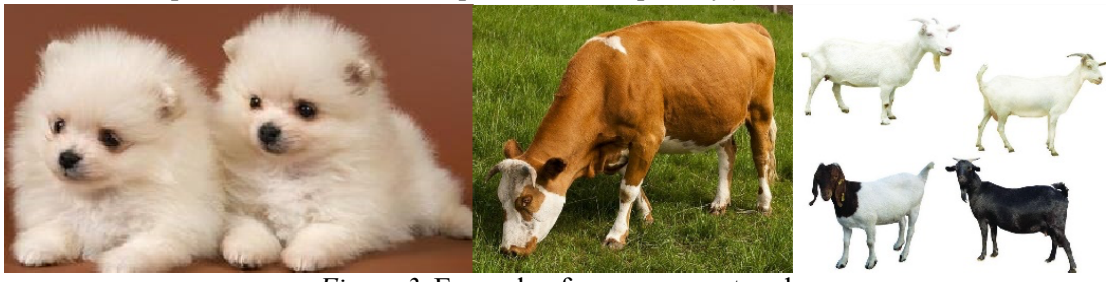


Figure 3. Example of an assessment scale

This test scale is aimed at three aspects: the test of children's oral language ability, the test of children's listening and comprehension ability and the test of children's ability to read and write. The test scale consists of 8 items with a total score of 100 points. Among them, the Children's Oral Ability Test contains 4 items with a total score of 48 points; the Children's Listening and Comprehension Ability Test contains 3 items with a total score of 40 points; the Children's Follow-up Reading Ability Test contains 1 item with a total score of 12.

3.3 Result analysis

By collecting and arranging the data of the test scale and analyzing the average value, from the overall level, the experimental group and the control group showed improvement from pre-experiment to

post-experiment. On average, there was almost no difference between the experimental group and the control group in the pre-test, but a great difference in the post-test.

Table 1. *Average Statistics Table*

	Pre-test	Post-test
Experimental-group	19.8667	61.26667
Control-group	17.1333	33.53333

3.3.1 Paired samples t-test analysis

According to Table 2, 15 students in the experimental group and 15 students in the control group were analyzed by paired sample T test.

Table 2. *Paired Sample T-Test in Experimental Group and Control Group*

	N	M	SD	df	t	Sig.
Experimental group pre-test	15	19.87	11.10	14	-7.758	0.000
Experimental group post-test	15	61.27	18.02			
Control group pre-test	15	17.13	12.11	14	-5.833	0.000
Control group post-test	15	33.53	20.20			

The pretest and posttest sig values of the experimental group and the control group is 0.000, which is less than 0.05, indicating that there is a significant difference between the pre-test and post-test values of the experimental group and the control group. Comparing with the mean results in Table 1, it can be found that the language level of students has been significantly improved after the implementation of teaching.

3.3.2 Independent sample t-test analysis

In order to verify the differences between the experimental group and the control group in the pre-test and post-test, the independent sample T test was used for analysis, as shown in Table 3 below.

Table 3. *Pre- Test and Post-Test Independent Samples T-Test*

		N	M	SD	df	t	Sig.
Pre test	Experimental group	15	19.87	11.10	28	-0.644	0.525
	Control group	15	17.13	12.11			
Post test	Experimental group	15	61.27	18.02	28	-4.236	0.000
	Control group	15	33.53	20.20			

It can be seen from the first group of data in Table 3 that through the homogeneity test, the significance of the pre-test data is 0.272, which is greater than 0.05, indicating that the variances are equal. Then according to the mean test, the sig value of the mean comparison between the two groups was 0.525, which was greater than 0.05, indicating that there was no significant difference in pretest data between the two groups. It can be concluded that there is little difference between the mean values of the experimental group and the control group at the beginning of the experiment, indicating that the ability levels of the subjects in the two groups are basically on the same level.

From the second group of post-test data, we can see that through the homogeneity test, the pre-test data is significant at 0.618, which is greater than 0.05, indicating that the variances are equal. Then according to the mean test, the sig value of the mean comparison between the two groups was 0.000, which was less than 0.05, indicating that the two groups of children had significant differences in the post-test data.

From this, it can be concluded that the mean difference between the experimental group and the control group is large after the end of the experiment, indicating that there has been a large difference in the level of children's language ability between the two groups, and the effect is obvious.

4. CONCLUSION

This study proves that AR-based children's learning activities can effectively improve children's language learning level. From the perspective of the resources themselves, AR-based children's learning resources are intuitive, interactive, interesting, and are more in line with the laws of children's cognitive development. Using multi-modal AR resources to design learning activities can stimulate children's multi-sensory, promote the development of spatial cognition, and improve children's attention. There is less practice and research in the field of AR technology empowerment education, and less application in the field of early childhood education, so there is still huge room for exploration in this area. On the contrary, the use of AR resources requires the use of electronic devices such as tablets, mobile phones, and all-in-one computers. Using these devices for a long time has a certain negative impact on children's bodies (especially the development of children's eyes). Therefore, the length of resource use should be strictly controlled.

AR resources can satisfy children's curiosity and stimulate children's interest in learning, so as to achieve the effect of promoting language development. Realize children's game learning. With the development of new technologies, it is highly feasible to carry out education and teaching in a way of combining virtual and real. A rich pragmatic environment plays an important role in children's language development. Through AR and other new technologies, we can create a richer pragmatic environment for children to learn a second language and promote their language development.

Acknowledgements

This research was supported by the 2021 Youth Doctoral Fund Project "Research on the Intelligent Learning Environment and Human Computer Cooperative Interaction Mode of Children's National Common Language in Tibetan Areas" No.2021QB020 and the National Natural Science Foundation of China under grant No. 6206700.

This research is a phased achievement of the "Innovation Star" project for outstanding graduate students in Gansu Province in 2022.

References

The Ministry of Education and other three departments issued the "Implementation Plan of the National Common Language Popularization and Promotion Project and Promotion to Help Rural Revitalization Program" [EB/OL]. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202201/t20220107_593034.html

Wang Anran. (2021). Reality analysis of the relationship between real and virtual in the context of augmented reality (AR) technology. *Communications of Dialectics of Nature*, 43(04): 40-46.

Li Hongchen, & Ma Jie. (2022). Research on the Reconstruction of "People, Fields, and Things" in the Metaverse Library from the Perspective of Immersion Theory. *Information Science*, 40(01):10-15.

Chen Yuming, Guo Tianyou, He Liguo, Yan Liangshi. (2014). A Review of Embodied Cognition Research. *Psychological Exploration*, 34(06):483-487.

Ye Haosheng. (2010). Embodied Cognition: A New Orientation of Cognitive Psychology. *Progress in Psychological Science*, 18(05):705-710.

Li Zhihe, Li Pengyuan, Zhou Nana, Liu Zhixiu. (2018). Design of Embodied Cognitive Learning Environment: Features, Elements, Applications and Development Trends. *Distance Education Journal*, 36(05):81-90.

Weng Ziyu. (2022). Research on the Application Effect of VR/AR Technology in Children's Picture Books——and Analysis of Technology Empowerment Path in Fusion Publishing. *Communication and Copyright*, (04):35-37.

Shen Hsinchu, & Zhang Guohong. (2021). The Significance and Value of AR Education in the Application of Early Childhood Life Education. *Foreign Economic and Trade*, (05):46-48.

Chen Yumei. (2016). Principles of Early Childhood Education Activity Design. *Xue Weekly*, (36):105-106.

Wang Xiaoxia. (2017). The Significance and Design Principles of Developing Early Childhood Education Activities. *Gansu Education*, (04):74.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Ask4Summary: A Summary Generation Moodle Plugin Using Natural Language Processing Techniques

Mohammed SALEH, Maria F. IRIARTE & Maiga CHANG*

Faculty of Science and Technology, Athabasca University, Canada
*maiga.chang@gmail.com

Abstract: Online learning management systems (e.g., Moodle) provide students with a variety of materials in a course. If learning materials include large amount of text, students may find it difficult to gain insights. Natural Language Processing (NLP) uses computational language models to obtain knowledge from text and speech. This study develops a Moodle plugin, Ask4Summary plugin, which summarizes relevant course content for the questions asked by students in a discussion forum. First, the system parses all course text-based content documents to extract and store the information. As students ask in a discussion forum, the system parses the questions and creates responses for them. The response is a customized summary generated via matching and retrieving from the stored information. Ask4Summary plugin is an open-source project and has been approved by Moodle. The tool is available for users to download from the Moodle plugins directory. For the next step, the research team has an evaluation plan to implement the plugin in a graduate level academic writing English course in collaboration with an Asian university. The questionnaire covers the perceived usability of the plugin and teacher´s and student's perceptions. It is expected that evaluation results demonstrate that the plugin generates an adequate summary for each question. The plugin might extend beyond the use in online learning (e.g., medical records, legal documents).

Keywords: Language Learning, N-gram, Part of Speech, Cosine Similarity, Discussion Forum

1. Introduction

Due to the pandemic, on-site education programs moved to the online environment to ensure education delivery (Dhawan, 2020). Nowadays, online learning resources in educational systems keeps growth (García-Morales et al., 2021; Setiawan et al., 2021). Nevertheless, information overload caused by technological development might influence learning performance. Feroz et al. (2021) conclude that an excess of information hinders students' knowledge acquisition.

This study addresses students' difficulties in handling extensive course materials across of different document types. The research team develops a Natural Language Processing (NLP) plugin for Moodle, a popular open-source learning management system. The main research contribution is the generation of a summary based on a student's question about a topic covered in the course materials.

The following sections are structured as follows: Section 2 elaborates on the related research background. Section 3 introduces the Ask4Summary Moodle plugin and explains the summary generation process. Section 4 discusses the document scanning workflow. Section 5 describes the technical issues of the summary generation. Section 6 proposes an evaluation plan for assessing the usability and satisfaction on the generated summary. Section 7 makes a conclusion and suggests future research.

2. Background

Natural Language Processing focuses on obtaining knowledge from the speech and text data with computational language models. Over time, the introduction of technology in language learning has resulted in favorable outcomes (Ahmadi, 2018). Research shows the advantage of NLP feedback systems for language learning over traditional technologies (Dodigovic, 2007; Nagata, 1993) and

suggest evolving from corrective feedback to interactive tutoring systems (Slavuj et al., 2015). Recent research confirms that automated personalized feedback provided by Intelligent Tutoring Systems (ITSs) promotes student knowledge acquisition (Al-Bastami & Naser, 2017; Kochmar et al., 2020). This study meets new trends in Technology Enhanced Language Learning (TELL) by generating a customized summary for each student's question.

N-gram is an NLP model based on a Markov chain that only accounts for the sequence of *n* components of a text (Russell & Norvig, 2022). To reduce the ambiguity inherent in natural language, the n-grams can be grammatically classified and labelled the words according to their part of speech (PoS) tags (Wilks & Stevenson, 1998). Ask4Summary plugin relies on two NLP open access web services: (1) *N-gram & PoS Identifier and Verifier* to select valid n-gram and PoS tags and (2) *Automatic Answering Service for Coronavirus Question*. The latter adopts both of NLP techniques and cosine similarity method (Pal, Chang, & Iriarte, 2022). Cosine similarity is a common metric that compares two vectors via the cosine of the angle they form, being 1 complete similarity and -1 total dissimilarity.

Moodle is a widespread open-source learning management system used in diverse educational options (Al-Ajlan & Zedan, 2008). Research highlights the system's comprehensive range of capabilities (Lisnani et al., 2020). Teachers can design a course and select the features that best fit their needs.

Moodle allows researchers to extend the standard version by creating plugins in various areas. New plugins require submission for reviewing and approval to become available to all Moodle sites. The research team designs Ask4Summary plugin as a block plugin. “Blocks” are a common plugin type that host functionalities in boxes distributed on the sides of Moodle webpages.

3. Ask4Summary Moodle Plugin

Ask4Summary plugin is a free tool approved by Moodle and available to users in the Moodle plugins directory (https://moodle.org/plugins/block_ask4summary). More documentation, workflow charts and tutorials can be found on the project website (https://ask4summary.vipresearch.ca/) and the research team YouTube channel (https://youtube.vipresearchgroup.ca). Tutorials cover installation, configuration, teacher use and student use including examples and results. Teachers can configure and add the block to their courses and have access to the service with a brief setup. Once the Ask4Summary Moodle plugin is installed and added to a course, teachers can choose from a variety of settings to control the summary generation service and the learning database.

The general process of Ask4Summary is to utilize the built-in Moodle tables to access the posts in a certain forum and use the course content to produce a response. Figure 1 shows the overall process flow. Once students have been told which forum to use, they can simply ask their questions with a “Helper Name” – a name that allows Ask4Summary to distinguish which post is a question. On a pre-set interval, a scheduled “Forum Scanning” task is queued. The task checks within the specific discussion forum in the Moodle database for the student’s posts that contains the “Helper Name”.

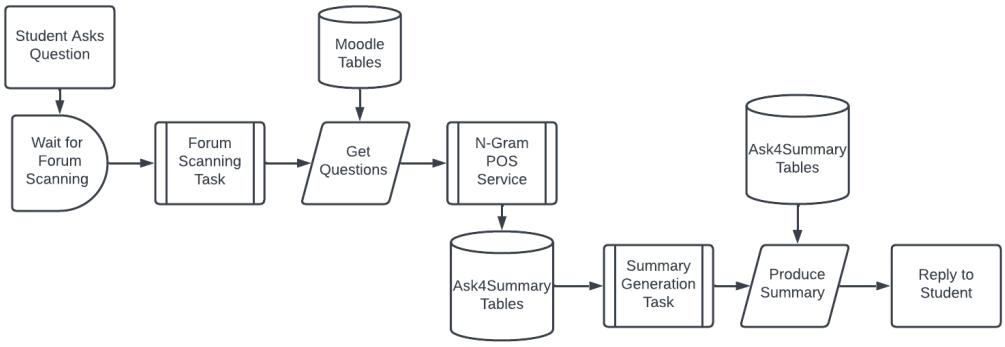


Figure 1. Ask4Summary Process.

With any post identified to be a student's question, another "N-Gram POS" service gathers and stores the n-grams and PoS tags from the post into Ask4Summary tables, by using the N-Gram & PoS Identifier and Verifier. From there, Ask4Summary has the "Summary Generation" ad-hoc task queued for producing a summary as the response of the student's post with the n-grams and PoS tags generated from course content. The ad-hoc task then uses the generated summary to respond the student's post in the designated forum.

There are settings that handle many of these processes and teachers can configure according to their needs. In the "Block Configuration" there are three general sections, for forum scanning, document scanning, and summary generation. Teachers can choose to enable/disable the plugin for responding questions, and how (i.e., the Helper Name) and where (i.e., a chosen discussion forum) students can ask their question as Figure 2 shows. They can also choose which file types they would like the plugin to consider as the base of its learning content database, and how many of these resources should be considered for responses as Figure 3 displays.

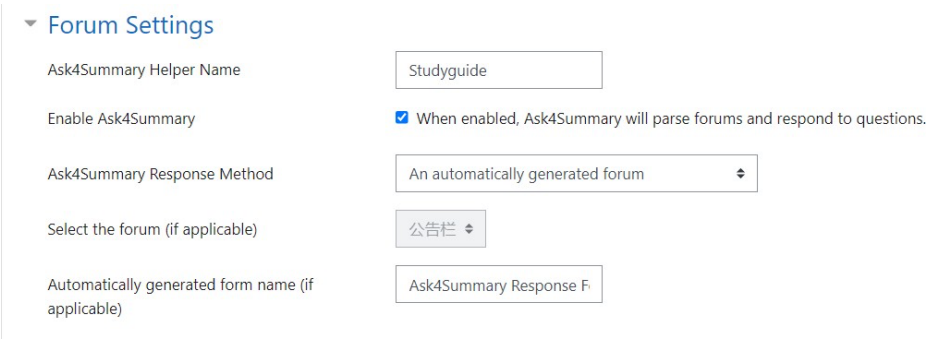


Figure 2. Forum Scanning.

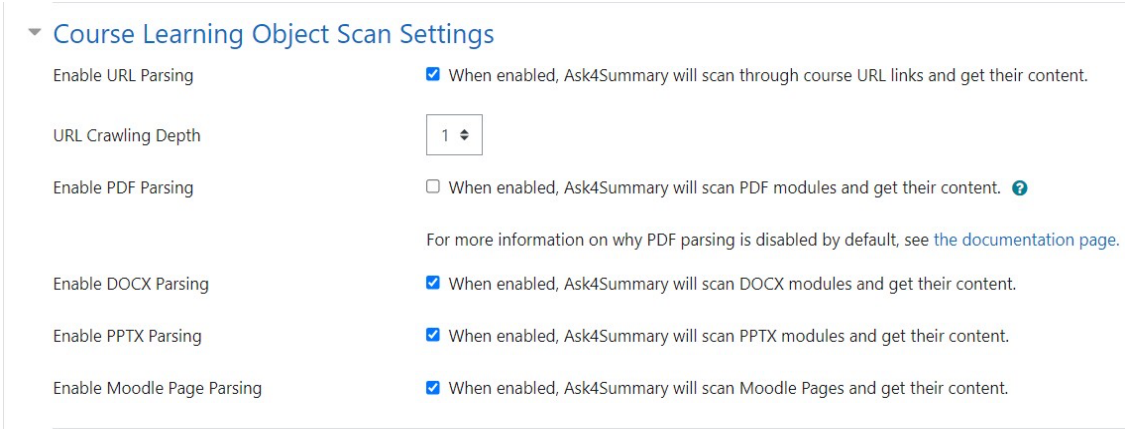


Figure 3. Document Scanning Settings.

Outside of the settings and in the block itself, there are two links that only the teachers can see: "Documentation" and "Logistics". The "Documentation" page has a variety of information on the plugin itself, and its range of functionality. The "Logistics" page shows a table for the scanning progress of each enabled file type, including the amount of unscanned and scanned course modules. It provides the names of each resource that were considered for the learning database and showcases all the questions that Ask4Summary was either able or unable to answer. The page also displays the average time that the plugin took to answer, and each question with who asked it and when.

4. Reading Content with Natural Language Processing

Before a summary can be generated, Ask4Summary needs a content base to choose sentences from. This content base is built from the course's material, where the plugin attempts to scan and parse course modules – activities and resources a course has – for their text content. Once the sentences in the

selected resources – teachers can choose which type(s) of materials should be scanned by the plugin as Figure 3 shows earlier – are retrieved, the plugin then uses the *N-Gram & PoS Identifier and Verifier* to find the valid n-grams based on their part of speech. Afterwards, this content is stored into Ask4Summary's learning tables. The document scanning and other system workflow scan be found in the documentation section of the project website (Vip Research Group, 2022).

During later hours of the day, Ask4Summary looks through the course materials that have not already considered to scan. Currently, the plugin supports the scanning of Word and PDF documents, PowerPoint presentation slides, URL web pages, and those built-in Moodle page modules. Depending on the filetype, Ask4Summary handles the parsing differently. For example, Word and PDF documents are already considered to be in sentence format, whereas PowerPoint slides are mostly sentence fragments. Regardless of type, Ask4Summary attempts to gather text content from these course modules.

Once the sentences are retrieved, Ask4Summary begins to parse the sentences for their n-grams and PoS tags. This is done by packaging a query that includes a sentence and the preferred settings for sending to the N-Gram & PoS Identifier and Verifier to get the top valid N-Grams. These n-grams and PoS tags and the corresponding sentence are stored in a variety of created XML based tables the plugin has. The tables form the content base.

5. Summary Generation

Forum scanning and summary generation work off one another to produce a response – summary generation in Ask4Summary can never commence if a question is not sent to the answering service. Once Ask4Summary begins its "Forum Scanning" task, it retrieves all the course's forum posts that contain the specified "Helper Name". Then, following the same process it reads and processes the course materials, the plugin goes through each forum post, breaks the question into sentences, and sends to the online service to get the valid n-grams and their part of speech tags.

Prior to starting the answering service, Ask4Summary compares the n-grams to its "response" table to see if similar question has the same identified n-gram set before was already answered. If so, it is not adding that post to the queue for summary generation but simply replies to the forum post with the previously generated summary. Once all the questions in the course were considered, if any questions were not already answered, Ask4Summary queues the background task "Question Answering".

The summary generation workflow is as follows. The "Summary Generation" task looks at all the forum posts that are considered to be valid. It gets the count of each n-gram the question has and creates a vector to represent the question. It also goes through each course learning object (e.g., a Moodle course page, a reading article in PDF, a teaching material in PowerPoint PPTX, an exercise document in DOCX, a supplemental webpage, etc.) and creates frequency vectors based on the same n-gram set for the learning object.

With the vectors, the task performs Cosine Similarity Calculation to find the top "X" documents, where the value of "X" can also be predefined and set by the teacher of the course. Now, these top documents are analyzed at the sentence level where all sentences from the documents are reviewed and the correspondent vectors for the same n-gram set are created. The Cosine Similarity Calculation is done again to compare how similar each sentence vector is to the original question vector. The top "Y" sentences will be considered, where "Y" value is also predefined and set by the teacher. With the identified "Y" sentences Ask4Summary can generate a summary to the student's question by replying his or her forum post. Figure 4 shows an example of a question asked by a student in the forum and Ask4Summary generates a summary to respond the student's question.

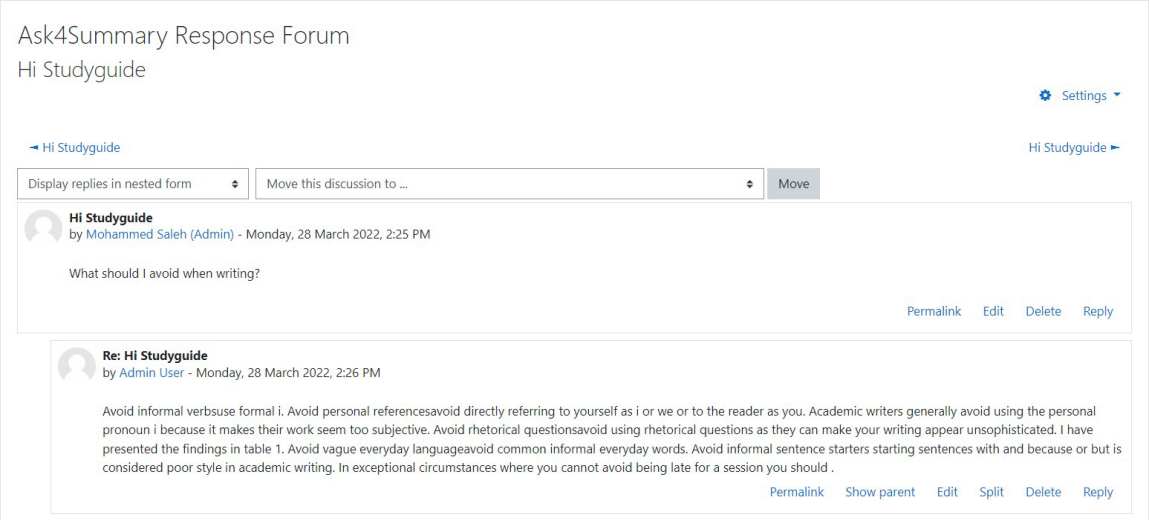


Figure 4. Student's question and the correspondent summary generated by Ask4Summary.

6. Evaluation Plan

To verify the usability and effectiveness of Ask4Summary, the research team plans to have the Ask4Summary plugin adopted and used in a graduate-level Academic English course through a partnership with an Asian university. The post-questionnaire focuses on several aspects of the system: usability, perceived relevance and satisfaction, and is preceded by an introduction and instructions.

The data analysis starts a review of the asked questions and the correspondent summary generated by Ask4Summary to check the teacher's and students´ perceived relevance and satisfaction degree. The research team is going to study the relevance and satisfaction degrees (in score 1 to 10) with the means of descriptive statistical analysis. The evaluation also foresees consultation with the teacher on discrepancies between perceptions on the same summary and opportunities for system improvement.

For the perceived relevance and satisfaction, the analysis will devote special attention to:

- Low relevance and low satisfaction scores: they require a detailed review on the question and the summary to identify possible reasons for the low score and system refinements.
- Relevance and satisfaction scores are not correlated: in general, the two scores on a question and its summary generated should not far apart. A relevant summary may generate satisfaction and vice versa. For this reason, inconsistent scores lead to investigation of the causes.

7. Conclusion

This study addresses the difficulty for students to extract information from the online learning materials in a course. The research team develops Ask4Summary, an NLP plugin for the open-source online learning management system Moodle. The plugin responds student questions posed in a forum by providing customized summaries from the text-based course materials (e.g., PDF, DOCX, PPTX, webpages, video transcripts).

There are further aspects of the study that have not yet been undertaken:

- The Ask4Summary plugin evaluation is planned on a Graduated Academic English course via a partnership with an Asian university.
- Influence of study on the improvement of students' knowledge acquisition and its possible contribution to learning success.
- Reduction of the summary generation time by applying other NLP techniques (e.g., distributed representations)

The research team has also observed that quality of a summary declines when generated from a presentation file (e.g., PowerPoint slides in PPTX format). This finding is understandable because the

presentation document might structure the content in tables and bullet points instead of sentences. Analysis of the collected data may provide more solid evidence to verify the observation.

This research is part of a larger project where other members in the research group are investigating and designing more sophisticated NLP methods. The research took the n-gram model as a starting point only and was aware of the limitation of the n-gram model and the chosen Cosine similarity method. The future work is aiming to improve the plug-in through the evaluation results and the investigation and trial outcome on the innovative and comprehensive NLP methods.

Acknowledgements

The authors acknowledge the support of the National Science and Engineering Research Council of Canada (NSERC) [RGPIN-2018-05545], NSERC Undergraduate Student Research Awards and Mitacs (Globalink program).

References

Ahmadi, D. M. R. (2018). The Use of Technology in English Language Learning: A Literature Review. *International Journal of Research in English Education, 3*(2), 115–125. https://doi.org/10.29252/ijree.3.2.115

Al-Ajlan, A., & Zedan, H. (2008). Why Moodle. *2008 12th IEEE International Workshop on Future Trends of Distributed Computing Systems*, 58–64. https://doi.org/10.1109/FTDCS.2008.22

Al-Bastami, B. G., & Naser, S. S. A. (2017). Design and Development of an Intelligent Tutoring System for C# Language. *European Academic Research, 4*(10).

Dhawan, S. (2020). Online Learning: A Panacea in the Time of COVID-19 Crisis. *Journal of Educational Technology Systems, 49*(1), 5–22. https://doi.org/10.1177/0047239520934018

Dodigovic, M. (2007). Artificial Intelligence and Second Language Learning: An Efficient Approach to Error Remediation. *Language Awareness, 16*(2), 99–113. https://doi.org/10.2167/la416.0

Feroz, H. M. B., Zulfiqar, S., Noor, S., & Huo, C. (2021). Examining multiple engagements and their impact on students' knowledge acquisition: The moderating role of information overload. *Journal of Applied Research in Higher Education, 14*(1), 366–393. https://doi.org/10.1108/JARHE-11-2020-0422

García-Morales, V. J., Garrido-Moreno, A., & Martín-Rojas, R. (2021). The Transformation of Higher Education After the COVID Disruption: Emerging Challenges in an Online Learning Scenario. *Frontiers in Psychology, 12*, 616059. https://doi.org/10.3389/fpsyg.2021.616059

Kochmar, E., Vu, D. D., Belfer, R., Gupta, V., Serban, I. V., & Pineau, J. (2020). Automated Personalized Feedback Improves Learning Gains in An Intelligent Tutoring System. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 140–146). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_26

Lisnani, L., Putri, R. I. I., Zulkardi, & Somakim. (2020). Designing Moodle features as e-learning for learning mathematics in COVID-19 pandemic. *Journal of Physics: Conference Series, 1657*(1), 012024. https://doi.org/10.1088/1742-6596/1657/1/012024

Nagata, N. (1993). Intelligent Computer Feedback for Second Language Instruction. *The Modern Language Journal, 77*(3), 330–339. https://doi.org/10.1111/j.1540-4781.1993.tb01980.x

Pal, S., Chang, M., & Iriarte, M. F. (2022). Summary Generation Using Natural Language Processing Techniques and Cosine Similarity. In A. Abraham, N. Gandhi, T. Hanne, T.-P. Hong, T. Nogueira Rios, & W. Ding (Eds.), *Intelligent Systems Design and Applications* (pp. 508–517). Springer International Publishing. https://doi.org/10.1007/978-3-030-96308-8_47

Russell, S. J., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

Setiawan, A. M., Munzil, & Fitriyah, I. J. (2021). Trend of learning management system (LMS) platforms for science education before-after Covid-19 pandemic. *AIP Conference Proceedings, 2330*(1), 060005. https://doi.org/10.1063/5.0043196

Slavuj, V., Kovačić, B., & Jugo, I. (2015). Intelligent tutoring systems for language learning. 2015 38th *International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 814–819. https://doi.org/10.1109/MIPRO.2015.7160383

Wilks, Y., & Stevenson, M. (1998). The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering, 4*(2), 135–143. https://doi.org/10.1017/S1351324998001946

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

What Criteria are Important for Evaluating the Quality of English Language Learning Edtech Products? Evidence From Literature

Minu NANDAKUMAR[a*], Leena BHATTACHARYA[a] & Sahana MURTHY[a]

[a]*Indian Institute of Technology, Bombay, India*
* nandakumar.minu@gmail.com

Abstract: The development and use of educational technology (EdTech) products have increased substantially in the last decade. Among other countries, India has witnessed a surge in the supply of EdTech products at the school education level spanning different disciplines. Primary among them has been English language learning EdTech products. In parallel, there has been an increase in demand for the same from stakeholders in India. However, in absence of a reliable framework to evaluate the quality of the available EdTech products in the education landscape, it has become difficult for the stakeholders to make meaningful choices on which product to use or adopt. In this regard, through a review of the literature, this paper identifies criteria to determine the quality of English language learning EdTech products that intend to benefit English as a Second Language (ESL) or English as a Foreign Language (EFL) learners in schools. The identified criteria can benefit decision makers to evaluate the quality of English language learning EdTech products in countries where English is taught as a second or a foreign language.

Keywords: language learning, educational technology, literature review, criteria, evaluation

1. Introduction

The development and use of educational technology (EdTech) products have risen substantially in the last decade, and more recently, in the Asian countries of India and China (Burns, 2021). The rapid surge of products has given a wide range of products to choose from but, in turn, it has uncovered a challenge. Stakeholders, including parents, teachers, and government officials, are finding it increasingly difficult to choose effective products in absence of a reliable framework that can inform them on how to determine the quality of EdTech products (Patel et al, 2021). Given the rising demand and supply of English EdTech products catering to the school education space, there is a need for a framework that takes into account the unique features of learning English as a second foreign language and is contextualised to enhance the learning of students in India.

Teaching and learning English as a second language in a multilingual classroom context is a complex process. Debates around 'Reading Wars' (Castles et al, 2018) concerning language development have been going on for decades. Standards and guidelines to set the quality of English language teaching and learning in a classroom exist in the ecosystem (Common Core State Standards, 2010; EdReports, 2022) but they are based on developed-country contexts where English is taught as the first language. Hence, they cannot be adopted to evaluate English language learning EdTech products that are designed to benefit English as a Second Language (ESL) or English as a Foreign Language (EFL) learners in school. The criteria highlighted by the Government of India (Department of School Education and Literacy, n.d) to evaluate e-Content fall short of incorporating English language learning specific criteria. This highlights the need to identify and design context-specific criteria to evaluate English language learning EdTech products, catering to ESL or EFL learners.

This paper identifies a set of concrete, reliable, evidence-informed features that can allow stakeholders to evaluate the quality and pedagogical suitability for curricular integration of EFL or ESL EdTech products. We define what these features mean and specify what to look for in EdTech products.

The findings of this paper contribute to the field of learning English using technology in school education.

2. Methodology

Literature review was done to identify the relevant criteria to determine the quality of English language learning EdTech products and the features of implementing them. A keyword search was done for each criterion in content, pedagogy and technology. Google Scholar, Science Direct, and Education Resources Information Center (ERIC) were used as the major source for finding papers. Papers were also selected through snowballing technique from the reference sections of the selected papers. Reviewed journals are at the intersection of three disciplines: Technology (Example, Computers & Education), Language learning (Example, Computer Assisted Language Learning) and Psychology (Example, Journal of Educational Psychology). The inclusion criteria of the review papers were as follows. First, all the studies were published in peer-reviewed journals between 2012-22. Seminal papers with theories related to language learning, and policy reports move beyond this range. Second, only the studies that explicitly focused on aspects of language learning were considered. Third, journal articles were based on empirical research covering students of the school-going age. Studies could be from multi-country contexts but cater to EFL or ESL learners.

3. Important criteria to determine quality: Evidence from the Literature

The section reviews literature and outlines criteria across content quality, pedagogical alignment and technology and design, to evaluate quality of English language learning EdTech products.

3.1 Content Quality

Besides ensuring content accuracy, the product design should also consider the comprehensibility of the content by the intended learners. It depends on the accent, vocabulary, and sentence structure used. Machine voice or foreign-accented voice may add to the cognitive load of the learners (Mayer, 2014). The use of figurative language, such as metaphors and analogies as well as strong conceptual material and large academic vocabulary, make texts difficult to understand (Common Core State Standards, 2010). Next, representation of race, gender and other forms of identity can affect learning positively (Dore, 2022), but there should not be a reinforcement of stereotypes associated with these identities. Texts of adequate complexity (measured by layout, layers of meaning and ideas of the visual text, font size, spacing, and illustration) for particular grades should be included (Common Core State Standards, 2010). For instance, a text with small fonts, dense language, multiple layers of meaning, time shifts, multiple characters, and storylines is typically difficult to follow, especially for students in lower grades. Exposure to multiple forms of content such as digital storytelling (Quah and Ng, 2021), poems, and drama (Zhang et al, 2018) can also improve language skills. Technology-mediated balanced approaches to teaching important components of literacy can improve reading achievement (Lysenko et al, 2014). All essential literacy components, including phonological awareness, decoding, oral language development, vocabulary, comprehension, and writing are covered in it. Products should cater to both lower-order (decoding, sound-symbol code) and higher-order (comprehension, metacognition) skills including guided reading and repeated reading, as well as writing skills. Good products should consider explicit skill instruction and practice to reach all learners (Castles et al, 2018). The criteria and what to look for in products are given in Table 1.

Table 1. *Criteria under Content Quality to check the quality of English EdTech product*

Criterion name and meaning	**What to look for in EdTech products**
Content Accuracy: Correctness of all forms of content in the product	Presence of proper facts, explanations, spellings, grammatical usage, pronunciation and punctuations
Comprehensibility: The language can be understood by intended learners	The accent, vocabulary is easy to understand by intended learners

Inclusivity: Include representation of the diverse learner groups	Unbiased representation of multiple sections of the society in the content that learners are familiar with
Content complexity: The text is at the level of the child	The layout, layers of meaning and ideas discussed are age appropriate
Types of texts: Texts of diverse nature	Availability of varied texts to engage with
Skill coverage: Covers language-specific skills	Explicit instruction and coverage of English language skills aligned to the national curriculum and policy

3.2 Pedagogical Alignment

The capacity to keep learners engaged with the content is important while evaluating the quality of an EdTech product. Emphasis on a conversational style in first-or-second person rather than a third-person impacts friendliness, retention and cognitive processing (Ginns et al, 2013). Cues should direct learners' attention to the relevant parts of the content or highlight the organisation of the material (Van, 2014). The presence of content in meaningful, coherent, and learner-paced segments rather than continuous units contributes to overall learning (Mayer and Pilegard 2014). Short and engaging content should be placed in a context that is aligned with their lives, needs and interests to have a strong influence on their learning (Zhang et al, 2018). Authentic context, in which language can be used in real life, helps learners master the content (Hwang et al, 2014). Context-aware technologies can also be used (Shadiev et al, 2017). It is important to check if the product helps the learners to construct the meaning of the content rather than merely transmitting information. The presence of guiding questions that help learners make connections, construct and test mental models can turn watching videos into an active learning experience. Appropriate scaffolds can help (Sysoev et al, 2022). Learning objectives and assessment questions should be aligned (Biggs, 1996) and the rigour of questions should meet the level of the learners (Hsu, 2016). Ensuring collaboration with other learners improves language skills and the aspect of anonymity can support a learner towards this (Jiang and Zhang, 2020). There is consensus in the literature that receiving feedback following the completion of a task is an effective technique to enhance learning (Wilson and Czik, 2016). Feedback should be easy for learners to interpret and should state ways in which learners can rectify their errors. However, simply providing feedback is not enough to alter a learner's post-feedback behaviour. Detailed feedback with motivational messages can help (De Sixte et al, 2020). Adult support features, for teachers and parents, to help learners use and effectively integrate the product into learning are essential. The criteria and what to look for in products are represented in Table 2.

Table 2. *Criteria under Pedagogical Alignment to check the quality of English EdTech product*

Criterion name and meaning	**What to look for in EdTech products**
Constructivism: Allowing construction of meaning of content on their own	Examples, guiding questions to help learners reflect on learning, provision of scaffolds when required
Rigour: Activities, exercises at the level of rigour appropriate for learners	The activities challenge the learners to an appropriate level, without demotivating them
Alignment: Assessment questions matched to the learning objective	The assessment questions check for the skill mentioned in the learning objective
Engagement: Content presentation style and graphics to engage learners	Providing cues, illustrations, highlighting important points, using an active voice for discussion
Feedback: Responses to learners' performance in assessments or activities	Verbal or visual feedback that guides learners to identify what went wrong, and how it can be rectified
Motivation: Motivate learners to interact further with the content	Motivation can be seen through stars, badges, and visual or verbal warm messages
Context: Relevance of the context used to the learner	The context is relatable to the learners in terms of setting, choice of vocabulary, and examples
Collaboration: Opportunities for the learners for social interaction	Possibilities to enhance anonymous conversation, peer feedback, writing and others
Support: Support provided to the adults to help learners use the product	Support to adults to guide learners to meaningfully engage with the content

3.3 Technology and Design

Aside from ensuring an intuitive and easy-to-use interface, effectively used technology can reduce hurdles faced in language learning by children with varying abilities (Berninger et al, 2015). For example, a full-caption video or text would not only benefit learners with hearing disorders but can help all learners to improve their language skills (Teng, 2019). The ease with which the learners can navigate the content offered in the products, and the pace at which they can engage with the content also affect their sense of autonomy, and their learning experience (Reiber-Kuijpers et al, 2021). Tools like Learning analytics dashboards (LADs) help students (and teachers) make sense of their learning by offering insights into their learning patterns. It can help with self-regulation in team activities as well as assist educators to identify cognitive constraints and provide actionable pedagogical support (Zamecnik et al, 2022). The criteria and what to look for in products are represented in Table 3.

Table 3. *Criteria under Technology and Design to check the quality of English EdTech product*

Criterion name and meaning	What to look for in EdTech products
Interface design: Product features that would make the interface meaningful	Intuitive and easy-to-use interface design where the operations can be performed without hardship
Universal design: The features are designed in a way to reach all learners	Presence of full captions in videos, voice-overs for texts to reach all learners meaningfully
Analytics: Features to capture performance of learners	Easy and interpretable format to display learner's performance and areas of improvement
Learner navigation and pace: Features to support self-paced learning	Technology supports learners to explore content at their own pace and select relevant content

4. Designing for different use cases: Additional criteria for consideration

In this section, the aforementioned criteria are extended to PAL and DGBLL English products.

4.1 Personalised Adaptive Learning (PAL) products for English language

In an effective language learning curriculum, learners should receive content that meets their interests and learning needs. Adaptive learning systems that recommend articles based on learners' preferences and reading proficiency levels, significantly contribute to better reading comprehension. Personalised recommendation mechanisms help motivate learners and thereby improve their reading effectiveness (Hsu et al, 2013). Personalisation can be offered in multiple ways. For instance, there is a strong effect of providing personalised and motivational messages after formative tests (De Sixte et al, 2020). Personalised feedback that compares the learner's current performance to their past performance can enhance learning (Maier, 2021). Personalisation can also be done by providing content in a context relevant to the learners. Recent innovation also allows devices to provide contextualised content by tracking the coordinates of the learners (Hwang et al, 2014).

4.2 Digital Game-Based Language Learning (DGBLL) products

Studies have observed that the 'badge mechanism' in DGBLL, which gives learners instant feedback raises their self-efficacy and enhances their learning of English as a foreign language (Yang et al, 2016). The feeling of being a part of the game, created through sensory inputs such as auditory and visual feedback at the end of each question, can enhance engagement (Sandberg et al, 2014). This can be done by giving learners immediate feedback (both visual and auditory) on their performance. Quality DGBLL products should challenge learners while maintaining a balance between the learner's level and the challenge posed to them (Hsu, 2016; Sandberg et al, 2014). The game's rules should be easy to understand and should not have too many. A learner might get overwhelmed by difficult or too many rules (Acquah and Katz, 2020). Effective integration of competition and collaboration, synchronous and asynchronous, in the product design can be meaningful for learning (Hsu, 2016). Setting up goals, planning and Self-Directed Learning (SDL) is beneficial for learning in game-based settings (Li et al,

2021). Maintaining learners' anonymity and helping them in using *Avatars* and pseudonyms to express themselves is a strategy to improve engagement between learners (Chen and Kent, 2020). It empowers learners to interact with others and have open interactions without fearing mistakes. Additionally, a well-designed game allows learners to interact with it, and have adequate control over the gameplay, game environment and learning experience (Xu et al, 2020).

5. Discussion and Conclusion

The article outlines the evaluation criteria for English EdTech products for ESL or EFL learners. We identify the broad criteria to determine the quality of content, associated pedagogy and technology features but we do not claim that the criteria are exhaustive. Our choice of criteria aims to ensure that the quality of products can be measured but the number of criteria does not deter evaluators. Some of the criteria, such as content accuracy, constructivist approach, and universal design, can be used across country contexts. Others, including language comprehensibility, context, skill coverage need to be contextualised based on the context of the intended learner. We carry out this exercise for the Indian context, as a part of the EdTech *Tulna* project (EdTech Tulna, 2022). Based on these criteria, we developed the *Tulna-English* index, a three-scaled rubric with detailed reviewer guidelines to determine the quality of English language learning EdTech products. It has been contextualised to cater to the needs of learners in India. The skill-coverage criteria base itself on the skills defined by the National Council of Educational Research and Training (NCERT) for each grade and the overall index has been designed keeping National Education Policy 2020 into account. Similarly, language comprehensibility and context have been customised to include examples that are relevant to an Indian learner. The index will be validated through the evaluation of English language learning EdTech products. The contribution of this paper is to help stakeholders identify the features to evaluate the quality of English EdTech products. Similar exercises can be done by other countries for informed decision-making to effectively integrate technology into education.

Acknowledgement

The authors acknowledge the support of Central Square Foundation for funding and partnering in the *EdTech Tulna* project. The authors are grateful to the *Tulna* team members for their comments and suggestions, and to the Educational Technology interdisciplinary program at IIT Bombay for overall support.

References

Acquah, E. O., & Katz, H. T. (2020). Digital game-based L2 learning outcomes for primary through high-school students: A systematic literature review. *Computers & Education*, 143, 103667.

Berninger, V. W., Nagy, W., Tanimoto, S., Thompson, R., & Abbott, R. D. (2015). Computer instruction in handwriting, spelling, and composing for students with specific learning disabilities in grades 4–9. *Computers & Education*, 81, 154-168.

Biggs, J. (1996) Enhancing teaching through constructive alignment. *Higher Education* 32, no. 3: 347.

Burns, M (2021). Technology in Education. Paper commissioned for the 2023 Global Education Monitoring Report, Technology and education

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. Psychological Science in the Public Interest, 19(1), 5–51.

Chen, J. C., & Kent, S. (2020). Task engagement, learner motivation and avatar identities of struggling English language learners in the 3D virtual world. System, 88, 102168.

Common Core State Standards (2010). National Governors Association Center for Best Practices & Council of Chief State School Officers. Washington, DC

De Sixte, R., Mañá, A., Ávila, V., & Sánchez, E. (2020). Warm elaborated feedback. Exploring its benefits on post-feedback behaviour. *Educational Psychology*, 40(9), 1094-1112.

Department of School Education & Literacy. Retrieved 28 May 2022, from https://ncert.nic.in/pdf/GuidelinesforeContent3.pdf

Dore, R. A. (2022). The effect of character similarity on children's learning from fictional stories: The roles of race and gender. *Journal of Experimental Child Psychology*, 214, 105310.

EdReports. (2022). Retrieved 28 May 2022, from https://www.edreports.org/

EdTech Tulna (2022). https://edtechtulna.org/, retrieved 27 May 2022.
Ginns, P., Martin, A. J., & Marsh, H. W. (2013). Designing instructional text in a conversational style: A meta-analysis. *Educational Psychology Review*, 25(4), 445–472.
Hsu, C. K., Hwang, G. J., & Chang, C. K. (2013). A personalized recommendation-based mobile learning approach to improving the reading performance of EFL students. Computers & Education, 63, 327–336
Hsu, T. (2016). Learning English with augmented reality: Do learning styles matter? *Computers & Education*, 106, 137–149.
Hwang, W. Y., & Chen, H. S., Shadiev, R., Huang, R. Y. M., & Chen, C. Y. (2014). Improving English as a foreign language writing in elementary schools using mobile devices in familiar situational contexts. *Computer Assisted Language Learning*.
Jiang, D., & Zhang, L. J. (2020). Collaborating with 'familiar' strangers in mobile-assisted environments: The effect of socializing activities on learning EFL writing. *Computers & Education*, 150, 103841.
Li, H., Majumdar, R., Chen, M. R. A., & Ogata, H. (2021). Goal-oriented active learning (GOAL) system to promote reading engagement, self-directed learning behavior, and motivation in extensive reading. *Computers & Education*, 171, 104239.
Lysenko, L. V., & Abrami, P. C. (2014). Promoting reading comprehension with the use of technology. *Computers & Education*, 75, 162–172.
Maier, U. (2021). Self-referenced vs. reward-based feedback messages in online courses with formative mastery assessments: A randomized controlled trial in secondary classrooms. *Computers & Educatio*n, 174, 104306.
Mayer, R. (2014). Principles Based on Social Cues in Multimedia Learning: Personalization, Voice, Image, and Embodiment Principles. In R. Mayer (Ed.), The Cambridge Handbook of Multimedia Learning (Cambridge Handbooks in Psychology, pp. 345-368). Cambridge: Cambridge University Press.
Mayer, R. E., & Pilegard, C. (2014). Principles for managing essential processing in multimedia learning: segmenting, pre-training, and modality principles. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning (2nd ed., pp. 316–344). Cambridge: Cambridge University Press.
Patel, A., Dasgupta, C., Murthy, S., & Dhanani, R. (2021). Co-Designing for a Healthy Edtech Ecosystem: Lessons from the Tulna Research-Practice Partnership in India. Rodrigo, MMT et al.(Eds.).
Quah, C. Y., & Ng, K. H. (2021). A Systematic Literature Review on Digital Storytelling Authoring Tool in Education: January 2010 to January 2020. *International Journal of Human–Computer Interaction*, 1-17.
Reiber-Kuijpers, M., Kral, M., & Meijer, P. (2021). Digital reading in a second or foreign language: A systematic literature review. *Computers & Education*, 163, 104115.
Sandberg, J., Maris, M., & Hoogendoorn, P. (2014). The added value of a gaming context and intelligent adaptation for a mobile learning application for vocabulary learning. *Computers & Education*, 76, 119–130.
Shadiev, R., Hwang, W. Y., & Huang, Y. M. (2017). Review of research on mobile language learning in authentic environments. *Computer Assisted Language Learning,* 30:3-4, 284-303
Sysoev, I., Gray, J. H., Fine, S., Makini, S. P., & Roy, D. (2022). Child-driven, machine-guided: Automatic scaffolding of constructionist-inspired early literacy play. *Computers & Education*, 182, 104434.
Teng, F. (2019). Maximizing the potential of captions for primary school ESL students' comprehension of English-language videos. *Computer Assisted Language Learning*, 32(7), 665-691.
Van Gog, T. (2014). The signalling (or cueing) principle in multimedia learning. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning (2nd ed., pp. 263e278). New York, NY: Cambridge University Press.
Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94-109.
Xu, Z., Chen, Z., Eutsler, L., Geng, Z., & Kogut, A. (2020). A scoping review of digital game-based technology on English language learning. *Educational Technology Research and Development*, 68(3), 877-904.
Yang, J. C., Quadir, B., & Chen, N. S. (2016). Effects of the badge mechanism on self-efficacy and learning performance in a game-based English learning environment. *Journal of Educational Computing Research*, 54(3), 371-394.
Zamecnik, A., Kovanović, V., Grossmann, G., Joksimović, S., Jolliffe, G., Gibson, D., & Pardo, A. (2022). Team interactions with learning analytics dashboards. *Computers & Education*, *185*
Zhang, H., Hwang, W. Y., Tseng, S. Y., & Chen, H. S. (2018). Collaborative drama-based EFL learning in familiar contexts. *Journal of Educational Computing Research*, 57(3), 697-722.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Explainable English Material Recommendation Using an Information Retrieval Technique for EFL Learning

Kensuke TAKII[a*], Brendan FLANAGAN[b], Huiyong LI[b], Yuanyuan YANG[a] & Hiroaki OGATA[b]

[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*kensuke.takii96@gmail.com

Abstract: Learning material recommendation has been a common field in the recommendation in e-learning due to the difficulty the learners experience in choosing appropriate learning materials among many resources. However, few traditional recommendation methods can be applied to e-learning as they are because they do not consider the learners' characteristics. Such methods may not be persuasive enough for the learners and make them less motivated. In this study, we propose an explainable English material recommendation that can adapt to the changes of learners' learning state and can explain the basis of the recommendation by using an information retrieval technique. This aims to address learners' trust and motivation issues. The algorithm estimates the difficulty of materials and learners' English skills and makes material recommendations that fit their skill levels. A case study in the setting of extensive reading is also described. Lastly, this paper introduces plans for implementation using an e-learning system with this recommendation. In the future, we will conduct an experiment and improve the recommendation algorithms.

Keywords: e-learning, learning material recommendation, English as a foreign language, system transparency

1. Introduction

These days, as learning resources available online are exponentially increasing, learners are having difficulty in choosing appropriate learning materials due to information overload. This is a partial reason why learning material recommendation is one of the most common fields in the recommendation in e-learning (Tarus, Niu & Mustafa, 2018). Although various traditional recommendation methods have been proposed in the field of information retrieval in the past (e.g., collaborative filtering, and content-based filtering), they cannot be applied to e-learning material recommendation as they are because they do not take into account the chronological change of learners' cognitive states and learning contents (George & Lal, 2019). In addition, black-box recommendation, which has no transparency in its mechanism, may not be trusted by the users, especially when they cannot agree with the recommendations (Herlocker, Konstan & Riedl, 2000; Abdi, Khosravi, Sadiq & Gasevic, 2020). According to previous work, intelligent tutoring systems with prompt and feedback mechanisms can improve students' motivation in self-regulated learning and lead them to higher achievement (Duffy & Azevedo, 2015). Thus, if learning material recommendation is not persuasive enough and not trusted by the learners, it may make them less motivated and not be effective in their learning. To address these problems, we argue that the recommendation mechanisms in education should entail the explanation of the rationale behind the recommendation to foster learners' trust and motivation for learning.

In this study, targeting English as a Foreign Language (EFL) learning, we propose an explainable English learning material recommendation that takes into consideration learners' characteristics changing over time and makes explanations of its recommendation for the purpose of being persuasive for learners. This recommendation involves an e-book reader system and a vocabulary profile. It processes learners' reading logs and the difficulty of vocabulary with an information retrieval technique to estimate learners' English skills and the difficulty of materials. Using the estimated skills

and difficulties, the mechanism makes material recommendations so that the materials which have the closest difficulty to the learners' skills will be recommended. This mechanism can adapt to the chronological change of the learners' English skills by processing learners' learning logs with the materials in real-time. Besides, this recommendation can generate explanations about its basis according to the proximity between the difficulty of recommended materials and the learners' skills. This can contribute to improving learners' trust and motivation.

In addition, this paper proposes an implementation plan for evaluating the effect of the recommendation on learners' learning and motivation. This recommendation will be implemented in an existing e-learning platform including an e-book reader. Lastly, future work for the improvement of the recommendation will be described.

2. Related Work

2.1 Recommender System in E-learning

Many previous studies have stressed the importance of personalized systems in e-learning. Shishehchi, Banihashem, Zin & Noah (2011) stated that personalized environments in e-learning are very important since learning is a cognitive activity that differs from learner to learner. Many e-learning systems developed in the past provide a personalized learning experience based on the individual learner's needs, prior knowledge, preferences, and/or learning styles, and are more effective than non-personalized systems (Zou & Xie, 2018).

To provide personalized recommendation in e-learning systems, learners' knowledge and learning materials knowledge are necessary (Shishehchi, Banihashem, Zin & Noah, 2011). Most of these systems use learners' preferences (Hsu, 2008; Bourkoukou & Bachari, 2018), learning styles (Truong, 2016; Klašnja-Milićević, Vesin, Ivanović & Budimac, 2011), and knowledge levels (Bobadilla, Serradilla & Hernando, 2009). As a technique for generating recommendations, data mining is used these days to learn about students' behaviors in educational systems (Aher & Lobo, 2013). This is because a huge amount of educational data has been available. For example, Hsu (2008)'s English learning recommender system for English as a Second Language (ESL) students recommends reading lessons that suit their interests by using content-based analysis, collaborative filtering, and data mining techniques.

2.2 Generation of Explanations for Recommendations in E-learning

Explanations of the rationale behind the recommendations can be considered important for learners. Abdi, Khosravi, Sadiq & Gasevic (2020) mentioned the potential of a transparent educational recommender system, and these days this has been proven positively. Ooge, Kato & Verbert (2022)'s investigation on explaining exercise recommendations showed that explanations significantly increased initial trust for the recommendation when the trust was measured as a multidimensional construct. Flanagan and his colleagues proposed a system called EXAIT (Educational eXplainable AI Tools) (Flanagan, Takami, Takii, Dai, Majumdar & Ogata, 2021), which aims to tackle learners' trust and motivation issues behind recommendations made by e-learning systems. Their recent studies developed explainable math exercise recommenders with the Bayesian Knowledge Tracing algorithm (Takami, Dai, Flanagan & Ogata, 2022) and knowledge concepts extracted from textbooks (Dai, Flanagan, Takami & Ogata, 2022).

3. Recommendation Platform

3.1 Recommender Overview

The overview of the recommendation platform is shown in Figure 1.

First, a learner uses learning materials with an e-book reader system. This system is designed to convert the reading logs of the materials to the form of an Experience Application Programming Interface (xAPI) (Advanced Distributed Learning Initiative, 2013), and send them to a Learning Record Store (LRS) (xAPI.com, 2011), a general repository for learning/education records. The reading logs

stored in LRS are used together with information on the difficulty of each material to estimate the learner's English proficiency. The difficulty levels of the materials are calculated in advance from the information on the difficulty of vocabulary in the wordlist. Then, the recommender generates material recommendations with explanations about why the recommendations were made to the learner. These explanations contain recommendation weights (how highly the materials are recommended), and explanatory sentences that explain the basis of the recommendations. Lastly, the generated recommendations are shown to the learner through a recommendation UI.

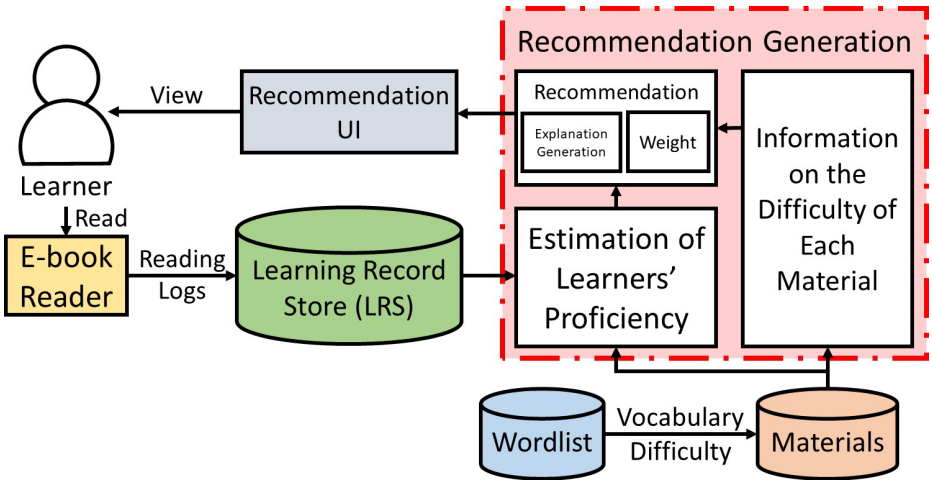


Figure 1. An overview of the recommendation platform.

3.2 Platform Components

3.2.1 GOAL – Recommendation User Interface

The user interface for the recommendation is implemented on the GOAL (Goal Oriented Active Learner) system (Majumdar, Yang, Li, Akçapınar, Flanagan & Ogata, 2018; Li, Majumdar, Chen & Ogata, 2021). The GOAL system is a platform to support students' development of data-informed self-directed learning (SDL) ability. SDL scaffoldings are implemented and provided to students in GOAL using the DAPER (Data collection – Analysis – Planning – Execution monitoring – Reflection) model. The DAPER model-based implementations systematically assist learners in taking initiatives to identify their status in contextual activities, set SMART goals, monitor their progress, and reflect their strategies. The self-directed activity context can be a set of learning activities where learners' learning trace data can be synchronized through learning behavior sensors. In this study it is the extensive reading activity. The SMART goals mean Specific, Measurable, Attainable, Relevant and Time-related goals.



Figure 2. UI of the picture book recommender system implemented in GOAL system

The interface shown in Figure 2 provides at most 5 recommended learning materials. Users can jump directly to the BookRoll, an e-book reader, by clicking the title, and read the recommended material. Each of the recommended materials is followed by the recommendation weight and explanatory sentences that provide the reason why the recommendation was made to the learner. The provided explanation can be shown/hidden when users click a button next to the explanation.

3.2.2 BookRoll – An E-book Reader System

BookRoll, developed by Flanagan & Ogata (2018) is an e-book reader system linked to Moodle (Moodle.org, 2017). Figure 3 shows a user interface of BookRoll.

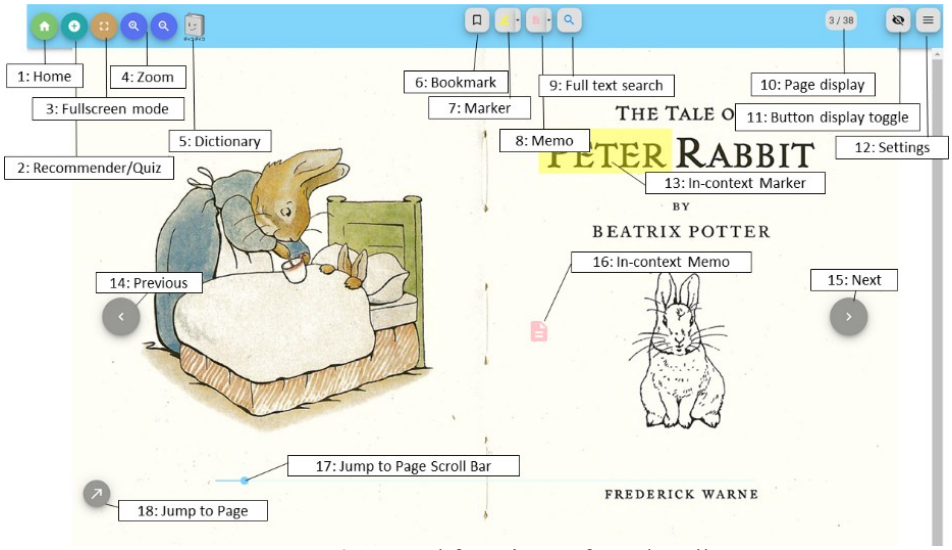


Figure 3. UI and functions of BookRoll

BookRoll provides an e-book page display and many additional functions including bookmark (6 in the figure), text highlighting (7), notes taking (8), text search (9), and recommendation/quiz function (2). When users raise events on BookRoll, such as moving pages, highlighting text, answering quizzes, etc., the detailed information on these events (e.g., the date and the page on which they occurred) is stored as reading logs in a Learning Record Store (LRS). From these features, BookRoll is suitable to gather English learners' learning logs. Thus, we will adopt it as an interface for reading e-books: that is, the whole reading activity in this study will be done through BookRoll and only reading logs performed through BookRoll will be recorded.

4. Recommendation Mechanism

4.1 Material Difficulty Evaluation

4.1.1 Difficulty of Vocabulary

The difficulty of materials should be determined by the difficulty of words consisting of them and their significance in the materials. A material which includes many difficult words is naturally difficult, but if they are not significant in the material, learners can ignore them and understand the content sufficiently. To estimate the difficulty of materials, we used a vocabulary profile and a ranking function in information retrieval.

In this study, the difficulty of vocabulary is used as a rationale for the difficulty of materials. We used "CEFR-J Wordlist Version 1.6" (Yukio Tono Lab., 2020) as a reference for the difficulty of vocabulary. This vocabulary list was constructed for English education in Japan by extracting common vocabulary used in CEFR level texts in each country/region based on corpora generated from English textbooks used in China, Taiwan, and Korea. The list contains 6868 headwords, each of which has 4 CEFR levels of difficulty, A1, A2, B1, and B2.

The difficulty of each word is converted to a numerical value, such as A1 to 1, A2 to 2, B1 to 3, and B2 to 4. Several words have multiple levels according to their multiple meanings, whose difficulties are set as the average of these multiple ones. Besides, identical words with different spellings are considered different words. Therefore, the difficulty of the word t, represented as $D(t)$, should be expressed as $D(t) \in [1, 4]$.

4.1.2 Difficulty of Materials

We computed the difficulty of materials by using the difficulty of each word introduced above, its TF-IDF score (Rajaraman & Ullman, 2011) in each material, and the length of each material. TF-IDF is a ranking function in information retrieval to estimate the relevance of documents to a query, and its score means how relevant the query is to the document. In this study, we assume that every word in the "CEFR-J Wordlist Version 1.6" is used as the query word and interpret the score of the document as the importance of the word in the document.

In this study, we deal with each material as a bag-of-words consisting of the words in the vocabulary list. Namely, when a set of all words in the vocabulary list is represented as Σ, the material d is expressed as $d' \cap \Sigma$, where d' is a bag-of-words of the material corresponding to d.

The difficulty of a material d is expressed as the following equation:

$$D(d) = f\left(c_1 \sum_{t \in d} D(t) \cdot \mathrm{TFIDF}(d, t) + c_2 \frac{|d|}{\max_{d' \in L} |d'|}\right) \quad (f(x) = \log_{10}(1 + x))$$

where the $\mathrm{TFIDF}(d, t)$ is the TF-IDF score of material d given a word t as a query, c_1 and c_2 are constants, and L is a set of all the materials. $|d|$ represents the length of the material d. This formula can be interpreted that the difficulty of each word and its significance in a certain document contributes to the overall difficulty of the document. Besides, since the length of materials is considered to affect its difficulty, the value obtained by normalizing it is also used. Thus, the difficulty of the document will be computed as a total of products of the difficulty of the word and its significance.

To make sure that this equation works, we measured the difficulty of English picture books prepared for extensive reading. These books have metadata about these difficulties. First, the books are classified into 8 groups by these difficulty levels. Each group was labeled in order from A1 to B2/C1 in advance, and the books in the group become more difficult in this order. These levels are according to CEFR levels (Council of Europe, 2001), which shows the achievement of foreign language learners mainly across Europe. Then, we computed the difficulty of each book with the equation above and found the average of the difficulty in each group. At this time, the constants c_1 and c_2 were respectively set as 0.1 and 1. Table 1 shows how many books there are and the average of the estimated difficulty in each group. The results show that as the difficulty level in the metadata increases, the estimated difficulty also increases. To verify that the estimated rank of difficulty matches the rank of difficulty in the metadata, we also computed Kendall's rank correlation coefficient between them, which was a statistically significant value of 0.9820 (p<0.001). We concluded that the equation above could calculate the difficulty of materials correctly from these features.

Table 1. *Results of an evaluation of an equation that computes the difficulty of books*

Difficulty	# of books	Average of estimated difficulty
A1	193	0.857
A1/A2	10	1.196
A2	70	1.210
A2/B1	92	1.210
B1	10	1.743
B1/B2	10	1.807
B2	10	1.908
B2/C1	10	1.995

4.2 Learner's Proficiency Estimation

In general, learners use English materials whose difficulty levels fit their English proficiency levels. According to previous work, there is a linear relationship between the degree of reading comprehension and the percentage of vocabulary in an English material known by the reader (Schmitt, Jiang & Grabe, 2011). This implies that the difficulty of materials the learner is using is a good indicator of the learner's English proficiency. In this study, we adopted the materials used by the learner as input information for estimating the learner's English proficiency level.

According to the difficulty of materials mentioned above, we extract the top 5 most difficult materials from the ones read by the learner. We consider that the difficulties of these 5 materials will best reflect the learner's English ability since learners are considered to change materials to use as their English skills improve. Then, the average difficulty of these 5 materials is computed as the learner's English proficiency. From now on, the learner s's estimated English proficiency will be denoted as $P(s)$.

4.3 Recommendation Generation

4.3.1 Recommendation Weight

The difficulty-based material recommendation is designed to recommend materials whose difficulties are the closest to the learner's estimated English proficiency. The recommended materials should not be too easy or difficult since such materials will have a bad influence on the learning effect. In this study, we introduce the material d's recommendation weight (denoted as $R(s, d)$), i.e., a value that shows how highly the material d is recommended to the learner s. This value should take a large value when the difficulty of the material is close to the learner's English proficiency level. Thus, we define this value as follows:

$$R(s, d) = \log \left(\frac{1}{P(s) - D(d)} \right).$$

This equation takes a large value when the difficulty of the material d (i.e., $D(d)$) is close to the learner s 's English proficiency (i.e., $P(s)$), and takes a small value when it is not. If the recommendation weight $R(s, d)$ takes a larger value, the material d is more highly recommended to the learner s. In the implementation, the value $P(s) - D(d)$ is adjusted so that it does not take a value of 0.

4.3.2 Explanation for the Recommendation

The recommendation algorithm we propose is designed to provide feedback about recommended materials and reasons why these materials were recommended to the learner. This is to make the recommendation persuasive for the learners so that they can agree with the recommendation and learn with self-motivation. This supplementary information includes the learner's estimated English proficiency, the recommendation weights of the recommended materials, and sentences that explain why these recommendations were suitable for the learner.

The recommended weights provided are normalized so that the minimum value is 0.0 and the maximum value is 100.0, to make them easier to understand. The learner's proficiency and the difficulty level of materials are similarly normalized, but the minimum and maximum difficulties of the materials are used at that time.

We prepared 5 types of sentences that explain the reasons why the materials were recommended to the learner. According to the theory of region of proximal learning, when learners make a decision of whether to study, they depend on their belief whether they already know the items: that is, they will choose to not study if they believe they know the item already, and vice versa (Metcalfe & Kornell, 2005). Thus, the explanatory sentences should lead the learners to select the recommended materials to use. The sentences depend on the difference between the difficulty of the recommended material and the learner's English proficiency, i.e., the value of $D(d) - P(s)$. This takes a value close to 0 when the difficulty of the recommended material d (i.e., $D(d)$) is close to the learner s's English proficiency. When the value is smaller than 0, it means that the recommended material is easy for the learner, and

when it is larger than 0, the material is difficult. The recommender provides different sentences according to the difficulty of recommended materials for the learner, as shown in Table 2.

Table 2. *Sentences that explain why the materials were recommended. They depend on the difference between the difficulty of the material and the learner's English proficiency.*

$D(d) - P(s)$	Explanatory Sentence
~ -0.3 (easy)	"This book is easy, but you can learn basic vocabulary with fun from this."
-0.3 ~ -0.1 (a little easy)	"This book is a little easy, but you can learn important vocabulary with this book."
-0.1 ~ 0.1 (average)	"This book is perfect for your English skills!"
0.1 ~ 0.25 (a little difficult)	"This book is a little difficult, but worth trying!"
0.25 ~ (difficult)	"This book is very difficult. Let's challenge!"

These sentences were written to motivate learners to use the recommended materials. The materials whose difficulty is close to the learner's proficiency are explained so that they perfectly fit the learner's level. Besides, even if the materials too easy or difficult are recommended, the recommender explains that they are worth reading.

5. Implementation and Case Study

5.1 Materials

As an example of actual implementations, we first suppose that this recommendation is used for an extensive reading (ER) program, and are going to use English picture books as the learning materials. As of May 2022, 534 English picture books for an ER program are stored at our e-book library, and they are classified into several categories by their difficulties according to the metadata based on the levels of CEFR. The numbers of books are shown in Table 3.

Table 3. *The numbers of picture books classified by CEFR levels*

Difficulty	# of books
pre-A1	27
A1 or A1+	205
A1/A2	10
A2 or A2+	75
A2/B1	126
B1 or B1+	37
B1/B2	27
B2	12
B2/C1	10
N/A	5
Total	534

The books which have difficulty level A1 or A1+ are the most common of all (for beginners, 205 books), and the books which have difficulty level A2/B1 are the second most common (for elementary-intermediate learners, 126 books). The 5 N/A level books did not have information about the level.

5.2 Case Stud

Using the picture books introduced above, we generated material recommendations with the proposed mechanism. We set 2 conditions, “easy” and “difficult”, and compared the recommendation result in each one. In the “easy” condition, a virtual reader was supposed to have read 9 easy books, and 9 difficult books in the “difficult” condition. The difficulty of the picture books read by the virtual reader was according to the metadata of the books. Then, material recommendations for the reader were generated in both conditions. The reader’s English skill level was also estimated.

Table 4. *The estimated English skill levels and the recommendation results*

Condition	Read books and its difficulty	Estimated English skill level	Average difficulty of recommended books
Easy	9 books with A1	42.69 / 100	42.702 / 100
Difficult	3 books with B2 & 6 books with B2/C1	81.11 / 100	81.222 / 100

Table 4 shows the reader’s estimated English skill level and average difficulty of the recommended picture books. In the “easy” condition, the reader’s estimated skill level was lower than that in the “difficult” condition. This shows that an English skill level of a reader who has read easy books was estimated to be low, and the opposite was true for difficult books. The same can be said about the difficulty of recommended books: that is, easy books were recommended in the “easy” condition, and difficult ones were in the “difficult” condition. The average difficulty of the recommended books was very close to the reader’s estimated skill level in both conditions. From this result, we can say that the proposed recommendation mechanism could recommend picture books whose difficulties were close to the reader’s English skills.

6. Discussion

This recommendation mechanism depends on 3 sources: difficulty of vocabulary, difficulty of materials, and learners’ learning logs. We consider that a source which is the most important for making the recommendations persuasive for learners, that is, for recommending materials whose difficulties are closest to the learners’ skills, is the learners’ learning logs. This mechanism estimates the learners’ skill levels from materials they have used before, but we admit there is still room for improvement on the algorithm. Learners’ personality including their skill levels is not so simple that this recommendation should take into account more detailed information about this. Besides, even if this exploitation of the further information is implemented, the estimated learners’ skills might conflict with their actual levels. The system then should allow the learners to modify their knowledge states so that the estimated skills correctly reflect it.

Moreover, the proposed recommendation can make one-way recommendations from the system to learners, but we should also consider introducing an interactive recommendation mechanism. This recommendation is designed to recommend materials that have moderate difficulty levels for each learner. However, sometimes it might be not enough for learners who are willing to improve their skills actively or relearn what they have learned before. In the current implementation, the difficulties of the recommended materials will be close to the learners’ skills. Therefore, the algorithm should allow the learners to adjust the difficulty levels of recommended materials according to the learners’ learning objectives.

The implementation and the experiment will be conducted in the setting of an ER program, but we believe that this recommendation will be effective even with more general English materials.

7. Conclusion and Future Work

In this paper, we proposed an English learning material recommendation that takes into account learners’ English skills changing over time and generates its explanations to make it persuasive. In this algorithm, we adopted TF-IDF, one of the information retrieval techniques, to realize these features of the recommendation. We also described the implementation of this recommendation and verified its learning and motivating effect.

The overall system using the proposed recommendation is made of multiple components. As all the components including front-end and back-end are fully implemented, conducting an experiment in an actual educational environment is our urgent task. We also have to prepare the plan for the evaluation of the experimental results, for instance, measuring learners' reading speed and comprehension and comparing the results before and after the experiment. The learners' system usage will be an important index for measuring the effect on their motivation.

For future work, we set two goals for the purpose of the improvement of the recommendation: 1) refinement of the algorithm and 2) utilizing more detailed information on learners. First, we recognize that there is still room for improvement in the estimation of the difficulty of learning materials. The estimation in the current implementation only depends on the vocabulary included in the materials and their length. The difficulty of the materials should be measured from more viewpoints, such as the complexity of grammar and how much expert knowledge will be needed to understand the material. We may be able to use a grammar profile to extract grammar items from English sentences in the materials.

Second, we may be able to combine the proposed recommendation with learner models. Learner models can manage learners' various characteristics including learning styles, preferences, and personality as well as learning skills. For example, the difficulty of one material may be different from learner to learner because of their background knowledge or preferences. By introducing the recommendation to a system using open learner models (Bull & Kay, 2010), the recommendation will be more persuasive for learners because it can utilize their learning states and provide them as more detailed explanations of the recommendation to the learners. This will improve learners' trust for the system and their motivation for learning. Moreover, by letting the learners check their learning states and the suitable levels of materials, the system can assist in improving learners' meta-cognition.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, NEDO JPNP20006 and JPNP18013. This work was also supported by JST, the establishment of university fellowships towards the creation of science technology innovation, Grant Number JPMJFS2123.

References

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2020, March). Complementing educational recommender systems with open learner models. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 360-365).

Advanced Distributed Learning Initiative. (2013). Experience API (xAPI) Standard. Retrieved May 26, 2022, from https://adlnet.gov/projects/xapi/.

Aher, S. B., & Lobo, L. M. R. J. (2013). Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data. *Knowledge-Based Systems, 51*, 1-14.

Bobadilla, J., Serradilla, F., & Hernando, A. (2009). Collaborative filtering adapted to recommender systems of e-learning. *Knowledge-Based Systems, 22*(4), 261-265.

Bourkoukou, O., & El Bachari, E. (2018). Toward a hybrid recommender system for e-learning personnalization based on data mining techniques. *JOIV: International Journal on Informatics Visualization, 2*(4), 271-278.

Bull, S., & Kay, J. (2010). Open learner models. In *Advances in intelligent tutoring systems* (pp. 301-322). Springer, Berlin, Heidelberg.

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR). Retrieved May 26, 2022, from https://www.coe.int/lang-CEFR.

Dai, Y., Flanagan, B., Takami, K., & Ogata, H. (2022, March). Design of a User-Interpretable Math Quiz Recommender System for Japanese High School Students. In *Proceedings of the 4th Workshop on Predicting Performance Based on the Analysis of Reading Behavior*.

Duffy, M. C., & Azevedo, R. (2015). Motivation matters: Interactions between achievement goals and agent scaffolding for self-regulated learning within an intelligent tutoring system. *Computers in Human Behavior*, 52, 338-348.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. *Knowledge Management & E-Learning: An International Journal, 10*(4), 469-484.

Flanagan, B., Takami, K., Takii, K., Dai, Y., Majumdar, R., & Ogata, H. (2021). EXAIT: A symbiotic Explanation Learning System. In *29th International Conference on Computers in Education Conference, ICCE 2021* (pp. 404-409).

George, G., & Lal, A. M. (2019). Review of ontology-based recommender systems in e-learning. *Computers & Education, 142*, 103642.

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000, December). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work* (pp. 241-250).

Hsu, M. H. (2008). A personalized English learning recommender system for ESL students. *Expert Systems with Applications, 34*(1), 683-688.

Klašnja-Milićević, A., Vesin, B., Ivanović, M., & Budimac, Z. (2011). E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & education, 56*(3), 885-899.

Li, H., Majumdar, R., Chen, M. R. A., & Ogata, H. (2021). Goal-oriented active learning (GOAL) system to promote reading engagement, self-directed learning behavior, and motivation in extensive reading. *Computers & Education, 171*, 104239.

Majumdar, R., Yang, Y. Y., Li, H., Akçapınar, G., Flanagan, B., & Ogata, H. (2018). GOAL: Supporting learner's development of self-direction skills using health and learning data. In *Proceedings of the 26th International Conference on Computers in Education (ICCE2018)* (pp. 406-415).

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of memory and language, 52*(4), 463-477.

Moodle.org. (2017). Moodle – open-source learning platform. Retrieved May 26, 2022, from https://moodle.org/TUG.

Ooge, J., Kato, S., & Verbert, K. (2022, March). Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th International Conference on Intelligent User Interfaces* (pp. 93-105).

Rajaraman, A., & Ullman, J. D. (2011). Mining of Massive Datasets. *Cambridge University Press*.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26-43.

Shishehchi, S., Banihashem, S. Y., Zin, N. A. M., & Noah, S. A. M. (2011, June). Review of personalized recommendation techniques for learners in e-learning systems. In *2011 international conference on semantic technology and information retrieval* (pp. 277-281). IEEE.

Takami, K., Dai, Y., Flanagan, B., & Ogata, H. (2022, March). Educational Explainable Recommender Usage and its Effectiveness in High School Summer Vacation Assignment. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 458-464).

Tarus, J. K., Niu, Z., & Mustafa, G. (2018). Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning. *Artificial intelligence review, 50*(1), 21-48.

Truong, H. M. (2016). Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. *Computers in human behavior, 55*, 1185-1193.

xAPI.com. (2011). What is an LRS? Learn more about Learning Record Stores. Retrieved May 26, 2022, from https://xapi.com/learning-record-store/.

Yukio Tono Lab., Tokyo University of Foreign Studies. (2020). CEFR-J Wordlist Version 1.6. Retrieved February 2022, from https://www.cefr-j.org/download.html.

Zou, D., & Xie, H. (2018). Personalized word-learning based on technique feature analysis and learning analytics. *Journal of Educational Technology & Society, 21*(2), 233-244.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

The Effects of Learning with Digital Storytelling on Classroom Engagement in a Grade 6 English Class

Yingxue LIU[a*] & Alex Wing Cheung TSE[b*]

The University of Hong Kong, Hong Kong SAR
[a]*u3586811@connect.hku.hk, [b]*awctse@hku.hk

Abstract: Digital storytelling (DST), an innovative pedagogical approach, has the potential to facilitate active learning in the classroom. Literature indicates that engagement is essential to effective learning. However, the investigation of the impacts of DST on learning engagement in English as a Foreign Language (EFL) education in China is scarce, especially in the K-12 context. Moreover, although related efforts have been made, few studies examine these effects with a clear focus on classroom engagement. To address this issue, this study adopted a 5-factor classroom engagement framework (affective engagement, behavioral engagement-compliance, behavioral engagement-effortful class participation, cognitive engagement, and disengagement) to investigate whether learning with DST affects multidimensional classroom engagement in a Grade 6 English class. A quasi-experimental approach with pretest and posttest was conducted, and data were collected from 70 Grade 6 students via Classroom Engagement Inventory. The quantitative results using ANCOVA revealed that there was a significant difference in overall classroom engagement between the experimental (n=35) and control groups (n=35) in the posttest, particularly in terms of the behavioral engagement-compliance and cognitive engagement. In other words, it was found that, when learning with DST in English class, students were more likely to adhere to the class norms and arrangements, as well as devote mental efforts to be strategic and self-regulated. Yet, no significant difference was manifested in affective engagement and behavioral engagement-effortful class participation, suggesting that the use of DST may not bring about more positive emotions and effortful involvement in English class. This study contributes to a realistic perspective for EFL educators and teachers to understand the potential of the DST approach through the lens of classroom engagement. The second stage of this project has been exploring the reason behind the research results, shedding light on possible mediators within DST that could facilitate or impair certain aspect(s) of classroom engagement. Further implications for DST integration in the EFL classroom can be proposed.

Keywords: Digital storytelling (DST), Classroom engagement, English as a Foreign Language (EFL), Quasi-experimental study

1. Introduction

In recent years, the pervasive technologies used in education give rise to an innovative pedagogical approach, digital storytelling (DST). As a merger of multimedia technologies and ancient storytelling (Rahimi & Yadollahi, 2017), this technology-integrated approach has shown potential to facilitate active learning in the classroom (Robin, 2016). When learning with DST, students are motivated to construct and represent their own meaning as a social practice (Bruner, 1990; Meadows, 2003; Van Gils, 2005).

In terms of classroom practice, however, the use of DST in English as a Foreign Language (EFL) education in China is rare (Niemi, Niu, Vivitsou, & Li, 2018), particularly in the K-12 context. There is a gap between the technology provision and its effective integration in the EFL classroom (Jacobsen, 2001). In fact, many English teachers are facing challenges in adopting the DST approach in their classes due to the lack of a shared vision of its potential to improve students' learning and effective professional training (Sadik, 2008).

Engagement is essential to learning success (Herrington, Oliver, & Reeves, 2003). Classroom engagement can serve as an indicator to measure the effectiveness of the DST practice in the classroom.

Recent studies have shown that effective DST integration in EFL class could bring about many learning benefits, such as increased learning motivation, enhanced student engagement, and better academic performance (e.g., Hung, 2019; Niemi & Multisilta, 2016). Nevertheless, there is still a lack of research investigating the effects of DST in EFL class with a clear focus on classroom engagement using a tailored theoretical framework and instrument. Therefore, it is meaningful and necessary to examine the impacts of learning with DST on classroom engagement in EFL education.

This study aims to explore to what extent and in what aspects learners are engaged in a Grade 6 DST English class. Based on the quantitative results yielded from this research, a more comprehensive analysis of how learning with DST affects various aspects of classroom engagement could be conducted in the second-stage study. It can contribute to a realistic perspective for EFL educators and teachers to understand the educational affordances of the DST approach to facilitate EFL learning in terms of classroom engagement (Sadik, 2008).

The following research questions guided the present study: *Does learning with digital storytelling affect multidimensional classroom engagement in a Grade 6 English class?*

2. Literature review

2.1 Digital Storytelling

In recent years, rapid advances in Information and Communication Technology have sparked many multimedia tools, providing opportunities for educators and learners to not only make use of but also create their own digital content (Churchill & Barratt-Pugh, 2020). DST, a combination of conventional storytelling with a mixture of digital content, including text, pictures, audio, music, and video (Robin, 2016), has gained popularity in many contexts, particularly in education. As a modern expression of ancient storytelling art, digital stories can be interpreted as multimodal artifacts that require students to make sense of, and then share their personal experiences through digital narratives as a social practice (Bruner, 1990; Meadows, 2003; Van Gils, 2005). Despite its promising educational benefits, to our knowledge, the effective integration of DST into K-12 EFL education in China is scarce.

2.2 Digital Storytelling in English as a Foreign Language Learning

As a student-centered approach (Barrett, 2006), DST is beneficial for student learning in various aspects. When learning with DST, students are situated in an active learning environment where they build knowledge constructively (Chubko, Morris, McKinnon, Slater, & Lummis, 2020). It facilitates collaboration and communication skills among learners as they need to go through continuous interaction and cooperation when weaving digital stories (Smeda, Dakich, & Sharda, 2014). Also, it provides opportunities to develop 21st-century skills (Niemi & Multisilta, 2016), such as critical thinking and problem-solving skills, since students as creators are expected to manage and represent ill-structured problems in their digital stories (Robin, 2008). When involved in the manipulation of a wide range of digital tools, students' digital literacy can also be enhanced (Churchill & Barratt-Pugh, 2020). Overall, the use of DST helps to facilitate a constructivist learning environment in the classroom (Smeda et al., 2014) and thus has great potential to provide better learning outcomes. So far, several empirical studies have confirmed this hypothesis (e.g., Sadik, 2008; Yang & Wu, 2012).

In EFL teaching and learning, the DST approach inherits advantages to improving EFL students' learning performance and skills. Many studies emphasized that learning with DST can promote EFL learners' language learning performance. For instance, it is suggested that students can significantly improve their writing and reading proficiencies when drafting English scripts for their digital stories (Yoon, 2014). The pervasive use of English in digital story creation and sharing provides ample opportunities for students to improve listening skills and practice speaking skills (e.g., Tsou, Wang, & Tzeng, 2006; Hwang et al., 2016; Yang & Wu, 2012). Apart from positive impacts on students' language capabilities, the DST approach in English class can also empower EFL learners' 21st-century skills as mentioned above, such as collaborative skills (Hung, 2019), critical thinking (Yang & Wu, 2012), creativity (Liu, Tai, & Liu, 2018), multiliteracy skills (Chubko et al., 2020). Thus, more practices in DST integration are needed in the EFL classroom.

Over the past decade, there has been an increasing number of researchers examining learning motivation and engagement in DST integrated EFL learning. Previous studies suggest that thanks to the meaningful scenario and interactive construction process in digital story creation, EFL learners are more likely to be engaged in deep and meaningful learning (Smeda et al., 2014) with increased motivation and interests (Sadik, 2008). However, these studies mostly center on students' long-term course and program-level engagement, which is overall student engagement, without proper contextualization (Lu, Xie, & Liu, 2022). There is a lack of research examining the effects of learning with DST in English class with a clear focus on classroom engagement. Literature on classroom engagement illustrates that learning motivation and classroom engagement are reciprocally related (Jang, Kim, & Reeve, 2012), and high-level classroom engagement can help to stimulate student motivation, thus triggering more educational benefits (Smeda et al., 2014).

To summarize, the DST approach is conducive to students' learning from multiple facets, especially in EFL education. It is worthwhile to examine the effectiveness of the DST approach in the EFL classroom and explore how it benefits students' learning through the lens of classroom engagement.

2.3 Why Classroom Engagement instead of Student Engagement

Engagement can be interpreted as the physical and mental efforts that students commit to learning (Astin, 1984). As the prerequisite of learning (Pittaway, 2012), engagement is essential to learning success (Herrington et al., 2003). The measurement of engagement has been one of the crucial considerations in educational research.

However, the conceptualization and contextualization of learning engagement in literature are quite blurred (Baron & Corbin, 2012). The term "student engagement" was used indiscriminately in many studies, and there is a dearth of well-articulated and context-specific measures of classroom-level engagement (Fredricks & McColskey, 2012).

Classroom engagement differs from student engagement in its nature. It refers to a student's active participation in classroom learning activities (Reeve, Jang, Carrell, Jeon, & Barch, 2004). That is to say, "student engagement", which is usually measured at the school level, cannot directly and precisely reflect engagement in the classroom setting. For instance, the same student may be highly engaged in one class, but not in others (Darr, 2012). Moreover, this student may be actively involved in in-class activities, but not be invested in after-class learning. Thus, the measurement of learning engagement should be context specific. Additionally, the investigation of classroom engagement and student engagement serve different functions (Wang et al., 2014). The measure should be at the classroom level when the research aims to evaluate the effectiveness of an intervention implemented in the classroom context (Wang et al., 2014). Hence, this study focuses on classroom engagement to examine the impacts of learning with DST in English class.

As stated above, according to the relevant literature, the use of DST in EFL classes has the potential to increase learning motivation and engagement (e.g., Hung, 2019; Niemi & Multisilta, 2016; Smeda et al., 2014). However, the understanding of the extent to which and in what aspects learners are engaged in such classes remains unexplored (Sadik, 2008). Moreover, although these studies were supposed to measure classroom-level engagement, the instruments employed were mostly centered on overall student engagement. It may lead to invalidities in research results and make it difficult for teachers and educators to understand the essence and effects of the DST practice in EFL class. Therefore, to investigate the impacts of learning with DST on EFL learning, it is necessary to have a clear focus on and tailored conceptual framework of classroom engagement.

2.4 Theoretical Framework of Classroom Engagement

There are three dimensions of classroom engagement: affective (emotional), cognitive, and behavioral (Fredricks, Blumenfeld, & Paris, 2004). This typology is well-grounded in the learning engagement across disciplines (Tang & Hew, 2022). At the classroom level, high-level affective engagement represents students' positive emotions, such as interest, enjoyment, and enthusiasm during class (Skinner, Kindermann, & Furrer, 2009). Behavioral engagement is often manifested by observable behaviors, such as overt attention, task participation, and question-asking in class (Fredricks et al.,

2004). Cognitive engagement refers to mental efforts devoted to classroom learning, such as meaningful processing and strategy use (Fredricks et al., 2004). It is often reflected through students' task investment in terms of being strategic or self-regulating (Lamborn, Newmann, & Wehlage, 1992). Based on the 3-dimension engagement model, Wang et al. (2014) designed a self-report instrument, Classroom Engagement Inventory, to measure multidimensional classroom engagement comprehensively. To ensure its internal consistency and reliability, they conducted several rounds of exploratory factor analysis and confirmatory factor analysis with large sample sizes. As a result, a 24-items, 5-factor Classroom Engagement Inventory was finalized with a stable structure that is invariant between groups (e.g., age, subject, gender, etc.). One of the significant contributions of this inventory is that it expands the traditional 3-dimension engagement model to a new 5-factor classroom engagement framework (Figure 1), including affective engagement (AE),

behavioral engagement-compliance (BEC), behavioral engagement-effortful class participation (BEE), cognitive engagement (CE), and disengagement (D). The advantage of this framework mainly resides in the behavioral level, as it differentiates students who are invested and fully immersed in class (effortful class participation) from those who just show obedience to the classroom norms but do not invest in and enjoy classroom learning (compliance). As Fredricks et al. (2004) once emphasized, it is of importance to separate compliant participation (e.g., classroom rules adherence) from autonomy involvement or self-directed learning behaviors. An articulated and customized instrument makes a thorough understanding of classroom engagement in DST EFL class a possibility.

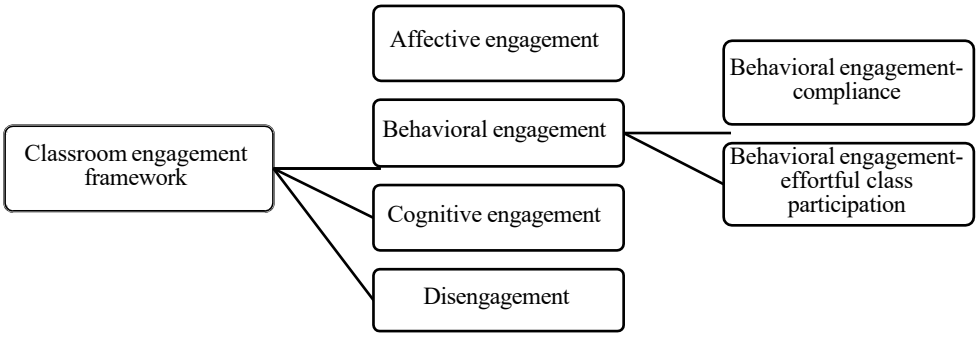


Figure 1. 5-factor Classroom Engagement Framework

3. Methodology

3.1 Research Context

This study recruited 70 students from two Grade 6 English classes at a public primary school in mainland China. Students with similar English academic performance attended as the experimental (n=35) and control groups (n=35). A weekly-2-hour (4 sessions in total) English class based on the central English curriculum was conducted with or without DST for one month. To control experimental conditions, these two classes were taught by the same English teacher with an identical teaching syllabus. Yet, the instructional design of English class for both groups varied from each other. For the control group, the instructor delivered a teacher-led content-based English class, where students were guided to revisit key knowledge learned in the first half term. Whereas for the experimental group, students needed to demonstrate and elaborate on a specific knowledge learned in the first half term via their digital story.

To ensure a relatively smooth DST English class, one school technology coordinator was invited to solve technological issues timely before and during class. The researcher was responsible for DST teacher professional training, DST learning materials design and development, and DST learning

environment support. Students in the experimental group took English class in the computer lab, and those in the control group joined English class in a traditional classroom. Multiple technological tools for learning with DST were available in the computer lab, such as digital cameras, projectors, microphones, video editing software, etc.

3.2 Research Design and Data Collection

To improve the reliability and validity of this research, a pretest-posttest quasi-experimental approach (Figure 2) was adopted. Quantitative data on classroom engagement were collected by Classroom Engagement Inventory (Wang et al., 2014) shortly before and after the intervention in both experimental and control groups. The sampling method for quantitative analysis was simple random sampling.

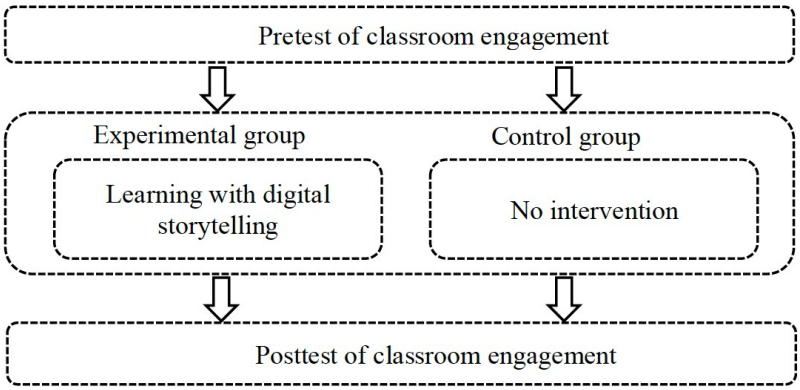


Figure 2. The Overall Research Design

3.3 Digital Storytelling Course Design

The researcher first explained the experiment process to all participants and obtained their consent before the study. Two teacher workshops were then conducted to ensure the English teacher's mastery of the basic technological pedagogical content knowledge to incorporate DST in teaching (Robin, 2016). The first workshop introduced the relevant concepts of DST, and the second was to demonstrate the use of Microsoft PowerPoint to create simple but clear digital stories. Previous studies have shown that PowerPoint is one of the most frequently used constructive tools for students' presentations and projects (Lim & Tay, 2003). For primary school students with limited digital literacy, an easy-to-use DST editing software plays a key role in its successful construction. In this study, students mainly turned to the recording function in Microsoft PowerPoint to combine text, pictures, animation, audio narration, and music for their digital stories.

Considering the nature of DST, it is powerful in instructing people on a particular concept or practice (Robin, 2006). Thus, students in the experimental group were expected to co-construct digital stories with the purpose of instructing viewers on a specific disciplinary concept or practice. In other words, students, acting as instructors, were asked to deliver explicit demonstration and elaboration on a specific knowledge they learned in English class in their digital stories. Following the Seven Elements of Digital Storytelling (Robin, 2006) and a 12-step Guideline for Digital Storytelling (Robin, 2016), this research made slight changes according to local context and adopted an 8-step 3-layer Digital Storytelling Learning Model (Figure 3) as the instructional design framework. The first layer comprises basic steps of digital story creation: 1) Choose a topic, 2) Research on the topic, 3) Draft the script, 4) Find and create images, 5) Create a storyboard, 6) Record audio narration, 7) Build the digital story and 8) Share the digital story (Robin, 2016, p23). The second layer is the guideline for students and the third layer is the guideline (Robin, 2016, p24) for the facilitator based on the ADDIE model (Analysis, Design, Development, Implementation, Evaluation).

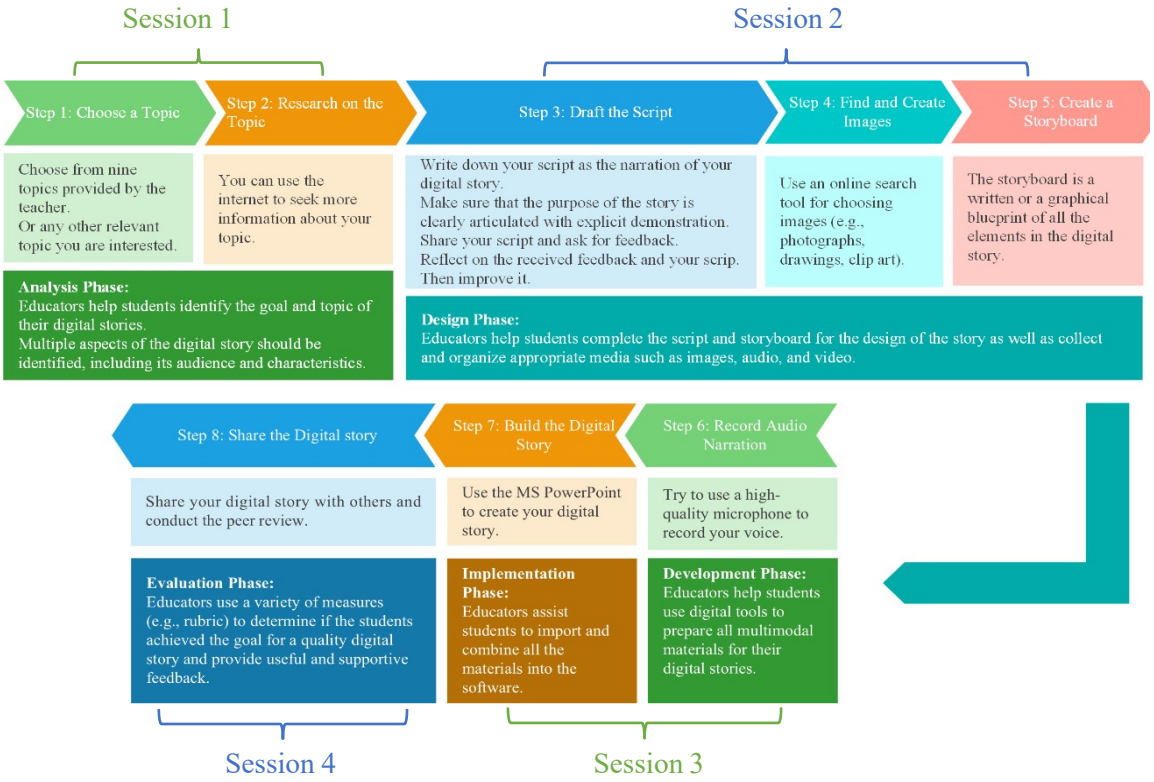


Figure 3. 8-step 3-layer Digital Storytelling Learning Model

3.4 Data Analysis

The quantitative data were analyzed using IBM SPSS 27.0. First, Cronbach's alpha of this inventory was calculated using the pretest data to ensure the internal consistency of each dimension of the Classroom Engagement Inventory in the research context. Next, descriptive analysis and repeated measures of ANCOVA were conducted to respectively provide background information and compare the overall classroom engagement and each specific factor of it between the experimental and control groups using both pretest and posttest data. In this study, the pretest scores of classroom engagement in two groups were considered as covariates that may affect the posttest results but cannot be controlled. Thus, ANCOVA with a confidence interval of 0.95 was adopted to eliminate these undesirable impacts and adjust posttest results according to the covariates. The independent variable was the intervention (learning with or without DST) or "group", and the dependent variable consisted of participants' overall classroom engagement and each specific aspect of it in the posttest. Before performing ANCOVA, all the basic assumptions of it were verified. The Kolmogorov Smirnov test was used to determine whether the data had a normal distribution. Homogeneity of the regression slopes was verified to make sure there was no interaction between the covariates and the intervention, and Levene's test was performed to check the homogeneity of variance across groups respectively.

4. Results

Before conducting the posttest, Cronbach's alpha was calculated using the pretest results to ensure the internal consistency for each factor within the CEI. The Cronbach's alpha for all factors were larger than .7 (α_{AE} = .708, α_{BEC} = .881, α_{BEE} = .889, α_{CE} = .959, α_{D} = .863), confirming that Classroom Engagement Inventory is a reliable instrument with internal consistency. The regression statistics (p < .05) of interaction between the covariates and "group" ($p_{\text{Group*Overall-Pretest}}$ = .992, $p_{\text{Group*Pre-AE}}$ = .143, $p_{\text{Group* Pre-BEC}}$ = .862, $p_{\text{Group* Pre-BEE}}$ = .326, $p_{\text{Group* Pre-CE}}$ = .78, $p_{\text{Group* Pre-D}}$ = .098) confirmed that there was a lack of interaction between them. Thus, our data met the homogeneity of regression slopes assumption. The homogeneity of variance was tested by Levene's test, and the results (F_{AE} = .154, p > .05; F_{BEC} = 3.865, p > .05; F_{BEE} = 3.605, p > .05; F_{CE} = .574, p > .05; F_{D} = 29.546, p < .05) suggested that there was

no difference between the variances across groups except for the factor, disengagement. Thus, we mainly refer to the overall engagement and the following specific factors of classroom engagement, including affective engagement, behavioral engagement-compliance, behavioral engagement-effortful class participation, and cognitive engagement in the following data analysis.

Descriptive analysis (Table 1) of the pretest and adjusted posttest results of classroom engagement is shown in the table below, suggesting that the overall engagement in both experimental and control groups increased in the posttest.

Table 1. *Descriptive and ANCOVA Results of Classroom Engagement in Pretest and Posttest*

Dimension	Pretest		Posttest		ANCOVA		
	Mean	SD	Mean	SD	F (1, 70)	*p*	η^2
Overall Engagement							
Experimental	4.36	0.52	4.44[a]	0.47	6.399	0.014*	0.087
Control Group	4.02	0.49	4.18[a]	0.45			
Affective Engagement (AE)							
Experimental	3.99	0.64	4.14[a]	0.73	0.814	0.370	0.012
Control Group	3.89	0.53	4.02[a]	0.48			
Behavioral Engagement-Compliance (BEC)							
Experimental	4.50	0.55	4.56[a]	0.48	8.172	0.006*	0.109
Control Group	4.16	0.52	4.26[a]	0.42			
Behavioral Engagement-Effortful Class Participation (BEE)							
Experimental	4.38	0.60	4.34[a]	0.60	1.981	0.164	0.029
Control Group	3.96	0.54	4.16[a]	0.53			
Cognitive Engagement (CE)							
Experimental	4.47	0.52	4.48[a]	0.48	5.753	0.019*	0.079
Control Group	4.03	0.54	4.22[a]	0.48			

Note: Experimental Group n = 35, Control Group n = 35. *p<0.05. Posttest means are adjusted according to the covariates.

Repeated measures of ANCOVA were performed to investigate whether there were significant differences in overall and each factor of classroom engagement between the experimental and control group in the posttest. The results (Table 1) revealed that, in overall, students in the experimental group exhibited significantly higher classroom engagement than those in the control group (F = 6.399, p = .014; η^2 = .087). Specifically, there were significant differences in behavioral engagement-compliance (F = 8.172, p = .006; η^2 = .109) and cognitive engagement (F = 5.753, p = .019; η^2 = .079) between two groups in the posttest. Therefore, the initial conclusion can be drawn that learning with DST in English class has significantly improved the overall classroom engagement in the experimental group, particularly in terms of behavioral-compliance engagement (e.g., listen carefully and complete assignments) and cognitive engagement (e.g., revisit key knowledge for deeper learning, and monitor and reflect on one's learning). Yet, no significant difference was found in affective engagement (F = .814, p = .37; η^2 = .012) and behavioral engagement-effortful class participation (F = 1.981, p = .164; η^2 = .029) between two groups in the posttest. This indicated that learning with DST in English class would not necessarily bring about more positive emotions (e.g., interested, excited, proud, happy, etc.) and effortful involvement in classroom learning (e.g., spontaneously form new questions and actively communicate and cooperate with others).

5. Conclusion and Discussion

This study adopted a pretest-posttest quasi-experimental approach to examine the effects of learning with DST on classroom engagement. The research findings showed that learning with DST could significantly improve the overall engagement in Grade 6 English class, particularly regarding behavioral engagement-compliance and cognitive engagement, though no significant difference was

manifested in affective engagement and behavioral engagement-effortful class participation. It deepens the existing understanding of the effectiveness of DST in EFL class and contributes to a realistic perspective for EFL educators and teachers to recognize the potential of this technology-integrated approach to facilitate classroom engagement.

The above findings are consistent with some studies on DST in EFL learning (e.g., Hung, 2019; Niemi & Multisilta, 2016; Sadik, 2008; Smeda et al., 2014), suggesting that learning with DST could enhance learning engagement in the classroom behaviorally (compliance) and cognitively. Students were motivated to listen carefully and learn deeply as active and self-regulated learners in digital story co-construction activities (Sadik, 2008). Yet, in terms of affective engagement, the research results may contradict some previous studies that positive emotions were not necessarily exhibited in a DST English class. In this study, struggling with problems beyond their reach in DST co-creation, some students became uninterested, frustrated, or even anxious. It was in line with one previous study that a few students expressed that learning activities in DST EFL class were unnecessary, difficult, and time-consuming (Hava, 2021). In addition, mixed results were reported regarding behavioral engagement. Despite many students conforming to the class arrangement, they may not be fully immersed in this class with effortful participation. Previous research has provided some clues for both low affective engagement and behavioural engagement-effortful class participation, indicating that the difficulty of the assigned task of DST may negatively influence students' involvement (Sugimoto, 2011), and students may have difficulties in maintaining group coordination when working cooperatively (Järvenoja, Volet, & Järvelä, 2013). The second-stage qualitative research of this project has been analyzing the student interview data to see whether can explain why learners became more or less engaged in certain aspect(s) in this class. Students' insufficient digital and language literacy, inadequate digital story creation time, collaborative failure of group work, ineffective instruction, scaffolding and feedback from the instructor, have been initially identified as potential moderators that may impair specific aspect(s) of classroom engagement. The detailed results and findings will be reported and discussed in another paper.

There were several limitations in this study. First and foremost, although we advanced the engagement study on DST EFL by looking at classroom engagement, how this engagement fluctuated and changed with time and context remains unanswered. Future research is expected to examine situational engagement in DST EFL class, taking dynamic environmental and individual factors embedded in real-time, fine-grained data into consideration (Xie, Heddy, & Vongkulluksn, 2019). Second, the duration of the intervention (one month) was relatively short, thus its impact on classroom engagement could be limited. Future study can be implemented with longitudinal experiments. Third, the sample size of this study was small with 70 students, and all data were collected by a self-report tool, which could be subjective and inaccurate, leading to limited generality of research findings. Large-scale research using diverse instruments, such as classroom observation and assessment of students' digital story artifacts as e-portfolios can be considered. Notably, though significant differences were manifested in some aspects of classroom engagement, the effect sizes of them were relatively small (e.g., $\eta^2_{BEE} = .029 < .06$), indicating insufficiently substantial differences (Cohen, 2013). More empirical studies with rigorous research design examining the effects of learning with DST in EFL classrooms are still in need.

Nevertheless, to the best of our knowledge, this is the first study examining the effects of learning with DST through the lens of multidimensional classroom engagement in an EFL class. To present a comprehensive picture of how learning with DST affects classroom engagement, the data of the second-stage qualitative research are expected to explain in what ways learners became more or less engaged in this class. After that, implications for future DST integration in EFL learning can be proposed.

Acknowledgements

We would like to thank all the people who participated in this research and helped review this paper.

References

Astin, A. W. (1984). Student involvement: A developmental theory for higher education. *Journal of college student personnel, 25*(4), 297-308.

Baron, P., & Corbin, L. (2012). Student engagement: rhetoric and reality. *Higher Education Research & Development, 31*(6), 759-772.

Barrett, H. (2006). *Researching and Evaluating Digital Storytelling as a Deep Learning Tool*. [online]Available at: https://electronic portfolios.com/portfolios/SITEStorytelling2006.pdf.

Bruner, J. (1990). *Acts of meaning*: Harvard university press.

Chubko, N., Morris, J. E., McKinnon, D. H., Slater, E. V., & Lummis, G. W. (2020). Digital storytelling as a disciplinary literacy enhancement tool for EFL students. *Educational technology research and development, 68*(6), 3587-3604.

Churchill, N., & Barratt-Pugh, C. (2020). The digital entanglement of humanities, literacy, and storytelling. In *Reconceptualizing the Digital Humanities in Asia* (pp. 141-154): Springer.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*: Routledge.

Darr, C. W. (2012). Measuring student engagement: The development of a scale for formative use. In *Handbook of research on student engagement* (pp. 707-723): Springer.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research, 74*(1), 59-109.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In *Handbook of research on student engagement* (pp. 763-782): Springer.

Hava, K. (2021). Exploring the role of digital storytelling in student motivation and satisfaction in EFL education. Computer Assisted Language Learning, 34(7), 958-978.

Herrington, J., Oliver, R., & Reeves, T. C. (2003). Patterns of engagement in authentic online learning environments. *Australasian Journal of Educational Technology, 19*(1).

Hung, S.-T. A. (2019). Creating Digital Stories EFL Learners' Engagement, Cognitive and Metacognitive Skills. *Journal of Educational Technology & Society, 22*(2), 26-37.

Hwang, W.-Y., Shadiev, R., Hsu, J.-L., Huang, Y.-M., Hsu, G.-L., & Lin, Y.-C. (2016). Effects of storytelling to facilitate EFL speaking using Web-based multimedia system. *Computer Assisted Language Learning, 29*(2), 215-241.

Jacobsen, D. M. (2001). Building Different Bridges: Technology Integration, Engaged Student Learning, and New Approaches to Professional Development.

Jang, H., Kim, E. J., & Reeve, J. (2012). Longitudinal test of self-determination theory's motivation mediation model in a naturally occurring classroom context. *Journal of educational psychology, 104*(4), 1175.

Järvenoja, H., Volet, S., & Järvelä, S. (2013). Regulation of emotions in socially challenging learning situations: an instrument to measure the adaptive and social nature of the regulation process. *Educational Psychology, 33*(1), 31-58.

Lamborn, S., Newmann, F., & Wehlage, G. (1992). The significance and sources of student engagement. *Student engagement and achievement in American secondary schools*, 11-39.

Lim, C. P., & Tay, L. Y. (2003). Information and communication technologies (ICT) in an elementary school: Students' engagement in higher order thinking. *Journal of Educational Multimedia and Hypermedia, 12*(4), 425-451.

Liu, K.-P., Tai, S.-J. D., & Liu, C.-C. (2018). Enhancing language learning through creation: The effect of digital storytelling on student learning motivation and performance in a school English course. *Educational technology research and development, 66*(4), 913-935.

Lu, G., Xie, K., & Liu, Q. (2022). What influences student situational engagement in smart classrooms: Perception of the learning environment and students' motivation. *British journal of educational technology,* doi:10.1111/bjet.13204

Meadows, D. (2003). Digital storytelling: Research-based practice in new media. *Visual communication, 2*(2), 189-193.

Niemi, H., & Multisilta, J. (2016). Digital storytelling promoting twenty-first century skills and student engagement. *Technology, Pedagogy and Education, 25*(4), 451-468.

Niemi, H., Niu, S., Vivitsou, M., & Li, B. (2018). Digital Storytelling for Twenty-First-Century Competencies with Math Literacy and Student Engagement in China and Finland. *Contemporary educational technology, 9*(4), 331-353.

Pittaway, S. M. (2012). Student and staff engagement: developing an engagement framework in a faculty of education. *Australian Journal of Teacher Education, 37*(4), 37-45.

Rahimi, M., & Yadollahi, S. (2017). Effects of offline vs. online digital storytelling on the development of EFL learners' literacy skills. *Cogent Education, 4*(1), 1285531.

Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and emotion, 28*(2), 147-169.
Robin, B. (2006). *The Educational Uses of Digital Storytelling*. Paper presented at the Society for Information Technology & Teacher Education International Conference 2006, Orlando, Florida, USA.
Robin, B. (2008). Digital storytelling: A powerful technology tool for the 21st century classroom. *Theory into practice, 47*(3), 220-228.
Robin, B. (2016). The power of digital storytelling to support teaching and learning. *Digital Education Review*(30), 17-29.
Sadik, A. (2008). Digital storytelling: A meaningful technology-integrated approach for engaged student learning. *Educational technology research and development, 56*(4), 487-506.
Skinner, E. A., Kindermann, T. A., & Furrer, C. J. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and psychological measurement, 69*(3), 493-525.
Smeda, N., Dakich, E., & Sharda, N. (2014). The effectiveness of digital storytelling in the classrooms: a comprehensive study. *Smart Learning Environments, 1*(1), 1-21.
Sugimoto, M. (2011). A mobile mixed-reality environment for children's storytelling using a handheld projector and a robot. *IEEE Transactions on Learning Technologies, 4*(3), 249-260.
Tang, Y., & Hew, K. (2022). Effects of using mobile instant messaging on student behavioral, emotional, and cognitive engagement: a quasi-experimental study. *International Journal of Educational Technology in Higher Education, 19*.
Tsou, W., Wang, W., & Tzeng, Y. (2006). Applying a multimedia storytelling website in foreign language learning. *Computers & Education, 47*(1), 17-28.
Van Gils, F. (2005). *Potential applications of digital storytelling in education.* Paper presented at the 3rd twente student conference on IT.
Wang, Z., Bergin, C., & Bergin, D. A. (2014). Measuring engagement in fourth to twelfth grade classrooms: The Classroom Engagement Inventory. *School Psychology Quarterly, 29*(4), 517.
Xie, K., Heddy, B. C., & Vongkulluksn, V. W. (2019). Examining engagement in context using experience-sampling method with mobile technology. Contemporary Educational Psychology.
Yang, Y.-T. C., & Wu, W.-C. I. (2012). Digital storytelling for enhancing student academic achievement, critical thinking, and learning motivation: A year-long experimental study. *Computers & Education, 59*(2), 339-352.
Yoon, T. (2014). Developing multimodal digital literacy: the application of digital storytelling as a new avenue for effective English learning with EFL elementary school students in Korea.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Learning Path Recommendation System for English Grammar Quiz Using Knowledge Map

Naomichi TANIMURA[a*], Kensuke TAKII[a], Brendan FLANAGAN[b] & Hiroaki OGATA[b]

[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*tanimuranaomichi@gmail.com

Abstract: Learning path recommendation systems assist learners in learning in an optimal sequence of knowledge items. Conventional systems do not take into account the knowledge that the learner has already acquired, and thus knowledge that has already been acquired may be included in the recommended learning path. Therefore, in this paper, we focus on English grammar and propose a system that recommends learning paths based on the learning history, excluding already acquired knowledge.

Keywords: English as a foreign language, Learning path, Knowledge Map, E-learning Recommender

1. Introduction

Systems that provide learning paths, in which learning materials are arranged so that learners can efficiently acquire multiple pieces of knowledge, have been previously studied (Shi et al. 2020; Zheng et al. 2013). Conventional learning path systems have the problem that the learning paths provided may include knowledge that has already been acquired.
In this study, we propose a learning path recommendation system that uses a graph structure called a "knowledge map" and learning history to recommend learning paths for English grammar knowledge that has not yet been learned. A knowledge map consists of nodes describing concepts or items and links describing relationships between nodes (Shishehchi et al., 2011; Lee and Segev, 2012). The proposed system recommends learning paths based on the English grammar knowledge map created based on the CEFR-J Grammar Profile, a measure of English learning achievement for Japanese learners developed by Tono et al. (2016) and the level of proficiency of each grammar based on the answer history of quizzes answered on the existing e-book reading system BookRoll (Ogata et al. 2015).

2. Related Research

2.1 Learning Path Recommendation Systems Using Knowledge Maps

Many previous studies have developed learning path recommendation systems, some of which use knowledge maps. However, few of them take into account whether the learner has already acquired the knowledge. The system proposed in this study recommends learning paths that take into account already acquired grammar knowledge.

2.2 Picture-book Recommendation Systems Using Vocabulary Knowledge Maps

The English Picture-book Recommender System (Takii, Flanagan, and Ogata, 2021) recommends picture books containing words that have not been learned yet that are related to previously learned words based on reading history and vocabulary knowledge maps (Flanagan et al., 2019). The system takes into account the relevance and proficiency of knowledge. The system proposed in this paper further takes into account the knowledge learning order.

3. Overview of the recommendation system

3.1 CEFR-J Grammar Profile

CEFR-J (Ishii and Tono, 2016) is an English proficiency framework developed by Tono et al. by applying the Common European Framework of Reference for Languages (CEFR) to Japan. The CEFR-J Grammar Profile is a classification of English grammar according to the CEFR-J levels, and covers the major grammar studied in junior and senior high school. Each grammar has a regular expression, by which we can extract it from given English sentences. In this study, we use this method to link each quiz with grammar knowledge items.

3.2 Generating English Grammar Knowledge Maps

The English Grammar Knowledge Map used in this study is generated by the following method.

1. Each item in the CEFR-J Grammar Profile is considered as a node, and the similarity between regular expression strings is calculated for each combination of two nodes. Weighted links are generated using the similarity as weights, and an undirected graph is obtained.
2. Based on the learning order of each English grammar in junior high and high schools, the graph obtained in step 1 is transformed into a directed graph with links starting with the grammar learned first and ending with the grammar learned later.
3. A maximum directed global tree is obtained by performing branch pruning on the graph obtained in step 2 using the minimum spanning tree algorithm.

3.3 System Configuration

Figure 1 shows an overview of the proposed quiz recommendation system. The knowledge map and links between quizzes and grammars are stored in a database. When learners answer quizzes on BookRoll, the answer logs are stored in the LRS after processing. These logs are used for quiz recommendation along with the information stored in the user's knowledge state store and the knowledge dependencies of the knowledge map.

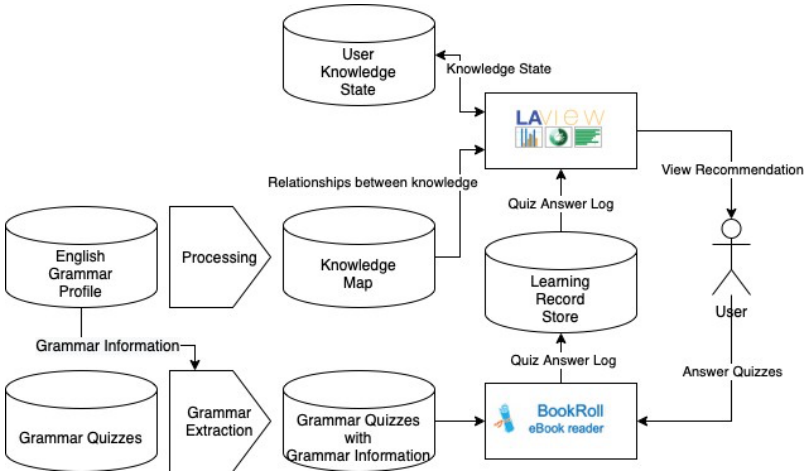


Figure 1. Structure of the English quiz recommendation system

3.4 Learning Path Recommendation Mechanism

The system generates a learning path as follows.

1. Obtain the grammar contained in quizzes that the learner has answered incorrectly in the past from the LRS.
2. Based on the knowledge map, search for some knowledge that should be acquired before mastering the knowledge obtained in step 1. At this time, only the knowledge that is yet to be acquired is selected with reference to the proficiency.

The system gives the quizzes corresponding to the knowledge in the learning pathway obtained in the above procedure a priority as follows, and encourages students to answer the quizzes starting with the quiz according to these priorities.

- If a quiz corresponds to one knowledge of a learning pathway, it is given the same priority as that knowledge.
- If a quiz corresponds to more than one piece of knowledge in a learning pathway, it is given the same priority as the knowledge to be learned later.

4. Conclusion and Future Work

In this paper, we proposed a quiz recommendation system for efficient English language learning in an optimal sequence. The system uses the English grammar knowledge map and the learner's proficiency level to sequence and recommend quizzes for learning based on knowledge that has yet to be learned. Compared to conventional systems, this system is expected to provide more efficient support by taking into account not only the learning order but also the learner's knowledge proficiency. We plan to conduct experiments with Japanese high school students to verify the effectiveness of the system.

Acknowledgements

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (Exploratory) 21K19824, and NEDO JPNP20006.

References

Flanagan, B., Chen, M. A., & Lecailliez, L., Majumdar, R., Akçapinar, G., Ocheja, P., & Ogata, H. (2019). Automatic vocabulary study map generation by semantic context and learning material analysis. Proceedings of the 27th International Conference on Computers in Education 698–702.

Lee, J. H., & Segev, A. (2012). Knowledge maps for e-learning. Computers & Education, 59(2), 353–364.

Ogata, H., Yin, C., Oi, M., Okubo, F., Shimada, A., Kojima, K., & Yamada, M. (2015). E-Book-based learning analytics in university education. In International conference on computer in education (ICCE 2015) (pp. 401–406).

Shi, D., Wang, T., Xing, H., & Xu, H. (2020). A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning. Knowledge-Based Systems, 195(105618), 105618.

Shishehchi, S., Banihashem, S. Y., Zin, N. A. M., & Shahrul, S. A. (2011). Review of personalized recommendation techniques for learners in e-learning systems. 2011 International Conference on Semantic Technology and Information Retrieval, STAIR 2011, 277–281.

Takii, K., Flanagan, B., & Ogata, H. (2021). An English Picture-book Recommender System for Extensive Reading Using Vocabulary Knowledge Map. Companion Proceedings of the 11th International Conference on Learning Analytics and Knowledge, 40-42.

Zheng, Q. H., Huang, X. J., & Zhang, C. (2013). A Learning Navigation Path Algorithm based on knowledge map in E-Learning. 2013 2nd International Conference on E-Learning and E-Technologies in Education, ICEEE 2013, 180–185.

Ishii, Y. & Tono, Y. (2016). A frequency analysis of grammar categories for the CEFR-J Grammar Profile. Proceedings of the Twenty-second Annual Meeting of the Association for Natural Language Processing, 777–780.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Role of Peer Assessment in Facilitating Computational Thinking among Pre-Service Teachers

Xin Pei VOON[a], Su Luan WONG[b*], Lung Hsiang WONG[c], Mas Nida MD KHAMBARI[d] & Sharifah Intan Sharina SYED ABDULLAH[e]
[a,b,d,e]*Faculty of Educational Studies, Universiti Putra Malaysia, Malaysia*
[c]*National Institute of Education, Nanyang Technological University, Singapore*
*suluan@upm.edu.my

Abstract: Peer assessment has been employed as an effective learning strategy to enhance cognitive practices such as problem-solving and reflection practices. A case study was conducted to explore the effects of online peer assessment as a learning strategy of computational thinking among pre-service teachers. The peer comments were analysed and coded by adopting a coding scheme of comments to investigate the significance of the peer-reviewing process in facilitating the learning of computational thinking. Each student was required to design a lesson by integrating the computational thinking facets into their lesson plan. Upon submitting the lesson plan to the instructor, they were engaged in a blind review process. The students worked in a group to review and provide constructive comments on their peers' lesson plans. By adapting a peer-reviewing cognitive process model, this article provides evidence that the peer-reviewing process played a critical role in facilitating the learning of Computational Thinking. The findings indicate that the peer assessment strategy can develop pre-service teachers' problem-solving competencies. It was suggested that all students be informed about the purposes and learning benefits of the peer-review process to optimise the learning outcomes.

Keywords: Peer assessment, peer-reviewing, learning strategy, computational thinking, pre-service teacher, teacher education

1. Introduction

According to Sondergaard (2009), peer assessment can increase student engagement and social interaction, encourage learning reflection, and develop diagnosis and evaluation skills. Peer reviewing that emphasises formative feedback fits under the category of peer assessment. Peer reviewing is an evaluative problem-solving process that engages students in detecting text problems, diagnosing them, and generating solutions to improve the problems (Cho & MacArthur, 2011). The higher-order thinking (HOT) is defined as "new information and information stored in memory interrelates and/or rearranges and extends this information to achieve a purpose or find possible solutions in complex situations." (Lewis & Smith, 1993, p.136). Problem-solving is a cognitive process through which knowledge, skills, and personal experiences are mobilised to identify problems, find solutions, and resolve conflicts effectively (Hoi et al., 2018). Computational thinking (CT) emphasises individuals' cognitive process (Shute et al., 2017), developing the pre-service teachers' CT competencies and enabling them to integrate CT into their teaching contexts. A case study was conducted to explore peer assessment as a learning strategy in facilitating pre-service teachers' learning of CT, guided by the research question: How does the peer-reviewing process foster problem-solving competencies in the learning of CT among pre-service teachers?

2. Theoretical framework: Reviewing as a problem-solving process

This study adapted the cognitive model developed by Flower et al. (1986) to investigate how the reviewing process fosters the students' HOT, particularly problem-solving skills. Figure 1 presents three features of the model: (1) the process results in the call for the use of knowledge; (2) various information flows between processes - in practice, the flow of information from one process to another result in a shift in the reviewer's attention, in which the reviewer's cognition shifts from one subprocess to another; (3) the flow of information between knowledge and process as well as between

different processes is highly reciprocal. The process of Evaluation, Diagnosis, and Strategy Selection has led the reviewer to re-think the goals and suggest making some necessary modifications, which is to redraft or revise the task based on the diagnosis of the problems. Finally, this may lead the reviewers out of the reviewing process, reflect on their task and conduct necessary modifications.

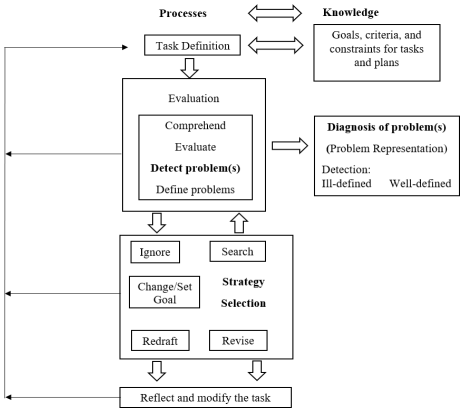


Figure 1. The cognitive model in the peer review process adapted from Flower et al. (1986)

3. Research design and data analysis

This study adopted a case study approach. The participants were selected based on the following criteria: (1) pursuing their Bachelor's Degree in Education; (2) enrolling in the Educational Technology course. Thirty-eight participants were in their third-year undergraduate studies and had no prior knowledge of CT. The participants attended a twelve-week CT module session, in which they learned about the six CT facets of cognitive process: decomposition, abstraction, algorithms, debugging, iteration and generalisation. Each participant was required to design a CT-integrated lesson plan: integrating CT facets based on the learning objective(s) of the lesson plan. Subsequently, the participants were divided into groups of four to review each other's lesson plans and submit their feedback through TEAMMATES. TEAMMATES is an online peer feedback system for student team projects. In implementing the anonymous peer assessment, each participant was given an account with a code as identification, allowing them to post and check their peers' feedback from TEAMMATES. Participants were given a CT framework (adapted from Shute et al., 2017) and guiding questions to facilitate the review process. The criteria of the review process: (1) identification of the integration of CT facets, (2) justification of the integration of CT facets based on the learning objective(s) of the lesson. The first author acted as the instructor who conducted the lessons online due to the campus closure during the Covid-19 pandemic. The lesson recordings were then uploaded onto Google Drive.

In data analysis, the comments generated by the participants in the reviewing process were coded. Each comment was assigned into three main categories: strength, weakness or off-task. First, the 'praise' category under the strength: good remarks on what constituted the strength of the task. Second, the weakness category consists of three sub-categories: (1) 'problem detection': identify a problem without diagnosing the problem; (2) 'problem diagnosis': detected and diagnosed a problem without suggesting any possible solution; (3) 'solution suggestion': the idea included a solution to improve the problem regardless of whether the idea had a statement of problem diagnosis. Third, off-task category: task-unrelated comments. The investigator triangulation involved two researchers, and the inter-coder reliability was conducted where one researcher coded all of the comments, and the second coder independently coded 50% of randomly selected comments. When there were discrepancies in the coding results, the discussion was conducted until a consensus was reached.

4. Results, discussion, and conclusion

The students' comments were coded based on the three main code categories. However, one comment was coded in more than one category. The inter-coder reliability for the coding was 0.83, which suggested a near-perfect agreement (Cohen, 1960). A total of 480 comments from all 38 students were analysed and assigned into five categories. The frequency of each code indicated that the students

were actively engaged in the problem-solving processes, in which they needed to review their peers' work by identifying and diagnosing the problem(s) and proposing appropriate solutions. 369 (76.9%) comments were categorised under the strength (praise). The comments that were categorised as weakness made up 256 (53.3%): 98 (20.4%) problem detection, 58 (12.1%) problem diagnosis, 100 (20.8%) solution suggestion, and 20 (4.2%) off-task.

First, the findings reported from the analysis of peer comments have demonstrated that peer-reviewing facilitates participants' CT learning experience while receiving feedback from peers and producing feedback for peers, concurred with the studies conducted by Cho and MacArthur (2010). The students' perspectives on the differences between receiving and generating reviews, teasing out the HOT processes activated through the learning-by-reviewing, and highlighting the role of these processes in enhancing their CT learning. This study's results concur with the findings of Nicol et al. (2014), i.e., the peer assessment process is perceived as beneficial because it actively engages students in HOT, for instance, critical thinking, reflection, and learning transfer.

Second, the findings of this study support the effectiveness of peer assessment in scaffolding problem-solving processes by considering problems, constructing the problem spot, and articulating contextual constraints. The problem-solving process stimulates the participants' HOT to identify the problems or deficiencies in the CT-integrated lesson design, thus prompting them to diagnose and propose appropriate solutions based on their interpretation of the lesson design. HOT processes were activated to develop the participants' problem-solving competencies. The findings are also congruent with the studies of Çevik (2015) and Nicol et al. (2014), in which the online peer assessment strategy positively contributes to students' problem-solving skills. The participants' problem-solving skills improved more substantially by providing feedback than receiving feedback because participants engaged in problem-solving processes while providing the feedback, such as identifying and analysing problems, developing, evaluating and justifying solutions.

To conclude, this study suggests two key insights: (1) peer assessment facilitates students' CT learning in designing CT-integrated lessons, and (2) peer assessment can be used as a learning strategy to develop students' HOT; in particular, reviewing processes develops students' problem-solving competencies. The results of this study provide important implications for teacher educators and educators, as teachers are key agents of change in teaching and learning (Voon et al., 2019). The researchers suggest that all students be informed about the purposes and benefits of the peer-review process. Limitations needed to take into account: First, the sample of this study comprises a homogenous group of participants, and the results may not be generalisable to other settings. Second, the modules were conducted online, and a few participants could only watch the lesson recordings due to the unstable internet connection.

References

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328-338.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Hoi, P. T., Bao, D. Q., Nghe, P. K., & Nga, N. T. (2018). Developing Problem-Solving Competency for Students in Teaching Biology at High School in Vietnam. *American Journal of Educational Research*, *6*(5), 539-545.

Çevik, Y. D. (2015). Assessor or assessee? Investigating the differential effects of online peer assessment roles in the development of students' problem-solving skills. *Computers in Human Behavior*, 52, 250-258.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*(1), 73-84.

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision. *College Composition and Communication*, *37*(1), 16-55.

Korkmaz, Ö., Çakir, R., & Özden, M. Y. (2017). A validity and reliability study of the computational thinking scales (CTS). *Computers in Human Behavior*, 72, 558-569.

Nicol, D., Thomson, A., & Breslin, C. (2014). Re-thinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, *39*(1), 102-122.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142-158.

Voon, X. P., Wong, L. H., Chen, W., & Looi, C. K. (2019). Principled practical knowledge in bridging practical and reflective experiential learning: Case studies of teachers' professional development. *Asia Pacific Education Review*, *20*(4), 641-656.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Authentic Assessments for Digital Education: Learning Technologies Shaping Assessment Practices

Tristan LIM[ab*], Swapna GOTTIPATI[b] & Michelle CHEONG[b]
[a]*Nanyang Polytechnic, Singapore*
[b]*Singapore Management University, Singapore*
*tristan_lim@nyp.edu.sg

Abstract: Assessment is a powerful lever that affects learning. To better inform educators on authentic assessment practices within digital education, this paper takes us through a meta-analysis of existing literature between 2011 to 2021. The study evaluates the following research question: "How are emerging technologies shaping authentic assessment practices within digital education for the period between 2011 and 2023". To aid with the forecasting, we utilize the EDUCAUSE Horizon Reports, which provide the predictions of emerging technology. This study affirms the importance of immersive learning technologies, followed by ubiquitous and adaptive learning technologies as key growth research interest areas impacting authentic assessments in the coming years. Results from this study found that the application of immersive learning technologies on authentic assessments is presently highly prioritized as a research area. However, this may likely be overtaken by applications of adaptive learning technologies in the near future. In addition, the application of ubiquitous learning technologies on authentic assessments offers much promise, having grown extensively in research intensity since 2011, overtaking learning objects. The study can help researchers understand past, present and future metatrends of emerging technologies shaping authentic assessment practices, and identify promising potential areas for research and/or education focus.

Keywords: Authentic Assessment, Assessment Design and Practice, Emerging Technology and Trends, Digital Education, Higher Education

1. Introduction

Digital education system is the incorporation of technologies and tools to assist the teaching and learning process (Williamson, 2016). Over the years, traditional education system is transformed into digitalized modes through computing tools, data mining algorithms, database infrastructures, enterprise architectures, cloud technologies, cyber-physical systems, and software and hardware platforms. Specifically, digital software and hardware systems influence the design and implementation of assessments to large extents (Dillenbourg, 2016).

Assessment is perceived with importance by many scholars as "*the most powerful lever teachers have to influence the way students respond to a course and behave as learners*" (Gibbs, 1999). Assessment practices have been under growing attention (Postareff et al. 2012), with increasing demand for the application of more innovative and effective practices to measure students' achievements, building upon higher levels of accountability and expectations for learning (Asghar, 2012). In digital education, there is a predilection towards connecting assessment activities to authentic and meaningful technology practices similar to the many manifestations of practicing technologists and innovators with diverse tools and processes. To formalize the definition of authentic assessments in this paper, we adopt the definition by National Research Council (1996), which defines authentic assessment as "*assessment exercises [that] require students to apply knowledge and reasoning to situations similar to those they will encounter in the world outside the classroom, as well as to situations that approximate how [practicing professionals] do their work*."

In the educational research setting, most of the questions studied by educators revolve around "*what works*" (Ross, 2017). Education technology is a complex and integrated process that includes defining and designing philosophical and ethical aspects, methodologies and approaches, and

theoretical and applied methods of using technology to support education. On the other hand, learning technology is related to the use of the technology to perform instruction or learning, within educational processes. In other words, learning technology has a more precise scope, and includes an explicit focus on technology, techniques, and strategies for the digital education system. The main aim of this study is to analyze "what works" in the design and implementation of assessments in the digital education system. In particular, we explore learning technologies for assessments to better inform educators on authentic assessment practices within digital education. This paper takes us through a meta-analysis of existing literature on learning technologies between 2011 to 2021. We first extracted the database of emerging technology from the EDUCAUSE Horizon reports (EDUCAUSE, n.d.) and applied bibliometric method to analyze trends and emerging technology, through the indicators including publications, citations, and occurrences of keywords or key phrases.

The study evaluates the following research question: *How are emerging technologies shaping authentic assessment practices within digital education for the period between 2011 and 2023?*

Through an evaluation of the key research question, the study hopes to (i) understand emerging learning technologies and trends, and their influence on authentic assessment practices, and (ii) provide an analysis of how authentic assessment practices may evolve in the coming years. This can help researchers decide where to focus the research efforts in the coming years and the educators on the various assessment models and methods.

2. Trend Analysis

From a well-cited research effort by Martin et al. (2011), the Horizon Report (HR) is a good starting point of reference for an annual technology prediction impacting education. HR explores and predicts technology developments impacting education from a broad perspective, inclusive of teaching, learning, and assessments. Since 2004, HR has released free and publicly accessible annual predictions of key global emergent technologies in education; no other resource has issued annual trend publications for a longer period of time. Published by EDUCAUSE (EDUCAUSE, n.d.), HR has been a widely read and highly regarded educational technology resource for education practitioners. A blunt (but useful) indicator of the volume of references, commentaries or discussions related to HR – with albeit a publication track record of 17 years, approximately 234,000 results can be found from a google search of HR (excluding contents from EDUCAUSE home site) (Google search entry: *+"horizon report" -site:educause.edu*).

The annual HR methodology involves surveying expert panelists to yield trend perspectives – in a short term (year of report), midterm (two to three years from year of report), and long term (four to five years from year of report). It is an ever-fruitful endeavor to continually improve survey methodologies – an undertaking EDUCAUSE undertook since its 2020 report. This resulted in a change in methodology. For each annual publication from 2020, the trend prediction time horizons were removed, in favour of discussions on broad trending technologies and practices impacting education by the expert panelists. In our study, we analysed HR reports from year 2011 to 2021, that cover technology metatrends between 2011 to 2023.

To evaluate metatrend hypotheses, methods including technology investment and sales, technology uptake reports, and patent analyses were considered by several studies. However, these methods yielded unreliable insights (Martin et al., 2011). The former two methods involve quantitative data which were difficult to validate. The latter method of patent analyses relates more to commercial and industry, rather than education-related research, which limited its usefulness. Bibliometric is a commonly applied method to analyze trends and emerging technology, via indicators including publications, citation and occurrences of words, among others (Daim et al., 2006).

In our study, to corroborate results from the metatrend hypotheses, bibliometric analysis is applied. At the time of this work, there are around 5,000 articles relating to education research involving bibliometric analysis on Google Scholar (Google Scholar search entry: *intitle:(education AND bibliometric)*). Recent education-related works that used bibliometric analysis include Huang et al. (2020), and Marín-Marín et al. (2019). Bibliometric analysis requires sufficient and verifiable data for temporal comparison. Google Scholar is a good tool for bibliometric analysis, as it provides a comprehensive, quantifiable, and verifiable search across a wide range of academic repositories, including JStor, Springer, Wiley, Mendeley and IEEE, among others.

3. Research Methodology

This study follows the methodology applied by Martin et al. (2011) for answering our research question. The steps are as follows:

1. *Metatrend review*: Record trends of technologies and/or practices affecting education from Horizon Report between 2011 to 2021. Create visual representation of the trends, by grouping technologies according to their similarities and creating detailed views for each group. Trends of technologies and/or practices that have no relation to assessments are excluded. Analyze the key metatrends observed and map against assessment practices in published literature within the period under study.
2. *Bibliometric analysis:* Perform bibliometric analysis as an objective evaluation of emerging technologies shaping assessment practices. This step includes measuring the impact of predicted technologies in assessment practices, by searching and computing the number of papers published between 2011 to 2021 for each relevant technology group in bibliographic databases. Using Google Scholar as a bibliometric analysis tool, we narrow the scope of selection to the selected technology group, assessment type, education approach, education level, publication title, and year of publication.
3. *Evaluation*: As the number of authentic assessment-related education papers varies each year, a weighting factor is applied to the results for each technology group to allow for comparison across different years. Weighting factor (WF_i), as shown in Equation (1), is computed by dividing the mean number of papers published across the period under study, against the number of publications at year *i*.

$$WF_i = \frac{\bar{p}}{p_i} = \frac{\frac{1}{N}\sum_{i=2011}^{2021} p_i}{p_i} \quad (1)$$

where $\bar{p}$ represents the mean number of publications from 2011 to 2021, p_i represents the number of publications at year *i*, *i* represents the period {2011, 2012… 2021}, and *N* represents the total number of years. With the application of the weighting factor, the years with lower publications of authentic assessment-related education papers will not be unduly penalized. Finally, we provide an objective assessment of the HR predictions against the bibliometric results.

4. Results

4.1 Metatrends and Technology Clustering

Results of step 1 of the methodology are compiled in Figure 1. In Figure 1, the rows represent the year of the horizon report, and the columns represent the prediction period of the technologies. Note that prior to 2020, the HR survey methodology predicts the technology trend for the year of the report, for two-to-three years ahead (mid-term), and four-to-five years ahead (long term). From 2020 and beyond, HR only discusses emerging technology trends based on the year of the report as such the forward period predictions of two years and beyond do not apply.

Each technology cluster is grouped by color based on similarity: (i) orange cluster represents technology trends related to ubiquitous learning technologies, (ii) yellow represents technology trends related to adaptive learning technologies (AI), (iii) blue represents technology trends related to immersive learning technologies, and (iv) grey represents technology trends related to learning objects.

4.2 Bibliometric Statistics

A detailed bibliometric analysis research is performed to assess if the defined technology group had predicted impact on authentic assessments in digital education. Figure 2 is the graphical representation of the number of papers available in Google Scholar for each technology group in each of the analyzed years, weighted based on Equation (1).

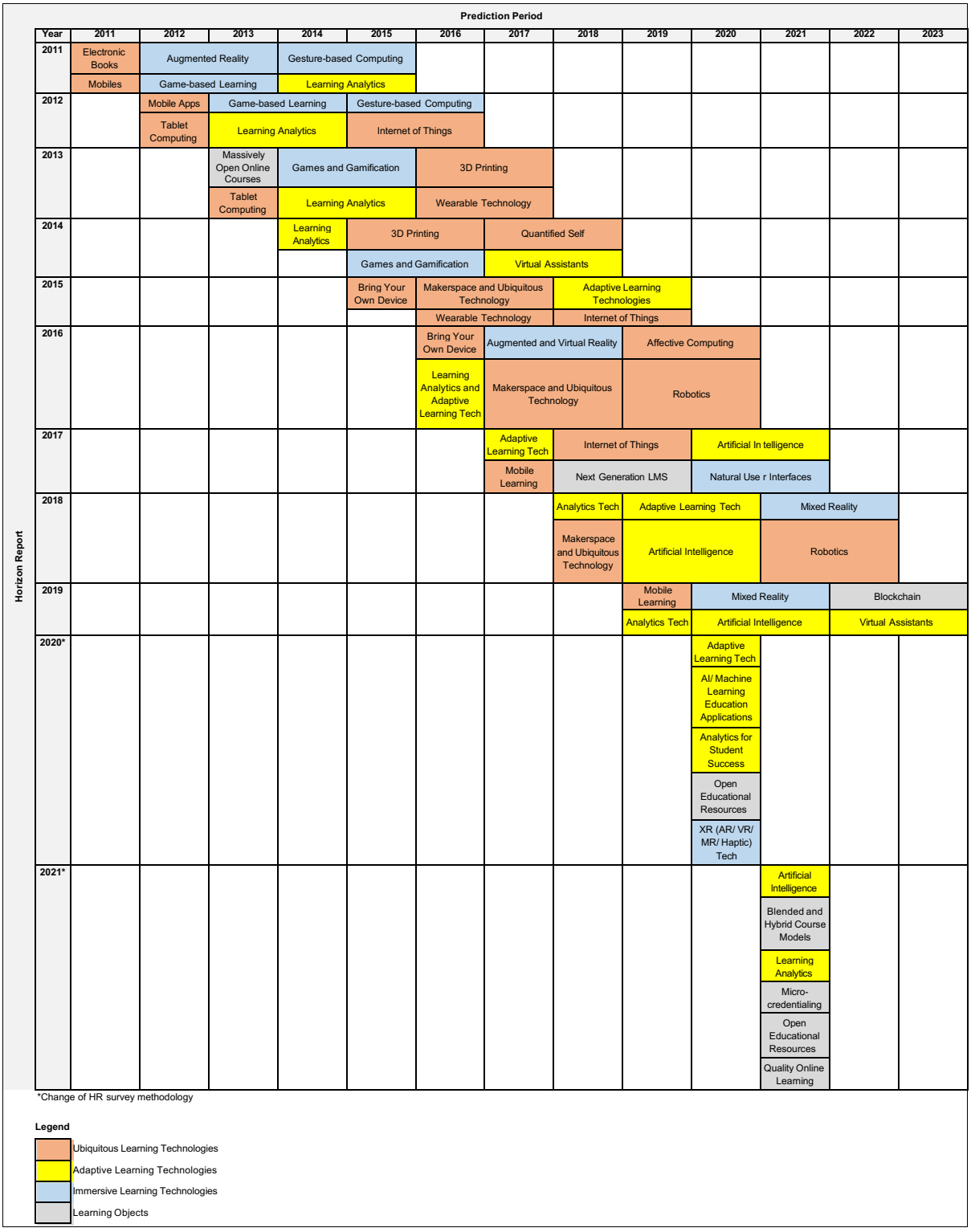


Figure 1: Emerging technologies impacting education based on Horizon Reports from 2011 to 2021.

5. Discussion

Immersive learning technologies generated the highest publication interest as of 2021, after weighing the effect of overall publication numbers. Its publication numbers were consistently around 800 to just under 950 per annum between 2011 to 2021. It was consistently mentioned in nine of the 11 years under study in HR. The consistency in research interest is evident in the continued HR predictions tied to immersive environment-related education and assessments, including game-based environments, and mixed reality environments (e.g., virtual reality, augmented reality). While the initial cost for equipment

and development of the immersive assessments may be high, this is arguably offset by the relatively cheap adaptation of the content for different authentic scenarios, the lower cost of maintenance, ease of portability and the relatively modest space requirement (Engelbrecht, Lindeman & Hoermann, 2019). This technology area has high ecological validity for authentic assessments but is still plagued by issues such as relative immaturity of technology, uncertain skill transfer, and lack of specialization and testing of systems.

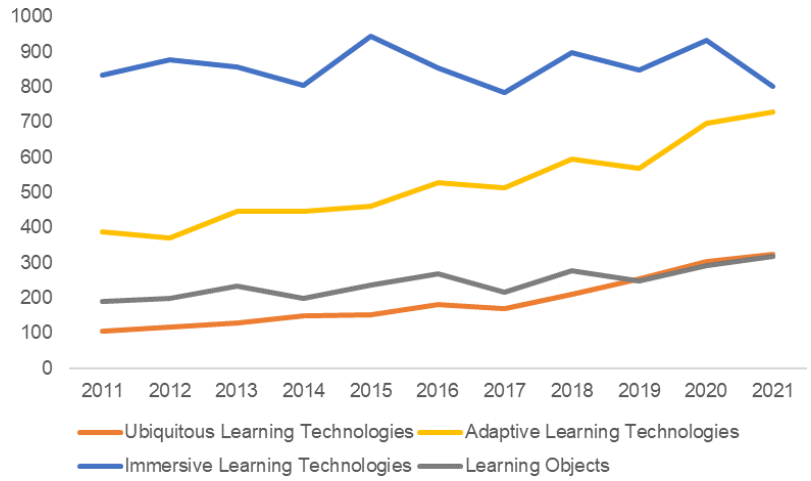


Figure 2: Number of papers published in Google Scholar in each technology group between 2011 to 2021, adjusted with weighting factor in Eq. (1).

Adaptive learning technologies generated the second highest publication interest as at 2021, after weighing for the effect of overall publication numbers. It is noteworthy that while adaptive learning technologies represented less than half of the research interest in immersive learning technologies in 2011, the gap between the two technology groups has shrunk to less than 10%. In fact, the weighted number of publications on adaptive learning technologies almost tripled in the 11 years under study. There are well documented ethical debates regarding the use of AI in assessments, and in particular, grading. However, among other benefits, the possibilities of the use of analytics and AI in assessments as part of an authentic learning intelligent tool to measure personalized cognitive skills, and provide analysis and feedback of learner performance, progress and potentials are highly useful (Chassignol et al., 2018). Adaptive learning technologies trended across all of 2011 to 2021 in HR. If allowed to continue in its current trend, it is expected that assessments driven by analytics and AI will take a larger role in the coming years.

Ubiquitous learning technologies generated the third highest publication interest as of 2021, after weighing the effect of overall publication numbers. The weighted number of publications started from a low base of 107 in 2011 and ended the period under study by tripling to 323, overtaking learning objects. Ubiquitous learning technologies trended across all of 2011 to 2019 in HR. While there is resistance to the adoption of assessments tied to ubiquitous learning technologies, e.g. possibility of cheating owing to poor monitoring mechanisms, changes in socialization dimensions from group to individual, and cost of implementation (Sophonhiranrak, 2021; Asiimwe & Khan, 2013), most threats are surmountable and ubiquitous learning technologies offer many opportunities in enhancing the authenticity of digital education. The tripling of weighted publication numbers during the period under study is a testament to the growing interest in assessments tied to ubiquitous learning technologies. This technology group is an area that has much potential to explore in the coming years.

When weighted for the effect of overall publication numbers, it is observed that assessments tied to learning objects generated the lowest publication interest as of 2021. Correspondingly, learning objects were mentioned the least in HR – only in five of the 11 years under study. The higher mentions in HR from 2019 onwards, is in line with the growing research since 2019. These coincided with the effects of Covid-19 on broad-based education, including assessments. It is observed that learning objects only started receiving more mentions in recent years in HR, owing partly to pandemic-related need for assessments tied to more quality online courses (including micro-credential short courses), blended and hybrid course models, and open educational resources.

6. Conclusion

The findings provide a clear evolution of trend which can be used for future trend predictions for focus research and/or investment areas. Adaptive learning technologies, albeit being the second most highly published area in 2021, is highly likely to surpass immersive learning technologies as the number one research area in the near future. Further, having tripled in publications over the past 11 years and having overtaken learning object in research intensity, ubiquitous learning technologies showed promising potential as focused research and/or investment area.

Technology is growing at a rapid pace at multiple directions. Future research can look to provide a more detailed, mapped and streamlined guidance on how each technology area (or sub-areas) can influence assessment practices, with illustrations of such assessment practices for educational or research practitioners. It will also be useful to consider research similar to the Gartner Hype Cycle (Perez & Kreinovich, 2018) for emerging technologies and trend that may significantly affect assessment practices, to inform and prepare educators and researchers for potential changes coming in the years ahead.

References

Asghar, M. (2012). The Lived Experience of Formative Assessment Practice in a British University. *Journal of Further and Higher Education, 36*(2), 205–223.

Asiimwe, E. N., & Khan, S. Z. (2013). Ubiquitous computing in education: A SWOT analysis by students and teachers. In *12th World Conference on Mobile and Contextual Learning (mLearn 2013), 3*, 18. Hamad bin Khalifa University Press (HBKU Press).

Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial Intelligence trends in education: a narrative overview. *Procedia Computer Science, 136,* 16-24.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological forecasting and social change, 73*(8), 981-1012.

Dillenbourg, P. (2016). The evolution of research on digital education. *International Journal of Artificial Intelligence in Education*, *26*(2), 544-560.

EDUCAUSE. (n.d.). EDUCAUSE Horizon Reports. https://library.educause.edu/resources/2021/2/horizon-reports.

Engelbrecht, H., Lindeman, R. W., & Hoermann, S. (2019). A SWOT analysis of the field of virtual reality for firefighter training. *Frontiers in Robotics and AI, 6*, 101.

Gibbs, G. (1999). Using Assessment Strategically to Change the Way Students Learn. *Assessment Matters in Higher Education*, 41–54.

Huang, C., Yang, C., Wang, S., Wu, W., Su, J., & Liang, C. (2020). Evolution of topics in education research: a systematic review using bibliometric analysis. *Educational Review, 72*(3), 281-297.

Marín-Marín, J. A., López-Belmonte, J., Fernández-Campoy, J. M., & Romero-Rodríguez, J. M. (2019). Big data in education. A bibliometric review. *Social Sciences, 8*(8), 223.

Martin, S., Diaz, G., Sancristobal, E., Gil, R., Castro, M., & Peire, J. (2011). New technology trends in education: Seven years of forecasts and convergence. *Computers & Education, 57*(3), 1893-1906.

National Research Council. (1996). *National science education standards*. National Academies Press.

Postareff, L., Virtanen, V., Katajavuori, N., & Lindblom-Ylänne, S. (2012). Academics' Conceptions of Assessment and Their Assessment Practices. *Studies in Educational Evaluation, 38*(3-4), 84–92.

Perez, J., & Kreinovich, V. (2018). Gartner's Hype Cycle: A Simple Explanation.

Ross, J. (2017). Speculative method in digital education research. *Learning, Media and Technology*, *42*(2), 214-229.

Sophonhiranrak, S. (2021). Features, barriers, and influencing factors of mobile learning in higher education: A systematic review. *Heliyon, 7*(4), e06696.

Williamson, B. (2016). Digital education governance: data visualization, predictive analytics, and 'real-time' policy instruments. *Journal of education policy*, *31*(2), 123-141.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

The Comparative Effects of 'What if/what if not' and 'Main Ideas' Scaffolding for Online Student Question-generation on Science Learning

Fu-Yun YU* & Hsiu-Chen CHEN
Institute of Education, National Cheng Kung University, Taiwan
fuyun.ncku@gmail.com

Abstract: This study examined the effects of two scaffolding types with different levels of structures for student question-generation (SQG) on science learning. Specifically, the comparative effects of the structured 'what-if/what-if-not' scaffolding type and the semi-structured 'main ideas' scaffolding type on science academic achievement and SQG task performance were investigated by a non-equivalent quasi-experimental research method. Students from two sixth-grade classes (n = 56) participated and engaged in a weekly online SQG activity to support science learning for five weeks. An online system was adopted to support the integrated online SQG learning activity. Data analyzed by the multivariate analysis of covariance and analysis of covariance found that the participants in the 'what-if/what-if-not' scaffolding group performed significantly better than those in the 'main ideas' scaffolding group only in the flexibility dimension of student-generated questions. No significant differences were found between the two experimental groups in academic achievement and the other dimensions of student-generated questions task performance (including fluency, originality, and elaboration).

Keywords: Innovative learning and assessment approach, online learning activity, scaffolding, science learning, student-generated questions

1. Introduction

Student question-generation (SQG) is well recognized as an innovative approach to teaching, learning, and assessment (Yu & Wu, 2020). In practice, rather than resorting to teachers posing questions on the instructional material for the assessment of student learning, SQG entails engaging students in the process of question-generation on the study content they personally view as important, relevant, or interesting for self or peer assessment (Yu, 2019). By prompting self-reflection and meaning-making on the part of students, the inclusion and use of SQG in classrooms have been strongly recommended by researchers and practitioners.

Up till now, a wealth of empirical studies has generally substantiated the beneficial effects of SQG for the promotion of various cognitive and affective learning (Rosenshine, Meister, & Chapman, 1996; Rosli, Capraro, & Capraro, 2014; Zuya, 2017). Despite this, researchers have noted difficulties and challenges in the implementation of SQG in classrooms. In an attempt to alleviate SQG implementation hurdles, scaffolding of various types has been proposed. Among these, 'what if/what if not' proposed by Brown and Walter (2005) has been well received whereas 'main ideas' is among one of the most frequently adopted types by researchers (Rosenshine, Meister, & Chapman, 1996).

According to Stoyanova and Ellerton's SQG classification scheme (1996), 'what-if/what-if-not' is the structured SQG type, where students generate questions by the reformulation of given questions to help enhance understanding of the question structure and explore the relationship between questions and answers. 'Main ideas' would be the semi-structured type, where students construct a series of interrelated questions related to unfinished problem structures (e.g., based on a given set of keywords/main ideas) (Stoyanova & Ellerton, 1996). Despite that both 'what-if/what-if-not' and 'main ideas' aim to support students during the process SQG, and their facilitating effects on learning have been confirmed (Brown & Walter, 2005; Rosenshine, Meister, & Chapman, 1996), studies examined if and how students' learning could be affected by scaffolding with different levels of structures for SQG

were scarce. Additionally, for those that did, all adopted a less rigid research method (e.g., one-group experimental research method).

Given that 'what-if/what-if-not' and 'main ideas' as scaffolding for SQG are different in terms of levels of structures (Stoyanova & Ellerton, 1996), which should likely affect the focus, amount, and extent of attention of students, issues regarding if and how they affect learning are the focus of this study. Specifically, the comparative effects of 'what-if/what-if-not' and 'main ideas' scaffolding for online SQG on science academic achievement and task performance are examined.

2. Methods

2.1 The participants, research method, and experimental procedures

A non-equivalent quasi-experimental research design was adopted for this study, with two experimental treatment groups devised — the 'main ideas' online SQG group (i.e., Experimental I group), and the 'what-if/what-if-not' online SQG group (i.e., Experimental II group). Two sixth-grade classes (n=56) from the same elementary school in Tainan city, Taiwan taught by the same science teacher were randomly assigned to the two experimental groups. The SQG activity was introduced to support the teaching and learning of the science unit on 'Simple Machinery' with four subunits: leverages, axles, pulleys, gears, and power transmission.

This study consisted of two main stages and commenced right at the beginning of the spring semester. During Stage I, fundamental knowledge and skills on SQG and the operational procedures for navigating within the adopted online SQG system were introduced before students' hands-on practice session on the system. During this stage, the participants from both groups were directed to generate at least one question for each of the two chosen question types (i.e., short-answer and multiple-choice) on the last unit of the previous semester, and the signal words scaffolding type (i.e., who, what, where, when, what, and how), one of the easily learned scaffolding types for SQG, was introduced to support SQG. The participants' submitted SQG at this stage was analyzed and used as the covariant of SQG task performance whereas the participants' score at the end-of-semester science posttest administered school-wide was collected and analyzed as the covariant of academic achievement.

During Stage II, as a routine, following the instructional session, lab activity, and drill-and-practice sessions on each of the four subunits in the participating school's science lab, the participants reconvened in the school's computer lab during the weekly computer literacy class for the online SQG activity on the respective subunits. To equip the participants of the two experimental groups with the essential knowledge and skills on accessing the respective SQG scaffolding incorporated in the system (see next sub-section), a brief training session was arranged before the participants' first encounter with their respective SQG scaffolding types — the 'main ideas' for the Experimental I group and the 'what-if/what-if-not' for the Experimental II group. After the participants concludes their last online SQG activity on the fourth science subunit, they were tested on the covered content, and their performance at all SQG activity during this stage (i.e., a total of four activities) was assessed.

2.2 The online system

An online system developed under the supervision of the first author of this study (Yu, 2009) was adopted to support the integrated SQG activities. The system allows the content of different types of scaffolding to be dynamically adjusted by the implementing teacher according to the applied context. The space for the two chosen types of questions to be generated as well as examples illustrating how scaffolding for SQG works is depicted in Figures 1 and 2. As shown, for short-answer SQG, for successful submission, students need to complete the question-stem, the answer key, and annotation fields (the middle panel of Figure 1) whereas for multiple-choice SQG, they need to complete the question-stem, four options, the answer key, and annotation fields (the middle panel of Figure 2).

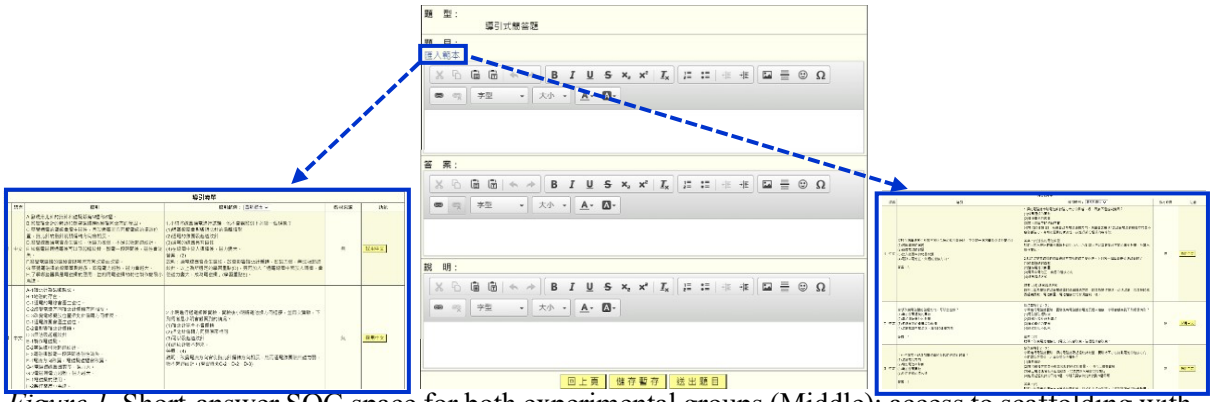


Figure 1. Short-answer SQG space for both experimental groups (Middle); access to scaffolding with examples from the prior unit during Stage I for the 'main ideas' scaffolding group (Left) and for 'what-if/what-if-not' scaffolding group (Right)

To access the scaffolding incorporated in the system for SQG reference, the participants simply click on the hyperlink placed on top of the question-stem field. A pop-up window with either a list of main ideas covered in the current study content (the left panel of Figures 1 and 2) or a list of questions they practiced during the drill-and-practice session developed by the implementing teacher (the right panel of Figures 1 and 2) can be viewed by the participants of the 'main ideas' and what-if/what-if-not' groups, respectively.

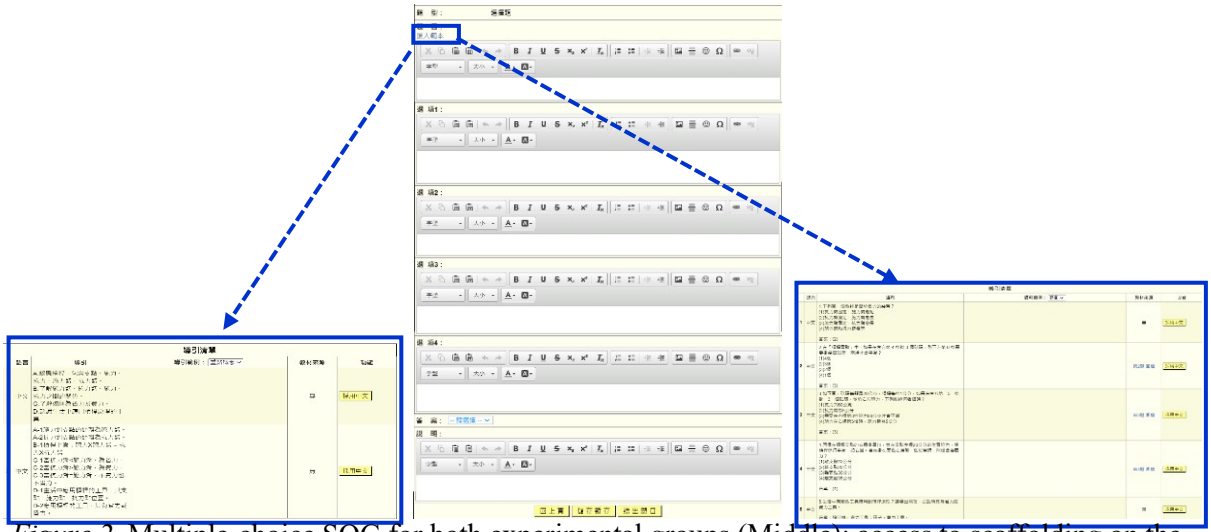


Figure 2. Multiple-choice SQG for both experimental groups (Middle); access to scaffolding on the current subunit during Stage II for the 'main ideas' scaffolding group (Left) and for 'what-if/what-if-not' scaffolding group (Right)

2.3 Measurement instruments

In light of the nature of the science topic, two-tier questions highlighting misconceptions frequently held by students on 'Simple Machinery' (i.e., posttest section I) in addition to the most frequently used question types in classrooms (i.e., posttest section II) were adopted at the posttest. Item analysis procedures with item difficulty and discrimination were conducted for both science academic achievement pretest and posttest to ensure their technical quality. The results evidenced satisfactory item difficulty and discrimination for both tests.

The quality of student-generated questions was assessed according to a set of criteria frequently used for the assessment of SQG performance. Explicitly, all questions the participants generated during Stages I and II of the study (i.e., a total of five SQG activities) were assessed along the four dimensions: fluency, flexibility, originality, and elaboration (Yu & Wu, 2016). Each of the dimensions was operationalized defined. Inter-rater reliability was established.

3. Results and Discussion

3.1 Results

The means and standard deviations of the two experimental groups on the examined variables are listed in Table 1. As a significant correlation was detected between posttest sections I and II, academic achievement was analyzed by the multivariate analysis of covariance technique. The quality of questions (on the four dimensions) was analyzed by the analysis of covariance when assumptions associated with the adopted data analysis methods were satisfied.

The results of data analysis indicated no significant differences between the two experimental groups in academic achievement, Wilks'λ= 0.99, $F(2,52) = 0.25$, $p = .78$, and the fluency, $F(1,53) = 0.12$, $p = .73$, originality, $F(1,53) = 0.07$, $p = .79$, and elaboration, $F(1,53) = 0.42$, $p = .52$ of SQG task performance. Only statistically significant differences in the flexibility dimension of SQG performance were found between the two experimental groups, $F(1,53) = 32.01$, $p = .00$.

Table 1. *Descriptive Statistics of the Two Treatment Groups on the Examined Variables*

Groups (*N*)	Examined variables	*M*	*SD*	Adjusted M
Main ideas (28)	Academic achievement — two-tier test	13.71	1.06	14.75
	posttest	59.64	15.23	62.26
	SQG performance — Fluence	24.93	10.43	25.01
	Flexibility	0.86	1.27	0.85
	Originality	2.61	2.36	2.53
	Elaboration	8.61	4.73	10.04
What-if/what-if (28)	Academic achievement — two-tier test	16.57	1.07	15.54
	posttest	65.79	13.84	63.17
	SQG performance — Fluence	24.25	7.41	24.17
	Flexibility	3.43	1.95	3.44
	Originality	2.29	2.89	2.36
	Elaboration	12.46	6.66	11.03

3.2 Discussion and Conclusion

Attending to and addressing students' misconceptions through innovative teaching strategies are emphasized in science education (Gomez-Zwiep, 2008). As noted by researchers, misconceptions are frequently observed in the teaching and learning of 'Simple Machinery'. In light of the facilitating effects of SQG for the promotion of a deeper understanding of the study content (Yu, 2009), and the overall positive effects of scaffolding on learning in general (Kim, Belland, & Walker, 2018) and SQG task performance in specific (e.g., Yu & Pan, 2014; Yu, Tsai, & Wu, 2013), the comparative effects of two scaffolding effects with different levels of structures, namely 'what-if/what-if-not' and 'main ideas' for SQG, for the support of student learning of 'Simple Machinery' were examined in this study.

Based on the results of data analysis from this study, the more structured 'what-if/what-if-not' scaffolding type did not exert differential learning effects on science learning and most of the indexes of SQG task performance as compared to the 'main ideas' scaffolding type. In fact, the participants exposed to the 'what-if/what-if-not' scaffolding were found to only perform better than the 'main ideas' scaffolding in the flexibility dimension of SQG task performance. In other words, the participants in the 'what-if/what-if-not' SQG group generated significantly more questions containing interlinks among related concepts (either within the current study subunit or to prior subunits) than those in the 'main ideas' group.

In light of the scarcity of studies examining this research question (i.e., if and how students' learning could be affected by scaffolding with different levels of structures for SQG), more studies are

called for. Also, researchers adopting different research method (e.g., qualitative research via in-depth interviews, and non-participant observation, etc.) to explore and understand the underlying mechanism resulting in any observed differences between the two scaffolding types for SQG would be a fruitful future research direction.

Acknowledgment

This study was supported by the Ministry of Science and Technology in Taiwan (MOST 108-2511-H-006-007-MY3).

References

Brown, S. I., & Walter, M. I. (2005). *The Art of Problem Posing* (3rd ed.). Lawrence Erlbaum Associates.

Gomez-Zwiep, S. (2008). Elementary teachers' understanding of students' science misconceptions: Implications for practice and teacher education. *Journal of Science Teacher Education, 19*(5), 437-454.

Kim, N. J., Belland, B. R., & Walker, A. E. (2018). Effectiveness of computer-based scaffolding in the context of problem-based learning for stem education: Bayesian meta-analysis. *Educational Psychological Review, 30*, 397–429.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of educational research, 66*(2), 181-221.

Rosli, R., Capraro, M. M., & Capraro, R. M. (2014). The effects of problem posing on student mathematical learning: A meta-analysis. *International Education Studies, 7*(13), 227-241.

Stoyanova, E., & Ellerton, N. F. (1996). A framework for research into students' problem posing in school mathematics. In P. C. Clarkson (Ed.), *Technology in Mathematics Education* (pp. 518-525). Mathematics Education Research Group of Australasia.

Yu, F. Y. (2019). The learning potential of online student-constructed tests with citing peer-generated questions. *Interactive Learning Environments, 27*(2), 226-241.

Yu, F. Y. (2009). Scaffolding student-generated questions: Design and development of a customizable online learning system. *Computers in Human Behavior, 25*(5), 1129-1138.

Yu, F. Y. & Pan, K-J (2014). Effects of student question-generation with online prompts on learning. *Educational Technology and Society, 17*(3), 267-279.

Yu, F. Y., Tsai, H. C. & Wu, H. L. (2013). Effects of online procedural scaffolds and the timing of scaffolding provision on elementary Taiwanese students' question-generation in a science class. *Australasian Journal of Educational Technology, 29*(3), 416-433.

Yu, F. Y. & Wu, C. P. (2016). Predictive effects of the quality of online peer-feedback provided and received on primary school students' quality of question-generation. *Educational Technology & Society, 19*(3), 234-246.

Yu, F. Y. & Wu, W. S. (2020). Effects of student-generated feedback corresponding to answers to online student-generated questions on learning: What, why, and how? *Computers & Education, 145*, 103723.

Zuya, H. E. (2017). The benefits of problem posing in the learning of mathematics: A systematic review. *International Journal of Advanced Research, 5*(3), 853-860.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

When the School Door Closes, Do Teachers Open a Window? Using Mixed Methods to Investigate Teachers' Online Teaching Practices and Momentary Experiences in Crisis

Jo-Chi HSIAO[ab], Chao-Yang CHENG[c] & Sunny S. J. LIN[d*]
[a]*Institute of Education, National Yang Ming Chiao Tung University, Taiwan*
[b]*National Hsinchu Senior High School, Taiwan*
[c]*Department of E-learning Design and Management, National Chiayi University, Taiwan*
[d]*Institute of Education, National Yang Ming Chiao Tung University, Taiwan*
* sunnylin@nycu.edu.tw

Abstract: The crisis of COVID-19 significantly disrupted global education and overturned our understanding of classroom setting in secondary education. This study adopted a mixed methods design to categorize the daily instructional practices of teachers in the early stages of school closure using content analysis and to test the resulting framework using ANOVA. A total of 48 teachers provided five-day diary entries in which they recorded episodic activities and subjective experiences, reported daily teaching techniques, and rated their subjective teaching quality. Four emergent online teaching orientations emerged from the data: 1 nonreciprocal knowledge understanding (1NRKU), 2 nonreciprocal knowledge construction (2NRKC), 3 reciprocal knowledge understanding (3RKU), and 4 reciprocal knowledge construction (4RKC). The results also showed that teachers adopting different online teaching orientations described their experiences differently. Teachers adopting 2NRKC, 3RKU, and 4RKC experienced significantly higher daily teaching satisfaction; teachers adopting 3RKU and 4RKC perceived significantly better student engagement. In terms of teaching readiness, teachers adopting 2NRKC, 3RKU, and 4RKC had significantly more experiences of online professional development and significantly better distance teaching TPCK. This study is innovative in presenting momentary data, which sheds light on future programs for customized professional development that will help prepare teachers for the next educational crisis.

Keywords: COVID-19 pandemic, online teaching orientations, mixed methods

1. Introduction

In a recent review study of the impact of the COVID-19 pandemic on teaching and learning worldwide, Pokhrel and Chhetri (2021) summarized the challenges and opportunities that teachers and learners faced. They concluded that despite the increase in pandemic-relevant studies, further exploration should be carried out to investigate what types of effective online pedagogies have been adopted to keep learning sustained. A further investigation into studies focusing on teachers' first-hand experiences and perceptions of online teaching under the threat of COVID-19 reveals that these direct reports have significant limitations. First, they were insufficiently grounded in existing theories or frameworks, so meaningful insights could not be drawn from previous endeavors on distance teaching or relevant issues to explain and interpret the apparent chaos. Second, they mainly collected cross-sectional data via open-ended surveys, interviews, questionnaires, or mixed-methods design (Alea et al., 2020; Baker et al., 2021). These data were sufficient for a preliminary investigation with quick summaries of the emergent situations but failed to zoom in to rich details that witnessed emergency remote teaching (Whittle et al., 2020) in a more complex way.

The purpose of this study is twofold. First, it aims to categorize teachers' pedagogy for five teaching days in the initial phase of school closure. Second, it intends to collect cross-sectional and repeated data to reveal teachers' experiences of the sudden and complete shift of instruction from the traditional face-to-face environment to the virtual world. In other words, we want to explore, when school doors were closed during the pandemic outbreak, did teachers effectively open a window for

students to access learning remotely? If so, what kinds of windows did they open? What did they feel about their teaching and student learning at the critical moments when they opened the window-a pedagogy for which they had received little training

2. Theoretical Background

2.1. Emergency Remote Teaching and Online Teaching Readiness

Emergency remote teaching is proposed by Whittle et al. (2020). They introduced a model to highlight the uniqueness of unplanned and responsive remote teaching resulting from emergent crises. They suggested including teachers with a broader range of teaching experiences in the study. A longitudinal study that could reveal changes in teachers should be carried out. The present study addresses these suggestions by collecting repeated data from both novice and experienced teachers.

Among the various TPCK questionnaires that have been developed, the one proposed by Chai et al. (2013) caught our attention for a few reasons. First, the instrument was validated with an Asian sample of 550 pre-service teachers in China, Hong Kong, Singapore, and Taiwan, so the items were presented in Chinese (p.45-48), the first language the participants knew. Besides, the instrument has been adopted to measure TPCK of in-service teachers (Dong et al., 2015), who are identical to the participants. So far, few previous studies have viewed TPCK in the unique context of complete online distance teaching, so it is worth exploring its role in supporting teachers to do emergency remote teaching. Since online teaching requires greater technology-related knowledge, we decided to narrow down and focus on teachers' holistic technology-centered knowledge of TK (technology knowledge), TCK (technological content knowledge), TPK (technological pedagogical knowledge), and TPCK (technological pedagogical content knowledge). We coined "distance teaching TPCK" to describe this context-specific construct.

2.2. Subjective Teaching Quality: Teachers' Optimal experience, Teaching Satisfaction, and Perceived Student Engagement

The sudden shift to virtual learning environment caused by the COVID-19 pandemic ushers in global concerns about online teaching quality (United Nations, 2021; Seufert et al., 2021). This study focuses on teachers' subjective evaluations of teaching quality, including their perceptions of the momentary experience, teaching, and students' learning. When it comes to perceptions of experience, Csikszentmihalyi's (1990) flow theory has attracted much attention. This theory explains how deep engagement occurs when people work on challenging but not overwhelming activities. Csikszentmihalyi (1997) later noted that if teachers are in flow, students tend to experience flow (p. 33). This contagious effect has been empirically demonstrated (Bakker, 2005; Culbertson et al., 2015), but direct investigation on teachers' optimal experience in the classroom has been limited (Basom & Frase, 2004).

Beard and Hoy (2010) were one of the few studies to provide empirical evidence of teachers' optimal experience. They viewed teachers' flow experience as a holistic integration of the nine dimensions proposed by Nakamura and Csikszentmihalyi (2009): three antecedent conditions (i.e., challenge-skill balance, clear goals, and immediate feedback) and six elements of a subjective state (i.e., concentration on the task, action-awareness merging, loss of self-consciousness, sense of control, time distortion, the activity as intrinsically rewarding).

In addition to teachers' optimal experience, faculty satisfaction has also been used to evaluate the quality of online courses. It has been shown to be positively correlated with student satisfaction (Bolliger & Wasilik, 2009; Hartman et al., 2000). Interestingly, when Bolliger and Wasilik (2009) conducted an online faculty satisfaction survey, about one-third of the 28 items were related to teachers' perception of student engagement in the classroom.

Another common criterion is student engagement. Martin and Bolliger (2018) found that how teachers selected materials, presented knowledge, and employed pedagogical strategies was almost determinant of how engaged students were in class. Besides, online learners reported a higher mean score of engagement in learner-to-instructor strategies than in learner-to-content and learner-to-learner

strategies. These findings are not consistent with the longitudinal study by Muir et al. (2019). In other words, pedagogies and student engagement are two sides of one coin. Daily variation of the instructions leads to fluctuation of student engagement over time.

2.4. Current Study

The following research questions are intentionally asked:

- RQ1: What emergent online teaching orientations could be found in teachers' teaching logs kept during the initial stage of emergency remote teaching?
- RQ2: Did teachers applying different emergent online teaching orientations report significant differences in subjective teaching quality and online teaching readiness?

We collected nested data that fell into three types (see Figure 1): (1) person-based variables, which included online teaching readiness (i.e., online professional development experience and distance teaching TPCK); (2) daily teaching variables, which were emergent online teaching orientations and two variables of subjective teaching quality (i.e., teaching satisfaction and perceived student engagement); (3) the episodic variable, which was the last variable of subjective teaching quality (i.e., the optimal experience of teaching-related activities).

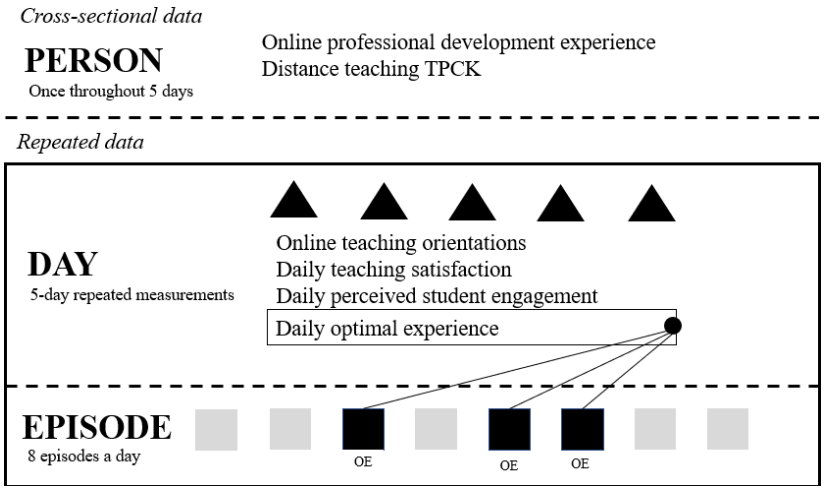


Figure 1. Three-level data structure of the current study.

Note. At the episode level, OE refers to Optimal experience. Black and grey cubes represent eight daily activities teachers reported per day. Black cubes represent teaching activities, which are the main source of this study. The lines and the framed square show episodic data are aggregated to form daily optimal experience at the Day level. Gray cubes refer to nonteaching activities (e.g., feeding children, watching movies with family, etc.), excluded from data analysis. At the day level, triangles represent five teaching days.

3. Method

3.1 Participants

Right after the government in Taiwan announced the closure of schools at the end of May 2021, an online invitation was sent to high school teachers. Initially, 50 teachers were recruited as participants, but two of them were excluded because they could not contribute a complete record of five working days. The majority of the 48 recruited teachers were female (73%) and formal teachers (88%), whose average teaching experience was 12.5 years (ranged from 1.5 to 30 years). Half of them (50%) had children staying at home with them while they were working at home.

3.2 Instruments

Three instruments were used in this study. The first instrument, Daily Diary, documenting participants' eight episodes of daily activities throughout the day, was adapted from Cheng et al. (2017). A total of

1,984 were collected from late May to early June 2021. This instrument consists of three parts. Part one is diary writing whose purpose was to help participants recall details of what they had done on the previous day from their memory (Kahneman et al., 2004). In part two, they divided the content written in part one into eight serial episodes. They had to give each episode a descriptive tag, such as "Teaching," "Preparing for the class," "Doing exercise," etc. Only episodes tagged "Teaching" were analyzed (1,092 events in total, 55%). In part three, participants answered structured questions about each episode. These questions can be divided into two types. The first type captures the characteristics of each episode, including *When* the episode began and ended, *What* they were doing, *Where* they were, *Whom* they were with, and *How* they perceived the episode. The other type measures participants' subjective experience, defined by nine indicators of optimal experience in Flow theory (Cheng et al., 2017) on a nine-point Likert scale.

The second instrument, Teaching Log, is a structured daily questionnaire consisting of two parts. In part one, participants described their online teaching practices by checking boxes from six categories (i.e., delivery types, teaching materials, platforms, websites, tools, and procedure of the class). They could provide additional text descriptions if necessary. Following Saldaña's (2015) manual, two coders coded the data into four types of measures: Delivery Types, Innovative Adds-On, Teaching Processes, and Interactive Activities. Delivery Types refer to how teachers delivered the class. Participants could choose between synchronous, asynchronous, and mixed modes. Innovative Adds-On indicates whether participants included additional online materials in addition to the digitized textbooks. This measure was coded from boxes regarding participants' teaching materials (e.g., self-made videos, worksheets other teachers designed, free interactive applications, etc.). Teaching Processes refer to the activity's agency (who initiates the behavior). This measure was coded into three categories (i.e., teacher-centered, student-centered understanding focused, and student-centered construction focused) from boxes concerning teaching activities exerted in class (e.g., teacher lecturing, teacher demonstration, group discussion, etc.). Interactive Activities focus on whether participants emphasized classroom interaction. This measure was coded from boxes about online learning software (e.g., Kami, Class Dojo, Kahoot, etc.) into three purposes: quizzes, games, and discussion. The coding of the data is further explained in Section 3.2.1. In part two, participants answered two questions on a 9-point Likert scale, one on teaching satisfaction ("Overall, were you satisfied with today's teaching?") and the other on perceived student engagement ("Overall, were your students engaged in today's learning?"). These two items were indicators of subjective teaching quality.

In the third instrument, Personal Profile, participants answered questions relevant to personal information and online teaching readiness (i.e., online professional development experience and distance-teaching TPCK). Online professional development experience refers to whether participants had experience with online instruction and participation in online learning activities prior to the COVID -19 pandemic. As for distance-teaching TPCK, it was measured by four subscales focusing on technology developed by Chai et al. (2013).

3.2.1 Coding Process for Emergent Online Teaching Orientations

Data collected by Teaching Log were initially coded into four measures (i.e., Delivery Types, Innovative Adds-On, Teaching Processes, and Interactive Activities). Based on a holistic evaluation of the measures, we developed a framework to categorize teaching days (237 teaching days reported by 48 teachers for five days; missing data were because some participants did not have classes on certain days) into four emergent online teaching orientations (see Figure 2). Two e-learning experts, following Saldaña's (2015) manual, conducted the coding. To test the reliability of their judgments, we relied on Cohen's (1960) Kappa coefficient κ as an inter-rater agreement and Landis and Koch's (1977) standard to explain the meaning of the Kappa value. The Kappa coefficient was 0.914, indicating near perfect agreement between raters (McHugh, 2012). See the Results section for a detailed explanation of the framework.

3.2.2 Statistical Analyses

We used SPSS 21 software for statistical analyses. First, Chi-square tests were applied to determine whether four types of measures (i.e., Delivery Types, Innovative Adds-On, Teaching Processes, and Interactive Activities) showed significant statistical differences among the four emergent online

teaching orientations. The results were used to answer RQ1. Second, to answer RQ2, we conducted one-way analyses of variance (ANOVA) with LSD post hoc tests to explore whether the types of emergent online teaching orientations were associated with participants' subjective teaching quality and with their online teaching readiness. It is important to note that daily optimal experience was aggregated from episodic optimal experiences (Figure 1).

4. Results

4.1. RQ1: What emergent online teaching orientations could be found in teachers' teaching logs kept during the initial stage of emergency remote teaching?

Based on the coded data from Teaching Log, we found a quadrant diagram for the emergent online teaching orientations (see Figure 2). Each emergent online teaching orientation (i.e., 1 nonreciprocal knowledge understanding (NRKU); 2 nonreciprocal knowledge construction (NRKC); 3 reciprocal knowledge understanding (RKU); and 4 reciprocal knowledge construction (RKC)) is followed by explanations and typical examples.

To determine whether the four measures of the teaching practices (e.g., *Delivery types*, *Innovative adds-on*, *Teaching processes*, and *Interactive activities*) were equally distributed in the four emergent online teaching orientations, we performed four chi-square tests of goodness-of-fit. As Table 1 shows, the four measures were not equally distributed across the emergent online teaching orientations ($\chi^2_{(6,237)} = 35.26$, $p < .001$; $\chi^2_{(3,74)} = 40.38$, $p < .001$; $\chi^2_{(6,1590)} = 183.81$, $p < .001$; $\chi^2_{(6,231)} = 61.17$, $p < .001$).

1NRKU (nonreciprocal knowledge understanding) was similar to face-to-face classroom setting where teachers aim to deliver knowledge and have limited interaction with students. As shown in Table 1, teachers adopting 1NRKU tended to use more synchronistic delivery type (64.5%) and some asynchronistic delivery type (25.8%). They used very few innovative adds-on materials (3.2%) but many teacher-centered activities (90.6%) and interactive quizzes (97%).

2NRKC (nonreciprocal knowledge construction) was featured by teachers providing videos and materials, asking students to engage in online self-directed learning activities where students worked alone to construct knowledge with limited interaction with the teacher (see Figure 2). As Table 1 shows, 2NRKC teachers chose more synchronistic delivery type (66.7%) and did not use many innovative adds-on materials (10.3%). They used similar amount of teacher-centered and student-centered activities (44.9% versus 55.1%) and many interactive quizzes (88.7%).

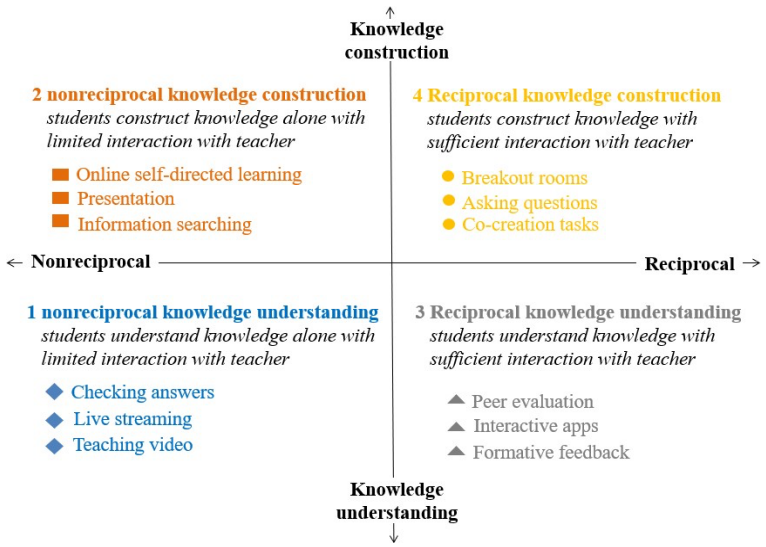


Figure 2. Quadrant Diagram of Four Emergent Online Teaching Orientations.

As shown in Figure 2, 3RKU (reciprocal knowledge understanding) was featured by highly interactive activities where students understand knowledge with immediate support or feedback from the teacher. As Table 1 shows, teachers adopting 1NRKU used similar amount of synchronistic and mixed delivery types (52.5% versus 45.9%). More than half of their teaching materials were innovative adds-on (55.2%), and their class was under their control (teacher-centered = 60.4%). They assigned many interactive quizzes (75.5%), but they also tried interactive games (15.8%) and discussion (9.2%).

4RKC (reciprocal knowledge construction) was featured by co-creation activities where students constructed knowledge with sufficient guidance from the teacher (see Figure 2). As shown in Table 1, teachers adopting 4RKC used more synchronistic and mixed delivery types (69.4% versus 25%). They incorporated many innovative adds-on materials (88.9%) and designed similar amount of teacher-centered and student-centered activities (46.6% versus 53.4%). In their class, interactive quizzes, interactive games, and discussions were commonly seen (37.7%, 26.1%, and 36.2%).

Table 1. *Chi-Square Tests of Percentage Distribution for Measures of Teaching Practices across Four Emergent Online Teaching Orientations*

Emergent online teaching orientations	**1NRKU** *DTA* [a] = 62 Count (%)	**2NRKC** *DTA* = 78 Count (%)	**3RKU** *DTA* = 61 Count (%)	**4RKC** *DTA* = 36 Count (%)	χ^2 *test*
Single response [b]					
Delivery types					$\chi^{22}_{(6,237)} = 35.26$***
Synchronistic	40(64.5)	52(66.7)	32(52.5)	25(69.4)	
Asynchronistic	16(25.8)	11(14.1)	1(1.6)	2(5.6)	
Mixed	6(9.7)	15(19.2)	28(45.9)	9(25.0)	
Innovative adds-on	2(3.2)	8(10.3)	32(52.5)	32(88.9)	$\chi^2_{(3,74)} = 40.38$***
Multiple response [b]					
Teaching Processes					$\chi^2_{(6,1590)} = 183.81$***
Teacher-centered	261(90.6)	264(44.9)	264(60.4)	129(46.6)	
Student-centered UF [c]	10(3.5)	116(19.7)	66(15.1)	49(17.7)	
Student-centered CF [d]	17(5.9)	208(35.4)	107(24.5)	99(35.7)	
Interactive activities					$\chi^2_{(6,231)} = 61.17$***
Quizzes	32(97.0)	47(88.7)	57(75.0)	26(37.7)	
Games	1(3.0)	5(9.4)	12(15.8)	18(26.1)	
Discussion	0(0.0)	1(1.9)	7(9.2)	25(36.2)	

Note. ***$p < .001$. 1NRKU = nonreciprocal knowledge understanding; 2NRKC = nonreciprocal knowledge construction; 3RKU = reciprocal knowledge understanding; and 4RKC = reciprocal knowledge construction. DTA [a] = Daily teaching activities; the number of the daily teaching activities that belong to a particular emergent online teaching orientation. 48 teachers report 237 daily teaching activities for five days (with missing). Single response [b]: Participants chose one option among choices. Multiple response b: Participants chose more than one option among choices. UF [c] = Understanding-focused. CF [d] = Construction-focused.

4.1 RQ2: Did teachers applying different emergent online teaching orientations report significant differences in subjective teaching quality and online teaching readiness?

The means and standard deviation of subjective teaching quality (i.e., daily optimal experience, daily teaching satisfaction, and daily perceived student engagement) and online teaching readiness (i.e., online professional development and distance teaching TPCK) according to emergent online teaching orientations are summarized in Figure 3. This figure also displays the results of one-way ANOVA that compares the effects of emergent online teaching orientations on subjective teaching quality and online teaching readiness.

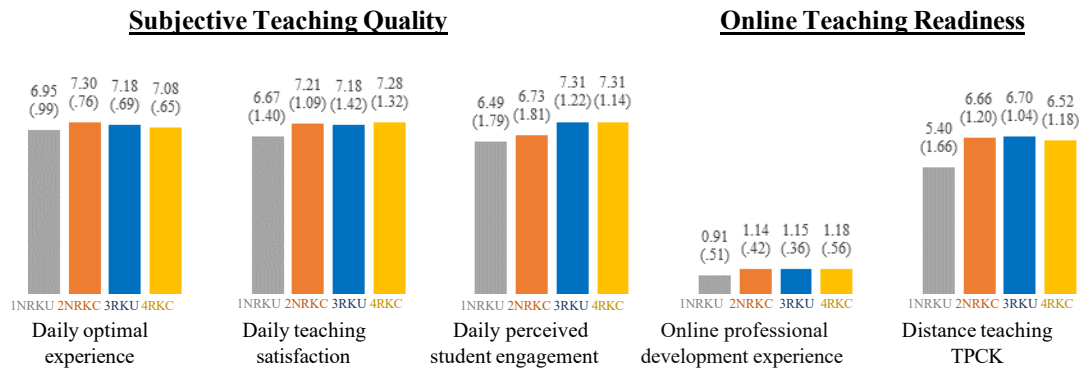
As shown in Figure 3, among the four emergent online teaching orientations, no significant difference was found in daily optimal experience ($F_{(3,232)} = 2.36$, $p = .072$). On the contrary, there was a statistically significant difference in daily teaching satisfaction ($F_{(3,232)} = 2.65$, $p = .05$, $\eta_p^2 = .03$).

According to Fisher's Least Significant Difference Test for multiple comparisons, in terms of the mean value of daily teaching satisfaction, 1NRKU was significantly lower than 2NRKC (p = .017, 95% *C.I.* = [-.97, -.10]), 3RKU (p = .032, 95% *C.I.* = [-.97, -.05]), and 4RKC (p = .027, 95% *C.I.* = [-1.14, -.07]).

Statistically significant difference was also found in daily perceived student engagement ($F_{(3,232)}$ = 3.82, p = .011, η_p^2 = .05). Results of Fisher's Least Significant Difference Test for multiple comparisons showed 1NRKU was significantly lower than 3RKU (p = .005, 95% *C.I.* = [-1.38, -.26]), and 4RKC (p = .015, 95% *C.I.* = [-1.47, -.16]). Besides, 2NRKU was significantly lower than 3RKU (p = .033, 95% *C.I.* = [-1.11, -.05]).

Another statistically significant difference was revealed in online professional development experience ($F_{(3,232)}$ = 4.41, p = .005, η_p^2 = .05). Post-hoc analyses using Fisher's Least Significant Difference Test for multiple comparisons indicated that the mean value of 1NRKU was significantly lower than 2NRKC (p = .003, 95% *C.I.* = [-.39, -.08]), 3RKU (p = .004, 95% *C.I.* = [-.40, -.08]), and 4RKC (p = .005, 95% *C.I.* = [-.46, -.08]).

The final statistically significant difference was found in distance teaching TPCK ($F_{(3,232)}$ = 13.97, p < .001, η_p^2 = .15). Reported by Post-hoc test of Fisher's Least Significant Difference Test for multiple comparisons, the mean value of 1NRKU was significantly lower than 2NRKC (p < .001, 95% *C.I.* = [-1.69, -.82]), 3RKU (p < .001, 95% *C.I.* = [-1.76, -.83]), and 4RKC (p < .001, 95% *C.I.* = [-1.66, -.59]).



Items	$F_{(3, 232)}$	Post-Hoc Test
Subjective teaching quality		
Daily optimal experience	*2.36*	-
Daily teaching satisfaction	*2.65*$^{\#}$	2NRKC, 3RKU, 4RKC > 1NRKU
Daily perceived student engageme	*3.82**	3RKU, 4RKC > 1NRKU; 3RKU > 2NRKC
Online teaching readiness		
Online professional development experience	*4.41***	2NRKC, 3RKU, 4RKC > 1NRKU
Distance teaching TPCK	*13.97****	2NRKC, 3RKU, 4RKC > 1NRKU

Figure 3. ANOVA Examinations on Differences of Subjective Teaching Quality and Online Teaching Readiness across Four Emergent Online Teaching Orientations.

Note. ***p < .001, **p < .01, *p < .05, $^{\#}p$ = .05. 1NRKU = nonreciprocal knowledge understanding (N = 61); 2NRKC = nonreciprocal knowledge construction (N = 78); 3RKU = reciprocal knowledge understanding (N = 61); 4RKC = reciprocal knowledge construction (N = 36).

5. Discussion and Conclusion

The most notable contribution of the present study is to identify four emergent online teaching orientations (Figure 2) and to examine whether significant differences existed among the four orientations in terms of subjective teaching quality and online teaching readiness. When schools were suddenly closed during the COVID-19 pandemic, teachers inevitably experienced anxiety and psychological stress (Chou & Chou, 2021; Kupers et al., 2022). Under pressure, teachers probably

could not focus on interpersonal interaction and knowledge construction at the same time. As long as teachers do not apply 1NRKU, they would feel satisfied with their emergency remote teaching. However, if they intend to increase their satisfaction of perceived student engagement, they should address social presence, as teachers applying 3RKU and 4RKC were found to have significantly higher perceived student engagement than those using 1NRKU and 2NRKC. The emphasis on social interaction has been well documented in previous studies (Martin & Bolliger, 2018; Muir et al., 2019). Based on the findings of the present study, 3RKU was proposed as an effective emergent online teaching orientation to achieve teaching satisfaction and perceived student engagement. Teachers still dominate the classroom to impart knowledge, but interactivity is emphasized through innovative adds-on, interactive games, and discussion.

Two implications can be drawn from this study. Theoretically, this study directly examined teachers' flow/optimal experience, which adds to the empirical findings on flow theory (Csikszentmihalyi, 1990). Future studies could compare teachers' flow/optimal experience across different types of episodes. In practice, teachers are encouraged to engage in sustained professional learning networks (Hsiao & Lin, 2022) in which they can develop their distance teaching TPCK and adapt their ongoing learning so that they can acquire and refine their competence in designing various learning activities in the virtual classroom.

Acknowledgments

This work was part of the first author's unpublished dissertation. It was supported by grants of Ministry of Science and Technology Taiwan to the second author (grant numbers: 109-2511-H-415-010-MY2) and the corresponding author (grant numbers: 108-2511-H-009-008-MY3). The content of this paper was part of the manuscript submitted to the journal Computers and Education.

References

Alea, L. A., Fabrea, M. F., Roldan, R. D. A., & Farooqi, A. Z. (2020). Teachers' Covid-19 awareness, distance learning education experiences and perceptions towards institutional readiness and challenges. *International Journal of Learning, Teaching and Educational Research, 19*(6), 127-144.

Baker, C. N., Peele, H., Daniels, M., Saybe, M., Whalen, K., Overstreet, S., & The New Orleans, T. I. S. L. C. (2021). The experience of COVID-19 and its impact on teachers' mental health, coping, and teaching. *School Psychology Review, 50*(4), 491-504.

Basom, M. R., & Frase, L. (2004). Creating optimal work environments: Exploring teacher flow experiences. *Mentoring & Tutoring: Partnership in Learning*, *12*(2), 241-258.

Bragg, L. A., Walsh, C., & Heyeres, M. (2021). Successful design and delivery of online professional development for teachers: A systematic review of the literature. *Computers & Education*, *166*, 104158.

Beard, K. S., & Hoy, W. K. (2010). The nature, meaning, and measure of teacher flow in elementary schools: A test of rival hypotheses. *Educational Administration Quarterly*, *46*(3), 426-458.

Bolliger, D. U., & Wasilik, O. (2009). Factors influencing faculty satisfaction with online teaching and learning in higher education. *Distance Education, 30*(1), 103-116.

Chai, C. S., Ng, E. M., Li, W., Hong, H. Y., & Koh, J. H. (2013). Validating and modelling technological pedagogical content knowledge framework among Asian preservice teachers. *Australasian Journal of Educational Technology*, *29*(1), 41-53.

Cheng, C. Y., Chen, S. Y., & Lin, S. S. J. (2017). Episodic and individual effects of elementary students' optimal experience: An HLM study. *The Journal of Educational Research*, *110*(6), 653-664.

Chou, H. L., & Chou, C. (2021). A multigroup analysis of factors underlying teachers' technostress and their continuance intention toward online teaching. *Computers & Education*, *175*, 104335.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience.* New York: Harper.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement with everyday life.* New York, NY: HarperCollins.

Dong, Y., Chai, C. S., Sang, G.-Y., Koh, H. L., & Tsai, C.-C. (2015). Exploring the profiles and interplays of pre-service and inservice teachers' Technological Pedagogical Content Knowledge (TPACK) in China. *Educational Technology & Society, 18*(1), 158–169.

Hartman, J., Dziuban, C., & Moskal, P. (2000). Faculty satisfaction in ALNs: A dependent or independent variable? *Journal of Asynchronous Learning Networks, 4*(3), 155-177.
Hsiao, J. C., & Lin, S. S. J. (2022). How energy maintains social sustainability of teachers' learning communities: New insights from a blended professional learning network. *Sustainability, 14*(6), 1-21.
Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A Survey Method for Characterizing Daily Life Experience: The Day Reconstruction Method. *Science*, *306*(5702), 1776-1780.
Koehler, M. J., & Mishra, P. (2009). What is technological pedagogical content knowledge? *Contemporary Issues in Technology and Teacher Education, 9*(1), 60-70.
Kupers, E., Mouw, J. M., & Fokkens-Bruinsma, M. (2022). Teaching in times of COVID-19: A mixed-method study into teachers' teaching practices, psychological needs, stress, and well-being. *Teaching and Teacher Education,115*, 103724.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.
Martin, F., & Bolliger, D.U. (2018). Engagement matters: Student perceptions on the importance of engagement strategies in the online learning environment. *Online Learning 22*(1), 205-222.
McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica*, *22*(3), 276-282.
Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017-1054.
Muir, T., Milthorpe, N., Stone, C., Dyment, J., Freeman, E., & Hopwood, B. (2019). Chronicling engagement: Students' experience of online learning over time. *Distance Education, 40*(2), 262-277.
Nakamura, J., & Csikszentmihalyi, M. (2009). Flow theory and research. *Handbook of Positive Psychology*, 195-206.
Pokhrel, S., & Chhetri, R. (2021). A literature review on impact of COVID-19 pandemic on teaching and learning. *Higher Education for the Future, 8*(1), 133-141.
Saldaña, J. (2015). *The coding manual for qualitative researchers.* Los Angeles, CA: Sage.
Seufert, S., Guggemos, J., & Sailer, M. (2021). Technology-related knowledge, skills, and attitudes of pre-and in-service teachers: The current situation and emerging trends. *Computers in Human Behavior*, *115*, 106552.
United Nations (2021, October 4). Teachers are driving force behind 'global education recovery' from COVID-19. *UN News, Global Perspective Human Stories*. https://news.un.org/en/story/2021/10/1102132
Whittle, C., Tiwari, S., Yan, S., & Williams, J. (2020). Emergency remote teaching environment: A conceptual framework for responsive online teaching in crises*. Information and Learning Sciences, 121*, 311-319.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Agile Learning as a Solution for Class Absence

Maria Francesca ROIG-MAIMÓ[a*], Ramon MAS-SANSÓ[a]
[a]*Universitat de les Illes Balears, Spain*
*xisca.roig@uib.es

Abstract: In this work, we have presented the results of the application of an agile learning methodology in a Computer Science Laboratory undergraduate course with the goal of decreasing the truancy rates. We have taught a Computer Science Laboratory course of a Telematics Engineering Degree over nine years. The first two academic years we followed a project-based approach. Although we had successfully used this methodology for two years, we noticed a significant increase in the students' absence in class. As an attempt to decrease that truancy rate, we adopted an agile learning methodology for the following seven years. Our primary hypothesis being that if a daily basis work is needed in order to deliver sprints on time, then class assistance should be a requirement to avoid falling behind. As we already expected, students easily went ahead in the subject and their satisfaction, commitment and involvement were greater. Moreover, we realized that truancy rates were lowered in a significant way.

Keywords: agile learning, computer science laboratory, agile methodologies, truancy rate, class absence

1. Introduction

The traditional learning approach, normally modelled as a one-way communication from the lecturer to the student, keeps playing an important role in high education institutions. However, in engineering degrees, due to its inherent practical nature, the project-based learning has been introduced as an alternative methodology to the conventional lectures.

Project-based learning tries to achieve optimal results by putting the student at the center of the learning process, emphasizing his autonomy and recognizing each particular learning style (Thomas 2000). When applied to the teaching of software development, project-based learning can be easily related to the sequential waterfall process, possibly due to its contemporaneity (Matković & Tumbas 2010). In such a model, the students are first proposed a project and then they are asked to analyze the requirements, design the product, implement the software and test the software to finally provide a running version. Each of the stages depends on the correct fulfillment of the previous one.

In 2013-14 we started to teach a Computer Science Laboratory (CSL) course of a Telematics Engineering Degree. The main goal of CSL is the strengthening of the programming skills attained throughout two previous programming courses. Therefore, the proposed program focuses mainly on the development of a computer application in all its stages so that programming skills are reinforced. Software and project management related competences are fulfilled in the curricula of the previous three years, so we mainly must promote the programming aspects in our program. We pursue the achievement of such learning outcome through the analysis, design, coding and validation of a mobile application. With this scenario, applying a project-based learning methodology looked like the most consequent.

The first two academic years of teaching this course (2013-14 and 2014-15) we followed a project-based approach. Although we had successfully used this linear approach for two years, we realized that in such a sequential process the students started the development stage very late. And, if we consider that this stage is the main road to achieve the course's goals, it might not be a suitable approach as a lot of time was devoted to previous steps. We also noticed a significant increase in the student's absence in class, possibly because, as they exactly knew what to do by the end of the semester, they thought they could do it without any supervision at their own pace. Assistance to class was not required but the main problem was that deviations from the project due to wrong design or implementation decisions were often detected too late. This sometimes led to students' demotivation, as they hat to rethink important items in their projects.

As a result from these concerns, we decided to shift to an agile learning methodology approach (from 2015-16 to 2021-22). The term agile learning, in contraposition to the cascade nature of project-

based learning, embodies the principles of agile programming included in the Manifesto for agile software development that was published in 2001 (Beck et al. 2001). This model, at first sight, could better fit the purpose of our course. In agile learning, the principles of agile software development are used in the context of learning. Agile software development uses an incremental model based on the continuous planning and team collaboration that follows the cycle of software development in an incremental way. The cycles of planning, design, building, testing and delivering a project are no longer large sequential stages that need to be completed before starting the new one but are incrementally updated to rapidly deliver working software to the customer.

In this work, we present the results of our experimental work on agile learning in a Computer Science Laboratory undergraduate course and compare them to our previous project-based approach. We particularly address the research question of whether the application of an agile learning methodology can decrease the truancy rates since a daily basis work is needed in order to deliver sprints on time.

In section 2 we first review the literature and in section 3 we compare both approaches and detail our implementation of the methodology. In subsequent sections we show and discuss the results obtained.

2. Previous work

Since the publication in the early years of the XXth century of The Project Method (Kilpatrick 1918) and its ulterior redefinitions as project-based learning (Guo et al. 2020), there have been many attempts to provide students with professional skills using a real problem-solve approach at different levels of education (Barron et al. 1988, Bell 2010, Kokotsaki et al. 2016, Morrison et al. 2020, Sharma et al. 2020, Tsybulsky et al. 2020).

Although Lang (2017) proposes the term agile learning to name the application of the processes and principles of agile development to learning, such a methodology has formerly been widely applied to divers education environments such as software engineering (Dewi et al. 2014, Scharf et all. 2013, Arcamone et al. 2013), interdisciplinary education (Gestwicki et al 2016), serious games (Lynch et al. 2011) or discrete mathematics (Duvall et all. 2017). These works show a growing interest in applying agile learning methodologies to allow students to work together in an energetic, targeted, and effective way in an educational context (Salza et al. 2019).

In his work, Lang states some of the advantages of agile learning. First, agile learning combines learning and application of learning as when new concepts are added, they are immediately applied in practice. And second, that due to the interactive nature of the methodology using sprints, students fail more and fail faster. He also states some disadvantages like that agile learning takes longer than project-based learning and that for students it is easier to fall behind.

Similar approaches and results can be found in the literature exploring the advantages of using an agile learning-based approach in computer science courses (Ju et al. 2020; Salleh et al. 2010; Roig-Maimó & Mas-Sansó 2022). In this work, we depict the importance of agile learning as a solution to the student class absence. As far as we know, there are no such results based on long term analysis (nine years) of class attendance when comparing both methodologies applied to programming reinforcing for non-computer scientist engineering students.

3. Implementation: project-based learning vs agile learning

The Computer Science Laboratory (CSL) course is a mandatory subject taught in the first semester of the last year of the Telematics Engineering Degree at our university. The number of students has important deviations depending on the academic year, with a mean of 15.5 students per year and a standard deviation of 4.2 along the CSL course evolution (see Fig. 1).

As previously mentioned, the main goal of the course is the strengthening of the programming skills attained throughout two previous programming courses. We seek the achievement of such learning outcome through the development of an Android mobile application. Android is chosen because the students are used to the JAVA language from the previous programming courses, so no additional time must be devoted to the learning of a new programming language.

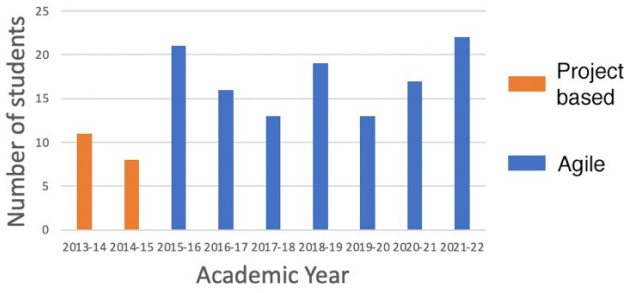


Figure 1: Students per academic year

The first two academic years (2013-14 and 2014-15) were we used a project-based methodology; students followed the traditional stages of software development in a cascade approach. At the beginning of the semester, the students were proposed a common programming project and they had to work out a solution. They first analyzed the requirements, then they fully designed the application, and then they started the project implementation and testing. Of course, they received the lecturers support all along the process.

The seven following academic years (from 2015-16 to 2021-22), were we used an agile learning methodology approach, we acted as the project customer, so we could change design criteria on the fly to force retrospective actions or to change initial requirements. When a retrospective analysis detected a requirements' change or that some modifications to the final product were needed, the students were asked to review the new application status. At the beginning of the semester, the students freely chose a team partner to work together all along the course.

4. The settings of the course following an agile learning methodology

In the first lecture of our agile learning methodology approach, we explained the main goals of the course and we introduced the agile learning methodology that we were going to use. We proposed a complete project that, as formerly stated, consisted in the development of a game application on an Android platform.

In that first lecture, we also provided the students with a proper view of the semester schedule. The semester duration of the course (15 weeks) was divided in a fixed number of sprints. A week had two class sessions of two hours each and six hours of autonomous work (which corresponds to 6 ECTS credits). Each sprint took one or two weeks, depending on its complexity, and resulted in the delivery of new functionalities to the final product. A sprint is one of the most characteristic trends in the Agile philosophy and it consists of a short term during which a certain task or set of tasks must be completed and the results obtained must be delivered. In the CSL course, a project was organized as a series of sprints that the students should go through.

The proposal of a sprint consisted of a written document specifying the title, the goal, the requirements and the delivery date of the result. The application requirements were presented as user stories to the students.

When a sprint finished, the students were asked to deliver their work and a new sprint was proposed. We reviewed the delivered work and gave feedback to the students before the new sprint began, so they had time to sort out detected problems.

Face-to-face class time was devoted to student guidance and support, and to the explaining of existing misunderstood concepts.

The frequent deliveries ensured fast and on time detection of programming errors or project concepts. Students' deviations were easily detected and, therefore, support efforts could be organized in order to correct such situations.

During the sprint's proposal, we often introduced specific points to be improved, forcing additional retrospective reflections.

By the end of the semester the students were asked to deliver the final product and a report documenting their work. The report consisted of a paper-like document where the students should describe what they did, the main troubles they faced and how they solved them. We also included a section where they could express their opinion about the course development and finally a self-assessment of their work.

5. Results and discussion

To analyze how the agile learning methodology has affected the teaching of our course along the years, we have collected qualitative and quantitative data:

- The qualitative data is extracted from the opinion section of the final reports of the students enrolled in the academic years where we applied the agile learning methodology approach (from 2015-16 to 2021-22). We have collected more than 350 qualitative comments from more than 70 different teams.
- The quantitative data is formed by the students' assistance and the students' assessments. It is obtained using the academic information from the academic students records over the complete longitudinal study (from 2013-14 to 2021-22).

5.1 Qualitative data: comments from the students reports

In this section, we present a summary of the results, extracting the ones that appeared the most. They are aligned with the results obtained in previous work:

- About working in pairs, a worth-mentioning aspect is that the students have felt very comfortable working in pairs, and they think it has allowed them to progress faster in the project. The importance of team working received a lot of comments.
- About working in sprints, the teams also point out that working with time spans of one or two weeks has been a very positive aspect, because they easily detected and corrected the designing and programming failures early and they could keep progressing without many problems.
- About consolidation of concepts, some students report that they have deeply understood basic concepts as object-oriented programming and top-down design. An easy adaptation to the agile learning methodology has also been widely reported.
- The feeling of satisfaction is also present in an important share of comments.

5.2 Quantitative data: students' assessments and assistance

Figure 3a depicts the graphical results of comparing what is the expected qualification of the student's pairs from their report (see orange color) in front of the actual assessment of their project (see blue color). We have used the data from all the academic courses, i.e., including the data of both approaches: project-based (2013-14 and 2014-15) and agile learning (from 2015-16 to 2021-22) bounded by a red vertical line in figure 3a.

As we can observe, the students' perceptions (self-assessment) are not too far from the assessment they get (actual assessment). However, we can see a noticeable overestimation in the first two years, which corresponds to the project-based approach methodology. That's an expected result because they don't get any feedback till the final assessment; while in the agile approach, they get feedback on their results after each sprint.

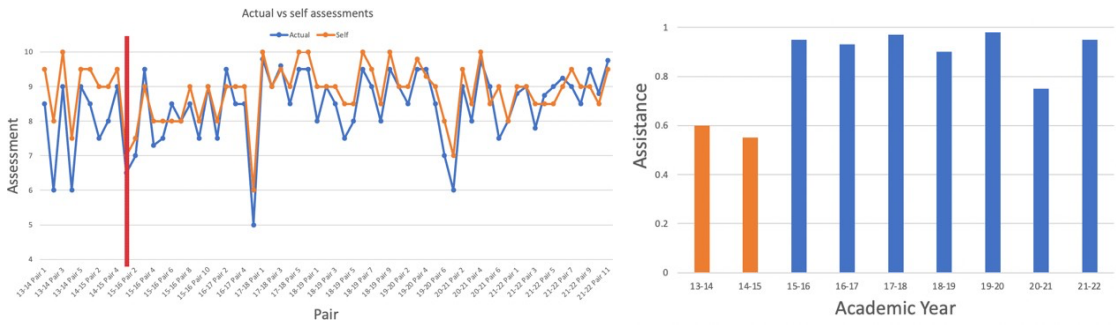


Figure 3: a) Perceived versus actual assessments and b) annual class assistance

Figure 3b shows the comparison of the assistance to class rate for both methodologies: the traditional project-based (see orange color) and the agile-based (see blue color) approaches. We should notice the data of the academic year 2020-21, which coincided with COVID pandemic, and consequently affected the assistance rate, which can be perceived as a decrease of around 20%. However, as we clearly observed, during the first two years, when a project-based learning approach was used, students' assistance to class was around 0.6, a 40% lower than the mean of 0.95 when using agile learning. The difference was statistically significant ($F_{1,7} = 33.35$, $p < .001$). The fact that short term deliveries are expected at the end of each sprint, motivates the teams' attendance as they can get immediate support and validation to their design and programming decisions. Moreover, we have observed that when somebody was absent from class, his teammate was attending the class. Rarely both team members were absent.

The data analyzed for Figure 3b encourage us to the application of agile learning approaches in order to motivate the students' assistance and, therefore, to decrease truancy rates.

6. Conclusion

In this work, we have presented the results of the application of an agile learning methodology in a Computer Science Laboratory undergraduate course with the goal of decreasing the truancy rates.

We have taught a Computer Science Laboratory course of a Telematics Engineering Degree over nine years. Although we had successfully used a traditional project-based methodology approach for two years, we noticed a significant increase in the students' absence in class. As an attempt to decrease that truancy rate, we adopted an agile learning methodology for the following seven years. Our primary hypothesis being that if a daily basis work is needed in order to deliver sprints on time, then class assistance should be a requirement to avoid falling behind.

At the end of every academic year, we collected qualitative and quantitative data.

From the qualitative data extracted from the final reports of the students we can stand out some positive aspects that are coherent with those previously pointed out in the literature. The students have felt very comfortable working in pairs; team working has been largely and positively commented. The time spans provided by the division of the project in sprints is another positive aspect as the students can detect and correct design and programming flaws in early stages of the project. They claim an easy adaptation to this learning model.

From the quantitative data formed by the students' assistance, it is remarkable that the truancy rate had significatively decreased, getting an assistance rate of 95%. This result led us to conclude that applying an agile learning approach is a good strategy to motivate the students' assistance and, therefore, to decrease truancy rates.

Acknowledgements

The authors acknowledge the Project EXPLainable Artificial INtelligence systems for health and well-beING (EXPLAINING) funded by PID2019-104829RA-I00 / MCIN/ AEI / 10.13039/501100011033. We also thank the University of the Balearic Islands, the Higher Polytechnic School (EPS) and the Department of Mathematics and Computer Science for their support.

References

Barron, B. J., Schwartz, D. L., Vye, N. J., Moore, A., Petrosino, A., Zech, L., & Bransford, J. D. (1998). Doing with understanding: Lessons from research on problem-and project-based learning. *Journal of the learning sciences*, 7(3-4), 271-311.

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Thomas, D. (2001). Manifesto for agile software development.

Bell, S. (2010). Project-based learning for the 21st century: Skills for the future. *The clearing house*: *A Journal of Educational Strategies, Issues and Ideas, 83*(2), 39-43

Dewi, D. A., & Muniandy, M. (2014, September). The agility of agile methodology for teaching and learning activities. In 2014 8th. *Malaysian Software Engineering Conference (MySEC*) (pp. 255-259). IEEE.

Duvall, S., Hutchings, D., & Kleckner, M. (2017). Changing perceptions of discrete mathematics through scrum-based course management practices. *Journal of Computing Sciences in Colleges*, 33(2), 182-189

Guo, P., Saab, N., Post, L. S., & Admiraal, W. (2020). A review of project-based learning in higher education: Student outcomes and measures. *International Journal of Educational Research, 102*, 101586.

Gestwicki, P., & McNely, B. (2016). Interdisciplinary projects in the academic studio. *ACM Transactions on Computing Education (TOCE)*, 16(2), 1-24.

Ju, A., Hemani, A., Dimitriadis, Y., & Fox, A. (2020, February). What Agile Processes Should We Use in Software Engineering Course Projects? In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 643-649).

Kilpatrick, W. (1918). The project method. *Teachers college record, 19*(4), 319-335.

Kokotsaki, D., Menzies, V., & Wiggins, A. (2016). Project-based learning: A review of the literature. *Improving schools*, 19(3), 267-277.

Lang, G. (2017). Agile learning: Sprinting through the semester. *Information Systems Education Journal*, 15(3), 14.

Matković, P., & Tumbas, P. (2010). A comparative overview of the evolution of software development models. *International Journal of Industrial Engineering and Management, 1*(4), 163.

Morrison, J., Frost, J., Gotch, C., McDuffie, A. R., Austin, B., & French, B. (2021). Teachers' role in students' learning at a project-based STEM high school: Implications for teacher education. *International Journal of Science and Mathematics Education, 19*(6), 1103-1123.

Roig-Maimó, M.F., & Mas-Sansó, R. (2022). Applying Agile Learning on Computer Science Laboratory: A Case Study. In T. Bastiaens (Ed.), *Proceedings of EdMedia + Innovate Learning* (pp. 455-464). New York City, NY, United States: Association for the Advancement of Computing in Education (AACE).

Salleh, N., Mendes, E., & Grundy, J. (2010). Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review. *IEEE Transactions on Software Engineering, 37*(4), 509-525.

Salza, P., Musmarra, P., & Ferrucci, F. (2019). Agile methodologies in education: A review. *Agile and lean concepts for teaching and learning*, 25-45.

Sharma, A., Dutt, H., Sai, C. N. V., & Naik, S. M. (2020). Impact of project-based learning methodology in engineering. *Procedia Computer Science, 172*, 922-926.

Scharf, A., & Koch, A. (2013, May). Scrum in a software engineering course: An in-depth praxis report. In 2013 *26th International Conference on Software Engineering Education and Training (CSEE&T)* (pp. 159-168). IEEE.

Thomas, JW. (2000). A review of research on project-based learning.

Tsybulsky, D. and Muchnik-Rozanov, Y. (2019). The development of student- teachers' professional identity while team-teaching science classes using a project-based learning approach: A multi-level analysis. *Teaching and Teacher Education*. 79, 48–59.

Werner, L., Arcamone, D., & Ross, B. (2012). Using Scrum in a quarter-length undergraduate software engineering course. *Journal of Computing Sciences in Colleges*, 27(4), 140-150.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Evaluation of the Quality of the Educational Software From the Perspective of Experiences of Italian and Polish Pre-Service Teachers

Łukasz **TOMCZYK**[a*], Laura **FEDELI**[b], Anna **WŁOCH**[c], Pierpaolo **LIMONE**[d], Monika **FRANIA**[e], Piergiorgio **GUARINI**[f], Michał **SZYSZKA**[g], Maria Lidia **MASCIA**[h], & Joanna **FALKOWSKA**[i]
[a] *Institute of Education, Jagiellonian University, Poland*
[b] *Department of Education, Cultural Heritage and Tourism, University of Macerata, Italy*
[c] *Faculty of Pedagogy and Psychology, Pedagogical University of Cracow, Kraków, Poland*
[d] *Rector Foggia University, Italy*
[e] *Faculty of Pedagogy and Psychology, Silesian University, Katowice, Poland*
[f] *Foggia University PhD student in Neuroscience and Education, Foggia, Italy*
[g] *Department of Pedagogy, WSB University, Dąbrowa-Górnicza, Poland*
[h] *Faculty of Humanistic Studies, University of Cagliari, Cagliari, Italy*
[i] *Institute of Educational Sciences, Nicolaus Copernicus University*
lukasz.tomczyk@uj.edu.pl

Abstract: The research involved evaluating the usability of modern educational software in two European countries (Italy, Poland). The research was conducted in the first quarter of 2022 by carrying out an online questionnaire using the LimeSurvey system. The online questionnaire was completed by 1209 prospective teachers. The software evaluation was limited to 22 types of teaching and learning support solutions, namely: Quizizz, Mentimeter, Wakelet, Padlet, Canva, Emaze, Answergarden, Jamboard, Coggle, Creately, Wordwall, LearningApps, Prezi, Kahoot, Plickers, Trimino, Dobble, Genial.ly, ClassDojo, Explain Everything, KhanAcademy, Easelly. The quantitative analysis of the collected data allowed us to draw the following conclusions: 1) In both countries, the most effective tools evaluated as: Canva, Quizizz, Prezi, Kahoot, Wordwall; 2) Canva software is the most well-known digital teaching tool in both countries; 3) There is a large variation in the evaluation of software in both countries; 4) In Poland and Italy pre-service teachers evaluate Canva, Answergarden, ClassDojo, KhanAcademy in a similar way; 5) The vast majority of respondents in both countries have no experience with educational software or rate the software low; 6) Among Italian respondents there is a constant tendency in the evaluation of educational software - a positive evaluation of one type of software is in a positive relation with the evaluation of another type of software; 7) Polish pre-service teachers are more diverse in their evaluation of software - a positive evaluation of one software does not necessarily condition a positive evaluation of another digital teaching resource.

Keywords: software, school, university, teachers, pre-service teachers, Italy, Poland

1. Introduction

Pre-service teachers (current students in the field of pedagogy) are one of the key groups for the development of education. Professional preparation of teachers is a priority issue in many countries (Romano, 2002; Melnyk et al., 2019). The formation of key competences in this group plays a special role for maintaining the quality of formal education. Therefore, due to the changes occurring through the intensive growth of the information society, more and more attention is being paid to the skills, knowledge and attitudes related to ICT use concerning pre-service teachers (Tomczyk & Fedeli, 2022). Digital competence, which encompasses a wide range of activities and extensive knowledge related to the use of ICT in different contexts of professional and private life, is now as much a matter of course as even handling analogue media three decades back. Therefore, the analysis of digital competences in particular professional groups requires taking into account the key conditions of a given profession. One of the strategic elements in the work of contemporary teachers is the use of educational software. This, in turn, is conditioned by knowledge of the potential of given e-solutions, attitude towards ICT (including assessment of software quality), previous experience of using given software (also in another

role, e.g. as a student), the level of basic - basic digital competences, or financial conditions of school institutions (Tomczyk, 2021).

This text is part of the debate on preparing new pedagogical staff in terms of the level of digital competence narrowed down to ICT-supported teaching activities. Such an analysis is necessary in the context of challenges related not only to crisis (pandemic) e-learning, but also relating to the level of attitude towards new media among pre-service teachers.

2. Methodology

2.1 Research aim and procedure

The aim of the research was to evaluate popular educational software among Polish and Italian pre-service teachers. The comparative research is in line with the challenges of educating modern teachers in the information society and is also linked to the modernisation of programmes related to the formation of digital competences in this key group. The aim of the research was clarified by the following five groups of research problems: RQ1: How do pre-service teachers in Poland and Italy assess modern educational software? RQ2: What are the differences and similarities between the evaluation of educational software among students from pedagogical faculties in two European countries?

The research involved students from two European countries, namely Italy (IT) and Poland (PL). 604 students from Italy (Macerata, Foggia, Genoa, Sardinia) and 605 students from Poland (Kraków, Katowice, Toruń, Dąbrowa-Górnicza) were included in the quantitative research. The selection of the study was similar to a random scheme (different regions in both countries). The research was conducted in the first half of 2022 using the online tool LimeSurvey in two languages.

2.2 Characteristics of the research sample

The vast majority of participants in the study were women (IT 85.4%, PL95.0%). This is due to the fact that in both countries the teaching profession is highly feminised. Both teachers and prospective teachers in Poland and Italy are predominantly women. The average age in both countries was similar, in IT=32.1 and in PL=25.5. The main participants of the study were students of the second level of pedagogical studies (Master's), IT = 86.42% PL = 69.25%. The remaining group consisted of undergraduate level students.

2.3 Research tool

The tool used in this study was the one constructed by a team of Polish-Italian media pedagogy researchers (Tomczyk & Fedeli, 2022). This text includes the analysis of one variable from the final version of the tool, which concerned the usability evaluation of educational software. The variable consisted of 22 indicators (of different types of software). Students in both countries could rate the software on a five-point likert scale from 1 - very low usefulness in education to 5 - very high usefulness in education. In addition, respondents could select option 0, which meant that they could not evaluate the software due to unfamiliarity with the software. The triangulation of qualitative and quantitative research results from, among others, the SELI project (Tomczyk & Sunday Oyelere, 2019; Arteaga et al., 2020), as well as the analysis of the literature, allowed the formulation of a software list. In addition, the authors of the tool are authors of academic courses aimed at increasing the level of digital competence in the group analysed. The cumulative teaching and project experience and expert knowledge formed the basis for the shortlisting of educational software. The respondents' own expert experience and also a literature review on current educational software: Quizizz (Zhao, 2019), Mentimeter (Rudolph, 2018), Wakelet (Graham, 2018), Padlet (Beltrán-Martín, 2019), Canva (Rahmonovna & Erkinovana, 2022), Emaze (Nazarenko et al., 2020), Answergarden (Jusmaya, 2022), Jamboard (Alanya-Beltran et al., 2021), Coggle (Debbag et al., 2021), Creately (Ren et al., 2020), Wordwall (Moorhouse & Kohnke, 2022), LearningApps (Behnamnia et al. 2020), Prezi (Strasser, 2014), Kahoot (Wang & Tahir, 2020), Plickers (Shana & Abd Al Baki, 2020), Trimino (Badan & Onishchenko, 2021), Dobble (Baćić Đuračković & Đuračković, 2020), Genial.ly (Kaźmierczak, 2020),

ClassDojo (Williamson, 2017), Explain Everything (Pelton & Pelton, 2013), KhanAcademy (More & Vankadara, 2022), Easelly (Weiner & Lorber, 2021).

3. Results

RQ1: Evaluation of educational software by pre-service teachers in Poland (descriptive statistics)

As mentioned in the introduction, ICT has changed the quality of education. New media have transformed both the quality of learning and the processes involved in teaching. Nowadays (in the post-covid era) there is a need to reflect on the frequency with which different software is used in education. In this section, ways of evaluating popular educational software by pre-service teachers are presented.

Italian students from pedagogical faculties, similarly to their Polish peers, have varying levels of evaluation of particular educational software. Among the most useful they rank: Quizizz, Mentimeter, Padlet, Canva, Jamboard, Wordwall, Kahoot. Also as in the case of Polish pre-service teachers, there is no single trend here. A varied type of software is highly rated, i.e. for creating: quizzes, real-time presentations, mind maps and diagrams, complex graphic forms, individual exercises or tasks based on gamification. However, the majority of students have a problem with evaluating educational software - similarly to Polish students. This means that they do not have sufficient knowledge and experience with the mentioned software. Lack of knowledge - ability to evaluate the mentioned software is also one of the factors blocking the use of digital teaching aids in their future professional work.

Based on the available data collected among Polish students, it was noted that most of the listed software is difficult to evaluate by students of pedagogical faculties. On the other hand, software such as Canva, Quizizz, LearningApps, Prezi, Kahoot, Wordwall, Genial.ly enjoy the greatest recognition - high ratings for particular software. Each of the listed software can be used for a different teaching purpose. Therefore, there is no tendency among the Polish respondents for one type of software. There is also no single type/type of software, which was very negatively evaluated by Polish students. Descriptive statistics (percentages) for the evaluation of particular software are presented in Table 1.

Table 1. *Evaluation of educational software by pre-service teachers in Italy and Poland*

	I have not dealt with this software		Very low usefulness in education		Low utility		Average usefulness		Good usefulness		Very good usefulness	
Country	IT	PL	IT	PL	IT	PL	IT	PL	IT	PL	IT	PL
1.Quizizz	69.2	33.4	1.2	0.5	0.7	1.5	7.5	6.9	13.4	31.2	8.1	26.4
2.Mentimeter	66.1	81.5	1.2	0.0	1.2	1.0	6.8	2.6	9.8	8.3	15.1	6.6
3.Wakelet	87.3	94.0	1.3	0.2	1.0	0.5	4.1	1.5	4.5	2.6	1.8	1.2
4.Padlet	51.3	75.4	1.3	0.5	2.0	0.3	9.8	4.0	14.9	11.4	20.7	8.4
5.Canva	18.7	34.0	2.8	0.2	2.0	1.5	14.2	4.3	30.1	17.9	32.1	42.1
6.Emaze	87.4	95.5	1.2	0.2	0.5	1.2	3.3	0.8	6.0	1.7	1.7	0.7
7.Answergarden	89.4	92.6	0.7	0.3	1.2	1.3	3.1	1.2	4.8	3.1	0.8	1.5
8.Jamboard	67.1	86.8	1.0	0.2	0.8	0.8	8.6	3.0	11.9	6.0	10.6	3.3
9.Coggle	80.0	86.9	1.7	0.3	0.7	0.8	4.0	2.0	7.3	4.6	6.5	5.3
10.Creately	85.8	93.1	1.5	0.5	0.8	0.3	3.8	3.3	6.8	1.8	1.3	1.0
11.Wordwall	60.1	51.7	1.5	0.5	1.5	2.0	7.0	3.8	12.3	18.3	17.7	23.6
12.LearningApps	73.5	56.2	1.5	0.2	1.2	0.7	5.8	4.3	8.6	18.0	9.4	20.7
13.Prezi	72.7	37.0	0.8	1.0	1.2	2.1	7.1	11.9	10.3	24.1	7.9	23.8
14.Kahoot	58.8	25.3	1.2	0.3	0.8	1.5	6.0	4.6	14.9	21.8	18.4	46.4
15.Plickers	89.1	94.0	1.0	0.8	0.2	0.7	3.8	2.1	4.6	1.0	1.3	1.3
16.Trimino	89.9	96.2	1.3	0.3	1.2	0.5	2.8	1.8	4.0	0.7	0.8	0.5
17.Dobble	86.1	75.2	0.7	0.3	1.0	1.5	4.8	3.3	5.6	9.1	1.8	10.6

18.Genial.ly	84.6	57.2	1.2	0.2	0.8	1.2	3.5	6.6	6.6	14.7	3.3	20.2
19.ClassDojo	88.6	87.3	1.2	0.5	0.8	1.0	3.3	3.1	5.0	3.8	1.2	4.3
20.Explain Everything	88.7	95.4	1.5	0.3	0.8	1.2	3.1	1.5	4.3	1.0	1.5	0.7
21.KhanAcademy	88.9	88.1	1.3	0.2	0.7	1.0	3.8	2.3	4.1	4.8	1.2	3.6
22.Easelly	89.4	93.7	1.0	0.2	0.7	1.3	3.5	1.2	4.5	2.6	1.0	1.0

RQ2: Differences in software evaluation between Italian and Polish students

Out of twenty-two types of educational software, only in four cases there are no differences in terms of usability evaluation depending on the country of the respondents. This situation refers to: Canva, Answergarden, ClassDojo, KhanAcademy. It should be mentioned here that Canva is the software most known in the surveyed group in both countries. In the remaining 18 cases, there is a difference in the evaluation of the software. There is no consistent trend in the evaluation of software according to its type. For example, applications such as: Quizizz, Wordwall, LearningApps, Prezi, Kahoot, Dobble, Genial.ly are evaluated higher by Polish pre-service teachers. The differences in effectiveness ratings by country are presented in Table 3. The non-parametric Mann-Whitney U test was used to assess the statistical significance of the differences.

Table 2. *Differences in software evaluation between Italian and Polish students*

	Mean Italy	Mean Poland	Std.Dev. Italy	Std.Dev. Poland	U	Z-adjusted	p-value
1.Quizizz	1.190	2.815	1.863	2.094	109171.0	-13.152	0.000
2.Mentimeter	1.382	0.760	2.023	1.637	154261.5	6.069	0.000
3.Wakelet	0.427	0.220	1.190	0.906	170519.5	3.977	0.000
4.Padlet	1.977	1.008	2.152	1.816	138460.5	8.474	0.000
5.Canva	3.306	2.982	1.820	2.227	181531.5	0.203	0.839
6.Emaze	0.442	0.149	1.226	0.729	167775.0	5.082	0.000
7.Answergarden	0.358	0.264	1.089	0.983	177106.5	1.858	0.063
8.Jamboard	1.291	0.511	1.927	1.350	146328.0	8.129	0.000
9.Coggle	0.763	0.529	1.613	1.415	170496.0	3.111	0.002
10.Creately	0.483	0.233	1.256	0.895	169244.0	4.155	0.000
11.Wordwall	1.629	2.074	2.108	2.231	163962.5	-3.426	0.001
12.LearningApps	1.028	1.898	1.808	2.207	146524.0	-7.019	0.000
13.Prezi	1.053	2.565	1.795	2.099	113595.5	-12.535	0.000
14.Kahoot	1.722	3.367	2.145	2.062	110220.5	-12.691	0.000
15.Plickers	0.379	0.192	1.136	0.826	173470.5	3.158	0.002
16.Trimino	0.321	0.119	1.027	0.635	171154.5	4.319	0.000
17.Dobble	0.487	1.025	1.265	1.848	160459.0	-5.320	0.000
18.Genial.ly	0.563	1.820	1.390	2.176	128855.0	-11.083	0.000
19.ClassDojo	0.384	0.486	1.132	1.333	179715.5	-0.872	0.383
20.Explain Everything	0.373	0.144	1.121	0.698	170475.0	4.299	0.000
21.KhanAcademy	0.364	0.464	1.097	1.309	180478.0	-0.664	0.507
22.Easelly	0.356	0.218	1.090	0.885	174898.5	2.670	0.008

4. Limitations of the study

The list of educational software and OER mentioned in the study is limited. This means that a wide range of other solutions to support the learning and teaching process were not included in the study. In general, creating a complete list is an extremely difficult (or even impossible) task at the current stage of development of the information society. Future editions of the study should feature an expanded list of educational software. In addition, the evaluation of the listed software was carried out on the basis of the

subjective feelings of future educators (e.g. based on their own complex educational experiences), which means that the scale used is not sharp and needs to be refined at a later stage. Due to the limitations of the length of the text, it is also postulated that the statistical analyses should be extended to include issues related to socio-demographic characteristics (level of study, gender, metric age, place of residence), which will be carried out in subsequent research projects. The study sample is highly feminised as highlighted by one of the reviewers. Nevertheless, it should be emphasised that this is a typical situation in this group in both countries and is not a methodological limitation in the strict sense.

5. Conclusions

Analyses of the young generation of users show that this is a generation that uses new media intensively in a way that differs from previous generations, e.g. those in middle adulthood. Many myths have grown up around the generation of digital natives - the group studied in this text (Margaryan et al., 2011). One of them is the assumption that the representatives of this group use all the possibilities offered by new media, or cyberspace, in an intentional and effective way. The data collected among future pedagogical staff in Poland and Italy, as well as previous comparative studies conducted in other countries (Eger et al., 2018), show that there is a gap in the constructive and full use of the potential of new media. The descriptive statistics presented clearly show that most educational software is unknown to young pedagogical students in the two European countries. Despite the fact that these are people who use ICT very intensively in the area of communication and entertainment, there is a gap among most of the respondents related to the evaluation of educational software due to the lack of digital competence oriented towards the teaching profession.

Of course, this article (the methodology used related to the selection of the software list) can be discussed in the area of adding or removing selected solutions based on the use of software in education. Nevertheless, the collected data clearly suggest a gap in the preparation of teachers in terms of digital teaching resources, as well as having undeveloped experiences related to the use of modern software in the learning and teaching process (Jagušt et al., 2018). Those who are able to fully evaluate educational software with any experience in this area are a minority. Thus, the collected data provide a signpost for educational institutions, which should continuously monitor the level of professional digital competence of future pedagogical staff and, at the same time, modify academic courses preparing for the profession in the context of challenges related to the digitalisation of education.

Acknowledgements

The article was written as part of the project "Teachers of the future in the informationsociety—between risk and opportunity paradigm" funded by the Polish National Agency for Academic Exchange under the Bekker programme Grant number: PPN/BEK/2020/1/00176.

References

Alanya-Beltran, J., Salvatierra, M. S. A., Espinoza, M. D., & Tataje, F. A. O. (2021). Educación durante la pandemia COVID-19. Uso de la tecnología en la nube: Jamboard. *Revista Ibérica de Sistemas e Tecnologias de Informação*, (E44), 39-48.

Arteaga, M., Tomczyk, Ł., Barros, G., & Sunday Oyelere, S. (2020). *ICT and education in the perspective of experts from business, government, academia and NGOs: in Europe, Latin America and Caribbean*. Universidad del Azuay.

Baćić Đuračković, L., & Đuračković, V. (2020). Matematički Dobble. *Poučak: časopis za metodiku i nastavu matematike*, 21(84), 52-57.

Badan, A., & Onishchenko, N. (2021). Multimedia technologies in foreign language learning under pandemic. In *CEUR Workshop Proceedings* (Vol. 2870, pp. 642-656).

Behnamnia, N., Kamsin, A., & Ismail, M. A. B. (2020). The landscape of research on the use of digital game-based learning apps to nurture creativity among young children: A review. *Thinking Skills and Creativity*, 37, 100666.

Beltrán-Martín, I. (2019, July). Using Padlet for collaborative learning. In *HEAD'19. 5th International Conference on Higher Education Advances* (pp. 201-211). Editorial Universitat Politècnica de València.

Debbag, M., Cukurbasi, B., & Fidan, M. (2021). Use of digital mind maps in technology education: a pilot study with pre-service science teachers. *Informatics in Education*, 20(1), 47-68.

Eger, L., Klement, M., Pisoňová, M., & Petrová, G. (2018). Different user groups of university students and their ICT competence: evidence from three countries in central Europe. *Journal of Baltic Science Education*, 17(5), 851.

Graham, K. (2018). TechMatters: Doing Digital Curation with Wakelet. *LOEX Quarterly*, 45(2), 3.

Jagušt, T., Botički, I., & So, H. J. (2018). Examining competitive, collaborative and adaptive gamification in young learners' math learning. *Computers & Education*, 125, 444-457.

Jusmaya, A. (2022). The Effectiveness of Using Answer Garden Application as Brainstorming Activity in Online Learning. *Komposisi: Jurnal Pendidikan Bahasa, Sastra, dan Seni*, 23(1), 70-81.

Kaźmierczak, P. (2020). Czy aplikacja Genial. ly rzeczywiście jest genialna? O potencjale narzędzi online podczas zdalnego nauczania języka polskiego jako obcego. *Acta Universitatis Lodziensis. Kształcenie Polonistyczne Cudzoziemców*, (27), 559-570.

Margaryan, A., Littlejohn, A., & Vojt, G. (2011). Are digital natives a myth or reality? University students' use of digital technologies. *Computers & Education*, 56(2), 429-440.

Melnyk, N., Bidyuk, N., Kalenskyi, A., Maksymchuk, B., Bakhmat, N., Matviienko, O., ... & Maksymchuk, I. (2019). Models and organisational characteristics of preschool teachers' professional training in some EU countries and Ukraine. *Zbornik Instituta za pedagoska istrazivanja*, 51(1), 46-93.

Moorhouse, B. L., & Kohnke, L. (2022). Creating the Conditions for Vocabulary Learning with Wordwall. *RELC Journal*, 00336882221092796.

More, N. S., & Vankadara, R. (2022). Online Teaching Tools: Challenges and Their Solutions During a Pandemic Available in India. In *ICT Analysis and Applications* (pp. 547-556). Springer, Singapore.

Nazarenko, L., Palamar, S., Vaskivska, H., & Nezhyva, L. (2020). Didactic Potential of New Generation ICT in Forming Information and Communication Competence of Upper Secondary School Pupils. In *ICTERI* (pp. 248-261).

Pelton, T., & Pelton, L. F. (2013, March). Using an iPad to Explain Everything by creating interactive activities and vignettes. In *Society for Information Technology & Teacher Education International Conference* (pp. 4843-4847). Association for the Advancement of Computing in Education (AACE).

Rahmonovna, O. D., & Erkinovana, B. D. (2022). Creator of Online Presentations Canva. com Possibilities of Application in All Areas of the Program. *International Journal of Development and Public Policy*, 2(2), 52-55.

Ren, R., Castro, J. W., Santos, A., Pérez-Soler, S., Acuña, S. T., & de Lara, J. (2020). Collaborative modelling: chatbots or on-line tools? An experimental study. In *Proceedings of the Evaluation and Assessment in Software Engineering* (pp. 260-269).

Romano, M. (2002). Training teachers for quality education in Europe. *European journal of teacher education*, 25(1), 11-17.

Rudolph, J. (2018). A brief review of Mentimeter–A student response system. *Journal of Applied Learning & Teaching*, 1(1), 35-37.

Shana, Z. A., & Abd Al Baki, S. (2020). Using Plickers in Formative Assessment to Augment Student Learning. *International Journal of Mobile and Blended Learning (IJMBL)*, 12(2), 57-76.

Strasser, N. (2014). Using Prezi in higher education. *Journal of College Teaching & Learning (TLC)*, 11(2), 95-98.

Tomczyk, Ł. & Fedeli, L. (2022). *Digital Literacy for Teachers*. Lecture Notes in Educational Technology. Singapore: Springer.

Tomczyk, Ł. & Fedeli, L., (2022). *Digital literacy pre-service teachers English tool*. 10.13140/RG.2.2.23634.84169

Tomczyk, Ł. (2021). Declared and real level of digital skills of future teaching staff. *Education Sciences*, 11(10), 619.

Tomczyk, Ł., & Sunday Oyelere, S. (2019). *ICT for Learning and Inclusion in Latin America and Europe. Case Study From Countries: Bolivia, Brazil, Cuba, Dominican Republic, Ecuador, Finland, Poland, Turkey, Uruguay*. Cracow: Pedagogical University.

Wang, A. I., & Tahir, R. (2020). The effect of using Kahoot! for learning–A literature review. *Computers & Education*, 149, 103818.

Weiner, A., & Lorber, K. (2021, March). Infographics: A Methodology for Student Research Presentations and Other Academic Projects. In *Society for Information Technology & Teacher Education International Conference* (pp. 649-652). Association for the Advancement of Computing in Education (AACE).

Williamson, B. (2017). Decoding ClassDojo: Psycho-policy, social-emotional learning, and persuasive educational technologies. *Learning, Media and Technology*, 42(4), 440-453.

Zhao, F. (2019). Using Quizizz to Integrate Fun Multiplayer Activity in the Accounting Classroom. *International Journal of Higher Education*, 8(1), 37-43.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

The Perspectives of Distance Learning in Higher Education: Lessons Learned from the COVID-19 Pandemic

Dariusz DYMEK[a], Mariusz GRABOWSKI[a] & Grażyna PALIWODA-PĘKOSZ[b*]
[a]*Department of Computational Systems, Cracow University of Economics, Poland*
[b]*Department of Informatics, Cracow University of Economics, Poland*
*paliwodg@uek.krakow.pl

Abstract: The COVID-19 pandemic has triggered changes in higher education and put to light the question of the future of education. The goal of the research was to gather the opinions of students, who are the main stakeholders of education, concerning their perception of various forms of education in the context of distance learning and their preferences and predictions on the future shape of education. 1005 students of the Cracow University of Economics, Poland, share their opinion participating in the survey. The results point to a diversification of the student population as far as their perception of distance learning is concerned: students with greater experience in traditional higher education and part time students favored distance learning more and foresee it as the future of education, which is in line with the necessity of facilitating life-long learning. Hence, universities have to adjust their offer to different student preferences in order to survive.

Keywords: Poland, COVID-19 pandemic, students, distance learning

1. Introduction

The COVID-19 pandemic forced educational institutions to elaborate emergency plans for conducting educational processes in a distance form. It seems that the pandemic speeded up some unavoidable processes, such as the transformation of education related to the enhancement of information and communication technologies (ICT) skills, changes in the youth lifestyle, and changes in labor market expectations. The current graduates will have to requalify a number of times during their working life (Regmi, 2015). Besides, the growing number of secondary school graduates prefer to start work early to quickly gain practical experience (some of them come back to formal education after a gap), or decide to combine work with studies (part-time students). This phenomenon is in line with employers' requirements who value candidates with experience (Helyer & Lee, 2014). These premises imply the necessity of lifelong learning and infrastructure that facilitates it. Hence, the lessons learned from the COVID-19 pandemic concerning the students' perception of distance learning might comprise a valuable foundation for the future teaching strategy development by universities. There is a number of research works that tackle the distance learning domain, including students' wellbeing (Herbert et al., 2021; Conrad et al., 2021; Liu et al., 2021), grading (Vander Schee & Birrittella, 2021), and new possibilities of distant communication (Ackerman & Gross, 2021). However, to the best of our knowledge, there is no comprehensive research that addresses the role of distance learning in a multifaceted way, taking into account students as the main stakeholders of the learning experience. Hence, the goal of our research was to gather students' feedback concerning the effectiveness of various components of learning (form of classes, assessment, and communication) as well as students' preferences concerning the forms of learning. More specifically, we would like to answer the following research questions:

- RQ1: What is the students' opinion concerning the different forms of teaching applied in traditional and distance education?
- RQ2: How do the students perceive the suitability of various forms of teaching for distance learning?
- RQ3: What are the students' preferences concerning distance and traditional learning?
- RQ4: How do the students predict the role of distance learning in education?

We answered the research questions on the basis of data gathered from 1005 students of the Cracow University of Economics, Poland.

2. Related Work

The spring of 2020 connected with the beginning of the COVID-19 pandemic put high pressure on the educational institutions all over the world. The educational organizations in general and the higher educational institution (HEIs) in particular, had to enter the new paradigm of educational content delivery called emergency remote teaching (ERT) (Code et al., 2020; Rapanta et al., 2021). This imposed high tension on both the students and the teachers. Therefore, it put to the attention the problem of personal comfort and mental wellbeing. During the pandemic, 52% of students declared depressive symptoms (Herbet et al., 2021). Loneliness and isolation have also been harmful factors before the pandemic, but during this period they have increased. The main reason for this was the closing of campuses (Rippe et al., 2021; Stewart & Lowenthal, 2021). Naddeo et al. (2021) have identified the factors that affected online teaching, learning, and general wellbeing during the pandemic in Italy. These factors include, among others, problems in interactions with colleagues, adaptation of homes for online work, malfunction of e-learning platforms due to high Internet traffic and/or large number of online classes, using multiple devices, or inappropriate posture and clothing. Stress, anxiety, depressive symptoms, self-reported worry, grief, generalized anxiety and post-traumatic stress disorder symptoms were confirmed by Conrad et al. (2021) and Lee et al. (2021). Liu et al. (2021) have proposed a model of students' psychological wellbeing. It consisted of six negative and three positive predictors. Negative ones include perceived worry about mental health, restriction stress, perceived sufficiency of distance communication, white/European ethnicity, dietary changes, and social isolation, whereas positive ones contain physical health status, emotional support, and resilience. The remaining factors, i.e. gender, age, educational level, and international status, do not reveal any significant influence on the analyzed phenomenon. The pandemic imposed negative effects also on the faculty. The teachers had to respond to the need of ERT even though their mother institutions were closed. For this reason, the level of technostress among the faculty increased during the pandemic (Boyer-Davis, 2020).

The time since the start of the pandemic allows us to draw some conclusions concerning the suitability of certain teaching and education management methods. One of such challenges is grading (Vander Schee & Birrittella, 2021). The authors found peer group grading as a fair and time saving assessment tool. The results proved that grades assigned by students were not significantly different from those assigned by a teacher. The other important finding is that students perceive group grading as a fair solution. Meshram et al. (2022) have examined the influence of a crisis, such as a pandemic, on students' self-regulated learning and grade expectations. The outcomes of the research indicate direct and significant impacts of self-regulated learning and its components on grade expectations. There are also examples of research that aim at the verification of certain theories in the context of distance learning. An interesting example of enriching commonly used educational methods is the use of a synchronous online discussion board as the primary mode of teaching and learning (Ackerman & Gross, 2021). The results showed that student achieved higher examination results compared with the learning delivered by Zoom. On the other hand, students experienced frustration with the interface. Another important factor referred to online teaching and learning is the analysis of lecturers' productivity (Afrianty et al., 2022). The results indicate that it is important to enhance digital skills of the academic staff to achieve better results of online teaching and learning.

After a certain period of online teaching during the pandemic a question arises, whether the online form is only a temporary trend or will it become the reality for education institutions. The study by Agarwal et al. (2021) showed that the proponents and opponents of distance learning are divided in half. Therefore, if the educational institutions want to convince the undecided, they should promote online courses and fine-tune their offer to make it more suitable for online education. Krishnamurthy (2020) advocates the same opinion. According to the author, the educational system in the post-COVID-19 world needs a major redesign. The components of the online offer include artificial intelligence, machine learning, and automation. Hence, in the future, teachers might be partially replaced by sophisticated software. Code et al. (2020) has conducted research on the impact of ERT on technology-oriented courses. They indicate the inadequacy of pure online strategies in this matter. The future of online education demands providing the answers to several questions (Rapanta et al., 2021).

For example, do HEIs need to be more digitized or pedagogically prepared for the new challenges? It is also vital to ask whether the educator should strengthen their 'hard' or 'soft' skills for innovative learning. The design of the post-COVID learning environments, whether they are online or hybrid, is a challenging and complex endeavor. Spais and Paul (2021) have studied a crisis management model on the teaching of marketing in the higher education landscape. The authors provide crisis management examples from the three stages of crisis management i.e., identification, confrontation, and reconfiguration. The main concerns coming from the presented approach are the balance between short-term defensive and preventive actions and long-term strategic activities in the education domain. Therefore, the main tasks for crisis management are to secure educational continuity planning.

3. Research Methods

The aim of the study was to assess the impact of the pandemic on attitudes towards distance learning. The situational advantage was that all students were forced into this form of learning, regardless of their previous preferences and experience in this field. This gives a deeper picture of the students' attitudes, as it encompasses the whole student population regardless of their skills and previous attitude towards distance learning.

The starting point for the research was the development, on the basis of a literature review and the authors' own over 20-year experience in higher education teaching, of a list of potential factors that may affect the students' attitude towards distance learning. The list includes the following factors: gender, age, form of studies, self-assessment of ICT skills, technical conditions (Internet availability), and previous experience with the traditional (on premise) form of study. In addition, we would also like to take into account the students' attitudes towards online communication and their self-perceived change in general wellbeing related to the pandemic. We developed a survey that addresses these factors. The vast majority of questions in the survey was designed following a five-point Likert balanced and symmetric scale principle (Joshi et al., 2015). Its items relevant to the presented research are listed in Appendix. It should be noted that the questionnaire items reflect the needs stated in the research questions. Their goal was to gather the students' opinions concerning various phenomena and not to specifically assess these phenomena.

The survey was implemented using the Microsoft Office 365 tool. The questionnaire was sent to all Cracow University of Economics students (in total about 12 190 students, including bachelor, master, and PhD students). Data were gathered in May and June of 2021. We received 1005 responses, which gives an 8% global response rate (among full-time students the response rate was 10% and among part time-students 5%). For such values, with a statistical significance of alpha equal to 0.95, the error margin is 3% (broken down into full-time and part time students it is 3% and 6%, respectively). The preliminary results of the research that tackled the social aspects of online learning of Polish students were published in (Dymek et al., 2021). In this research, we would like to focus on the differences between full-time and part-time students and students with different previous experience in traditional learning, in the perception of the effectiveness of various forms of learning and preferences concerning the future form of learning.

4. Respondent Structure

Table 1 presents the respondent structure. It is worth noting that part-time students comprise mostly of working people, with some professional experience. They are much older than full-time students (on average they are approximately 3 years older, they are at age 24+, i.e. after the traditional educational age). In Poland students usually finish bachelor studies at the age of 22, master studies at 24 (2 year master studies in Poland). Many people start working right after finishing bachelor studies, sometimes continuing their education outside of the universities or coming back to formal university education after a break. Part-time students might be perceived as people who follow the idea of life-long learning. Therefore, the majority of part-time students is above 22 years old, while a significant part of the full-time students is in the range 19-22.

On average, students reported a deterioration of self-perceived wellbeing during the pandemic. However, some differences might be observed between groups. Interestingly, students with high experience in traditional learning seem to be the less affected group as far as wellbeing is concerned (43 percent reported deterioration), whereas the medium experience group is highly affected (58% reported

deterioration) (Dymek et al., 2021). However, the biggest differences were noticed between full-time and part-time students, the latter reported more often improvement in their wellbeing (Figure 1).

Table 1. *Respondents' Structure*

Variable	Category	No.	%
Gender	Female	643	64.0%
	Male	357	35.5%
	Not specified	5	0.5%
Age	< 20	143	14.2%
	20 - 24	727	72.3%
	25 - 30	99	9.9%
	30+	36	3.6%
Form of study	Full-time	766	76.2%
	Part-time	239	23.8%
Experience in a higher education traditional form of studies	Low (lack of experience)	375	37.3%
	Medium (about half a year experience)	154	15.3%
	High (at least 1.5 year experience)	476	47.4%

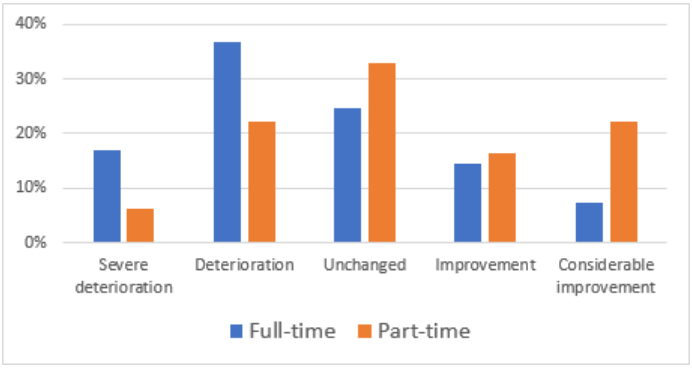


Figure 1. Students' Wellbeing in Relation to the Form of Study.

The preliminary data analysis shows that the number of respondents indicating an insufficient level of ICT skills or insufficient level of technical conditions was very small (2.4% and 2.8%, respectively), hence we excluded these criteria from further analysis.

5. Results

5.1 RQ1: Students' Evaluation of the Different Forms of Teaching in Distance Learning

Respondents assessed workshops and labs as being worse in distance learning, no matter how experienced they had been in the traditional form of learning (Table 2). However, more than half of the respondents with low or high experience perceived lectures as better performed in distance learning. As far as seminars are concerned, respondents were almost evenly distributed according to their opinions in this matter.

Table 2. *Students' Evaluation of the Different Forms of Teaching in the Distance Learning Form in Relation to Students' Experience in Traditional Learning*

	Workshops			Lectures			Seminars			Labs		
	L	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**
Worse	57%	58%	51%	23%	31%	25%	32%	28%	31%	52%	41%	51%
Neutral	19%	12%	17%	20%	30%	22%	35%	31%	31%	23%	29%	18%
Better	24%	30%	32%	57%	39%	53%	33%	40%	38%	25%	29%	31%

L – Low; M – Medium; H - High experience in traditional learning

The biggest differences in respondents' assessment were noticeable between full-time and part-time students, the latter favoring distance learning more (Table 3).

Table 3. *Students' Evaluation of the Different Forms of Teaching in the Distance Learning Form by Full-time and Part-time Students*

	Workshops		Lectures		Seminars		Labs	
	Full-time	**Part-time**	**Full-time**	**Part-time**	**Full-time**	**Part-time**	**Full-time**	**Part-time**
Worse	58%	44%	27%	20%	34%	22%	51%	44%
Neutral	16%	19%	23%	20%	30%	34%	24%	16%
Better	26%	38%	50%	60%	36%	44%	25%	40%

Similarly, part-time students generally assessed various aspects of learning performed in a distance mode better than full-time students (Table 4). Especially, they noticed the better comfort during the classes and did not perceive a deterioration in commitment as much as full-time students. Furthermore, students with high experience in traditional learning seemed to be neutral or in favor of all aspects of learning being performed in distance form.

Table 4. *The Influence of Distance Learning on the Various Aspects of Learning*

		Experience in a traditional form of learning			Form of study	
		L	**M**	**H**	**Full-time**	**Part-time**
Understanding of material	Worse	39%	53%	35%	43%	26%
	Neutral	42%	31%	42%	39%	45%
	Better	19%	17%	23%	18%	29%
Team work	Worse	58%	50%	44%	54%	38%
	Neutral	22%	22%	26%	22%	29%
	Better	20%	28%	30%	24%	32%
Interaction with the lecturer	Worse	57%	60%	52%	59%	41%
	Neutral	29%	23%	30%	27%	33%
	Better	14%	17%	18%	14%	25%
Interaction with other students	Worse	75%	72%	71%	77%	59%
	Neutral	17%	14%	18%	16%	23%
	Better	8%	14%	11%	7%	18%
Commitment during classes	Worse	54%	56%	49%	57%	37%
	Neutral	27%	23%	33%	27%	37%
	Better	18%	20%	18%	16%	26%
Feeling of comfort during classes	Worse	17%	24%	19%	21%	11%
	Neutral	15%	15%	15%	16%	13%
	Better	69%	61%	66%	63%	77%

L – Low; M – Medium; H - High experience in traditional learning

Table 5. *Effectiveness of Knowledge and Skills Verification in Distance Learning*

		Experience in a traditional form of learning			Form of study	
		L	**M**	**H**	**Full-time**	**Part-time**
Oral exam	Worse	25%	27%	28%	28%	23%
	Neutral	52%	50%	50%	50%	53%
	Better	23%	23%	21%	22%	25%
Multiple-choice test	Worse	26%	26%	27%	27%	23%
	Neutral	36%	40%	41%	40%	38%
	Better	38%	34%	32%	33%	40%
Exam with open questions	Worse	23%	21%	28%	26%	21%
	Neutral	46%	48%	46%	46%	46%
	Better	31%	31%	26%	28%	32%
Project and its presentation	Worse	13%	14%	12%	14%	8%
	Neutral	35%	32%	42%	37%	42%
	Better	52%	53%	46%	49%	49%
Written final work	Worse	14%	12%	15%	15%	13%
	Neutral	50%	45%	49%	49%	48%
	Better	36%	44%	36%	36%	39%

L – Low; M – Medium; H - High experience in traditional learning

Part-time students evaluate each form of knowledge and skills verification more positively than full-time students, among which the neutral evaluation dominates (Table 5).

5.2 RQ2: Suitability of Various Forms of Teaching for Distance Learning

More than 64% of highly experienced students and more than 70% of part-time students believed that all forms of teaching might be performed in the distance form apart from labs (about 20% and 28% respectively) (in Tables 6 and 7 we did not include the answer "I do not know" as it did not bring new content). Full-time students were less optimistic in this matter.

Table 6. *Suitability of Different Teaching Forms in Distance Learning According to Respondents with Different Experience in Traditional Learning*

	Workshops			Lectures			Seminars			Labs		
	L	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**
Unsuitable	47%	43%	35%	9%	9%	7%	12%	17%	16%	27%	34%	33%
"It is OK"	33%	33%	37%	26%	34%	21%	9%	21%	28%	9%	7%	7%
Optimal	16%	21%	27%	63%	55%	71%	15%	23%	45%	9%	9%	13%

L – Low; M – Medium; H - High experience in traditional learning

Table 7. *Suitability of Different Teaching Forms in Distance Learning According to Full-time and Part-time Students*

	Workshops		Lectures		Seminars		Labs	
	Full-time	**Part-time**	**Full-time**	**Part-time**	**Full-time**	**Part-time**	**Full-time**	**Part-time**
Unsuitable	46%	24%	8%	6%	16%	9%	31%	30%
"It is OK"	34%	36%	28%	14%	19%	21%	8%	7%
Optimal	17%	38%	61%	79%	25%	49%	8%	21%

5.3 RQ3: Students' Preferences Concerning Distance and Traditional Learning

About 20% of respondents would rather avoid distance learning in the future. Students with high experience in traditional learning prefer a continuation of distance learning (Figure 2a). Similarly, a greater share of part-time students than full-time students prefers a continuation of distance learning (Figure 2b). A similar conclusion might be drawn in relation to respondents' preferences towards the future form of studies, however in each subgroup hybrid learning dominates (Figure 3).

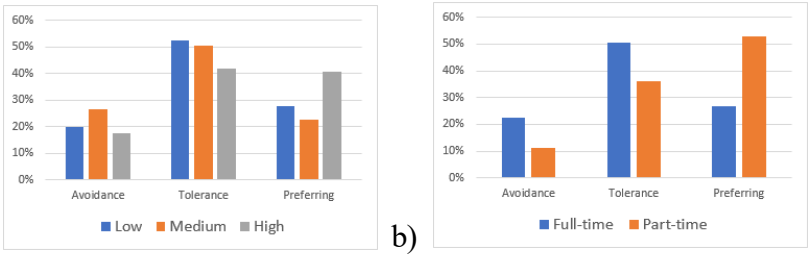


Figure 2. Attitude towards the Continuation of Distance Learning in Relation to (a) Experience with Traditional Higher Education and (b) Form of Studies.

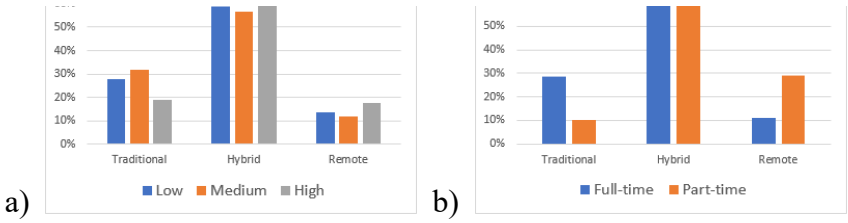


Figure 3. Students' Preferences Concerning Future Learning in Relation to (a) Experience with Traditional Higher Education and (b) Form of Studies.

5.4 RQ4: Students' Prediction of the Role of Distance Learning in Education

Respondents' predictions concerning the role of distance learning in education were diversified, with the domination of hybrid education (Figure 4). However, part-time students seem to be more inclined towards distance learning.

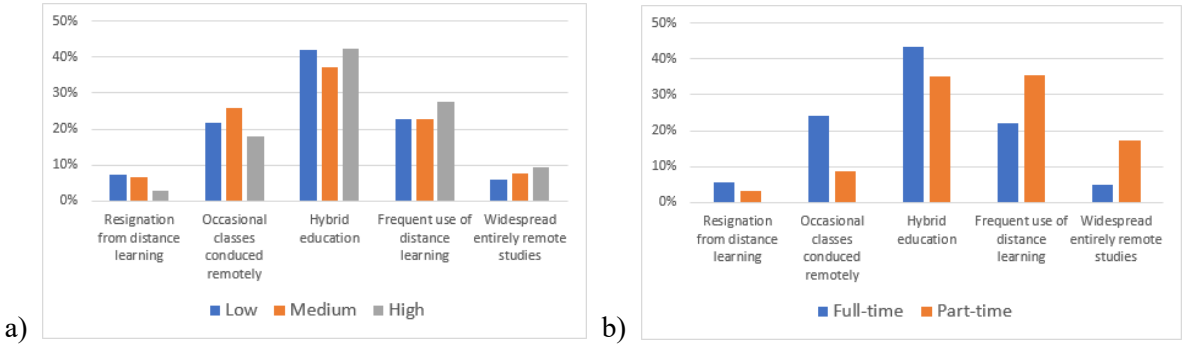


Figure 4. Students' Predictions of the Future of Learning in Relation to (a) Experience with Traditional Higher Education and (b) Form of Studies.

6. Discussion

The basic criteria for the data analysis were experience with learning in the traditional form (on premise) and the form of studies (full-time, part time). These criteria are not fully independent, since a large proportion of part-time students have extensive experience in traditional education, especially among those who study at master level. Such interdependence is visible in some convergence of the results for students with extensive experience in traditional learning and part-time students. There is no decrease in wellbeing among part-time students, on the contrary, an improvement is noticeable; the difference from full-time students is clearly visible. This might be caused by the following factors: (1) their work – part time students usually have an additional group in social contacts (work colleagues) and have activities other than studies; (2) global overload - studies in the case of part-time students require more effort and resources than in the case of the full-time variant (weekend meetings, the entire weekend at the university), hence the remote form of studies might be perceived by them as a benefit - a kind of rest for them.

Among the types of classes in which the students participated, all distinguished groups of respondents agreed that "lectures" in the distance form are a better solution compared to the traditional form. In the case of activities requiring more personal interaction, such as "workshops" or "labs", distance classes were rated as "worse". The assessments of "seminars" are more varied, as this form of learning requires a more individualized form of cooperation between students and teachers; the differences in the assessment of seminars might result from various experiences and preferences concerning individual contacts. Taking into account various aspects of learning, in all the groups, in all aspects except the "feeling of comfort during classes", the "neutral" and "worse" assessments dominated; this applies in particular to the aspects of teaching that are based on the interaction among students or between students and teachers. This confirms the previous assessment with regard to the forms of classes. It can also be noted that although the lectures in the distance form were considered optimal, their role is not dominant in the educational process since a high percentage of students perceived distance learning as "worse" in the context of understanding the material. Against this background, the assessment in the aspect of "feeling of comfort during classes" indicates that participation in a traditional form of classes is associated with additional stress for many students, which may be caused by a stronger pressure from both the teachers and other students for proper behavior, attitude, or activity in class. In the assessment of the forms of skills and knowledge verification, for all assessed forms, the "neutral" rating dominates, with a predominance of "better" over "worse" ratings. This can be caused by a greater sense of comfort during the distance form of knowledge verification, which allows students to focus more on the test/exam content and not to be distracted by the necessity of additional requirements of the traditional form, such as dress-code or posture.

The assessment of the suitability of various forms of classes for distance learning coincides with the students' opinions about these classes, in which the students participated, conducted in the distance form. We can notice a very positive assessment of distance lectures, for which, in the opinion of the majority of respondents, the distance form is optimal. Taking into account the students' opinions on the different forms of teaching, we can conclude that maybe some other forms of teaching should be developed that might fit more to distance learning, which has been already suggested in previous studies (Krishnamurthy, 2020; Ackerman & Gross, 2021). New technologies, such as virtual or augmented (mixed) reality open up new possibilities of interactivity, and thus might eliminate the lowest-rated aspects of distance learning.

Part-time students more often than full-time students indicate the distance form of learning as their preference. At the same time, it should be noted that a significant percentage of students would try to avoid this form of classes, which indicates that it is not in line with their expectations or preferences. Most respondents foresee the hybrid form of learning as optimal (some remote classes, some on premise). Taking into account their previous opinions concerning the form of classes, it seems that the form should be adjusted to the type of classes. However, it should be noted that there are two extreme opinions concerning the future studies: there are students who would prefer only one form of learning, traditional or distant.

The main limitation of the study is sometimes using only one survey item to gather students' opinions on some phenomena, e.g. wellbeing. However, when designing the survey we had to have in mind the potential respondents, who are usually not willing to fill in overly long surveys. Besides, we gathered data only from one university, therefore any conclusions should be drawn with caution.

7. Conclusion

The research conducted among the students of the Cracow University of Economics, Poland, shows that although we can generally speak of a positive attitude towards remote education imposed by the pandemic, a significant group (about 20%) of students intend to avoid this form of education in the future. There is also a visible difference in the opinions between full-time and part-time students, which may result from their different stages of life development. While full-time students are mostly still at the formal educational stage, part-time students have mostly finished this stage, started working, and in their case studies should be included in the lifelong learning category (adult learning). It also means that these two groups differ not only in their motivation and commitment to the learning process, but (e.g. due to their age) have more precise expectations based on their own experiences (they know what they

want), they are also more independent and disciplined, but have a more limited time capacity (they often have to combine full-time work with studying).

Educational institutions must find themselves in this rather complex situation. They should not exclude anyone from access to education, which means that their offers should meet the expectations of both those who prefer on premise forms and those positively oriented towards distance forms. With regard to the types of studies, it can be suggested that universities should expand the offer of fully or partially remote studies in the form of part-time studies. In full-time studies, however, there is room to increase the role of distance learning, but this should not be in the form of resignation from on premise classes. The optimal solution would be to organize the teaching process in such a way that a student can choose the form of realization of individual classes, tailored to his/her preferences and abilities. This may cause a significant number of organizational and administrative problems and may also affect the costs of education and put more demands on teachers (one subject delivered in two forms), but in the longer term, taking steps to implement distance learning adjusted to individual needs of students may determine the future of educational institutions.

Acknowledgments

The Project has been financed by the Ministry of Science and Higher Education within the "Regional Initiative of Excellence" Programme for 2019-2022. Project no.: 021/RID/2018/19. Total financing: 11 897 131.40 PLN.

References

Ackerman, D. S., & Gross, B. L. (2021). Synchronous online discussion board as a primary mode of delivering marketing education: Responding to the COVID-19 pandemic and beyond. *Marketing Education Review*, *31*(4), 284-293.

Afrianty, T. W., Artatanaya I. G., & Burgess J. (2022). Working from home effectiveness during COVID-19: Evidence from university staff in Indonesia. *Asia Pacific Management Review*, *27*(1), 50-57.

Agarwal, A., Bansal, K. M., Gupta, A., & Khurana, R. (2021). Online education: A booming product for institutes post-COVID-19?. *Marketing Education Review*, *31*(3), 262-272.

Boyer-Davis, S. (2020). Technostress in higher education: An examination of faculty perceptions before and during the COVID-19 pandemic. *Journal of Business and Accounting*, *13*(1), 42-58.

Code, J., Ralph, R., & Forde, K. (2020). Pandemic designs for the future: perspectives of technology education teachers during COVID-19. *Information and Learning Sciences*, *121*(5/6), 419-431.

Conrad, R. C., Hahm, H. C., Koirea, A., Pinder-Amaker, S., & Liu, C. H. (2021). College student mental health risks during the COVID-19 pandemic: Implications of campus relocation. *Journal of Psychiatric Research, 136*, 117-126.

Dymek, D., Grabowski, M., & Paliwoda-Pękosz, G. (2021). Change of students' online behavioural patterns during the COVID-19 pandemic: Insights from Poland. In: Soliman (Eds.), *Proceedings of the International Business Information Management Association Conference*, *38*, (pp. 3263–3272), Seville, Spain.

Helyer, R., & Lee, D. (2014). The role of work experience in the future employability of higher education graduates. *Higher Education Quarterly*, *68*(3), 348-372.

Herbert, C., El Bolock, A., & Abdennadher, S. (2021). How do you feel during the COVID-19 Pandemic? A survey using psychological and linguistic self-report measures, and machine learning to investigate mental health, subjective experience, personality, and behaviour during the COVID-19 pandemic among university students. *BMC Psychology*, *9*(1), 1-23.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396-403.

Krishnamurthy, S. (2020). The future of business education: A commentary in the shadow of the COVID-19 pandemic. *Journal of Business Research*, *117*, 1-5.

Lee, J., Jeong, H. J., & Kim, S. (2021). Stress, anxiety, and depression among undergraduate students during the COVID-19 pandemic and their use of mental health services. *Innovative Higher Education*, *46*, 519-538.

Liu, C., McCabe, M., Dawson, A., Cyrzon, C., Shankar, S., Gerges, N., Kellett-Renzella, S., Chye, Y, & Cornish, K. (2021). Identifying predictors of university students' wellbeing during the COVID-19 pandemic - A data-driven approach. *International Journal of Environmental Research and Public Health*, *18*(13), 1-10.

Meshram, K., Paladino, A., & Cotronei-Baird, V. S. (2022). Don't waste a crisis: COVID-19 and marketing students' self-regulated learning in the online environment. *Journal of Marketing Education*, January, 1-23.

Naddeo, A., Califano, R., & Fiorillo, I. (2021). Identifying factors that influenced wellbeing and learning effectiveness during the sudden transition into eLearning due to the COVID-19 lockdown. *Work*, *68*(1), 45-67.

Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2021). Balancing technology, pedagogy and the new normal: Post-pandemic challenges for higher education. *Postdigital Science and Education*, *3*, 715-742.

Regmi, K. D. (2015). Lifelong learning: Foundational models, underlying assumptions and critiques. *International Review of Education*, *61*(2), 133-151.

Rippé, C. B., Weisfeld-Spolter, S., Yurova, Y., & Kemp, A. (2021). Pandemic pedagogy for the new normal: Fostering perceived control during COVID-19. *Journal of Marketing Education*, *43*(2), 260-276.

Spais, G., & Paul, P. (2021). A crisis management model for marketing education: Reflections on marketing education system's transformation in view of the COVID-19 crisis. *Marketing Education Review*, *31*(4), 322-339.

Stewart, W. H., & Lowenthal, P. R. (2021). Distance education under duress: A case study of exchange students' experience with online learning during the COVID-19 pandemic in the Republic of Korea. *Journal of Research on Technology in Education*, *54*(sup1), 1-15.

Vander Schee, B. A., & Birrittella, T. D. (2021). Hybrid and online peer group grading: Adding assessment efficiency while maintaining perceived fairness. *Marketing Education Review*, (31), 275-283.

Appendix

1. Gender: male; female; prefer not to say
2. Study level: undergraduate; MA studies; doctoral studies
3. Form of studies: full time; part time
4. Year of study: first; second; third; fourth; fifth or higher
5. Age: under 20; 20 - 24; 25 - 30; 31+
6. I use information and communication technologies/tools at the level of: beginner; below average; average; above average; professional
7. My technical conditions (access to the Internet and equipment: PC, laptop, tablet, smartphone, etc.) I would describe as: I have constant problems; I have frequent problems; Sufficient for basic needs; I have occasional problems; I do not have any problems
8. How has your wellbeing changed after switching to distance learning? (Significantly deteriorated; worsened; no change; improved; significantly improved).
9. What is your attitude towards distance learning (not only organized forms but also e.g. training videos on YouTube)? (I don't like it and try to avoid/ I don't like it but I use it/neutral/ I prefer it/ I definitely prefer it); Before the pandemic | During the pandemic
10. Evaluate the effectiveness of the given forms of distance learning activities (workshops/lectures/seminars/laboratories/consultations) in relation to traditional learning on a scale (did not participate/unsuitable/slightly worse/same effectiveness/slightly better/optimal)
11. Which form of classes (workshops/lectures/seminars/laboratories/consultations) could be conducted remotely, in your opinion, after the pandemic is over? (Evaluate each of them on the scale: I don't have an opinion/unsuitable/"it's ok"/preferred form)
12. What do you think will be the trend in forms of education after the end of the pandemic? I think it will be (complete abandoning of the remote form of learning/frequent use of the distance learning form/hybrid form of learning/very frequent use of distance learning/spreading the popularity of using fully distance learning courses)
13. In your opinion, what form of studies will be preferred by future students after the pandemic is over? (Traditional (face-to-face classes with the occasional use of distance learning)/hybrid (face-to-face classes with significant use of distance learning)/remote (full studies using only distance learning))
14. In your opinion, how does the remote form of classes affect the quality of education (in the field of: understanding of the material/teamwork/interaction with the lecturer/interaction with other students/commitment during classes/feeling of comfort during classes). Rate it on the scale (much worse/worse/no difference/better/much better)
15. In your opinion, how do remote methods of completing the final exams (oral exam/multiple-choice test/exam with open questions/project and its presentation/written final work) affect the effectiveness of the verification of knowledge and skills? Rate on the scale (unsuitable/ decreases/no difference/improves/optimal)

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Media Exposure and Students' Attitude as Mediators Between Subjective Norms and Choice Intention for Science, Technology, Engineering, and Mathematics Careers

Tiny Chiu Yuen TEY[a,b*], Priscilla MOSES[a] & Phaik Kin CHEAH[a]
[a]*Universiti Tunku Abdul Rahman, Malaysia*
[b]*University of Sanya, China (current affiliation)*
*ttcyuen456@gmail.com

Abstract: Many countries have taken various forms of efforts and initiatives to attract more students to join the science, technology, engineering, and mathematics (STEM) workforce, but shortage of STEM talents remains a worldwide challenge. Concerning this issue, previous studies have frequently discussed the roles of parents, teachers, and peers (subjective norms), as well as media exposure and students' attitude as predictors of students' career choice in STEM. However, there are few studies that examined the mediating roles of media exposure and students' attitude in the relationship between subjective norms and career choice intention. It was indicated in previous studies that there could have been a shift in students' STEM choices because there are changes in their attitude and exposure to the media alongside technology advancement. Hence, this study aimed to determine the mediating roles of media exposure and students' attitude in the relationship between subjective norms and STEM career choice intention in addition to examining the direct causal relationships among the constructs. The hypothesised model was tested with 806 secondary school students in Malaysia. Results from partial least squares structural equation modelling found that subjective norms had both direct and indirect influence on career choice intention. Another highlight of this study is that media exposure and students' attitude were significant mediators between subjective norms and career choice intention. The findings drawn from this study provide insights into the design and development of STEM initiatives for parents, teachers, and students in Malaysia.

Keywords: Subjective norms, media exposure, students' attitude, career choice intention, mediator

1. Introduction

The science, technology, engineering, and mathematics (STEM) fields have been growing rapidly across the globe. STEM is often related to innovation and growth of a country because a high-capacity STEM workforce is anticipated to boost the country's competitiveness and productivity (Shin et al., 2018). Many countries have been investing heavily in the STEM fields as STEM workers are in high demand globally (Dockery et al., 2021; Shin et al., 2018). According to Li et al. (2021), shortage of competent STEM talents has been a worldwide challenge and has raised concerns if the existing education systems are compatible with the demand of STEM industries. It was mentioned that effective incorporation of STEM components to the respective education systems can be a solution to this global issue (Navy et al., 2020).

Warne et al. (2019) revealed that preparing middle school students for STEM careers is indeed one of the most critical educational challenges in the United States. The educational issue is pressing in the United States not only because it is important to ensure constant economic growth and technological innovation at large, but also to fill vacancies in the STEM workforce (Warne et al., 2019; Wyss et al., 2012). This is not uncommon in the Asian contexts as changes in national policies have been attempted in view of the challenges in Asia as well. In developing Asian countries like Indonesia, the government offered lucrative financial incentives for STEM employees to encourage more Indonesian students to pursue careers in STEM (Shin et al., 2018). Shin et al. (2018) also reported that Korea has implemented

a national policy for STEM enhancement to attract more students to STEM fields, but it was only effective in the short term.

In Malaysia, STEM reformation has also been done to encourage students' participation in STEM. One of the most substantial initiatives was the implementation of the new secondary school national curriculum. Alongside the implementation of the new national curriculum, STEM subjects have been made eligible to all upper secondary school students (Year 10 and 11) nationwide (Curriculum Development Division, 2016). The STEM-oriented curriculum aims to create STEM awareness among all upper secondary school students and to prepare them for STEM careers. STEM subjects offered in Malaysian schools are such as Computer Science, Invention, and Technical Graphic Communication (Curriculum Development Division, 2016).

In view of immense attempts to encourage students' participation in STEM, Li et al. (2021) indicated that there could have been a shift in perception of how students perceive STEM careers in tandem with technology advancement. It was suggested that media and students' close contacts could be particularly important for their STEM career decisions (Li et al., 2021). Besides, Wyss et al. (2012) suggested that adolescents' attitude and perceptions towards STEM are associated with their exposure to media. Batool and Ghayas (2020) also found that adolescents' career choice can be affected by multiple factors such as parents, teachers, peers, and media.

Findings from past research probes the authors' curiosity: (i) what is the influence of subjective norms (parents, teachers, and peers), media exposure and students' attitude towards STEM career choice intention? and (ii) do media exposure and students' attitude mediate the relationship between subjective norms and career choice intention? Based on the research questions, the purpose of the study was to examine the influence of subjective norms (parents, teachers, and peers), media exposure, and students' attitude on their STEM career choice intention. Besides, this study also aimed to investigate the mediating roles of media exposure and students' attitude between subjective norms and career choice intention in STEM fields.

2. Literature Review

According to Chen et al. (2020), students' career choice intention is an indication of their job pursuit after they complete their education. Mokhtar et al. (2016) proposed that students' career intention can be determined through factors such as subjective norms and attitude as presented in the Theory of Planned Behaviour. In this theory, subjective norms refer to an individual's perceived social pressure from significant others to perform or avoid a specific behaviour (Ajzen, 2002). Based on the review of literature, it was noticed that the most frequently mentioned significant others were parents (Razali, 2021; Sheehan et al., 2018), teachers (Avargil et al., 2020; Kong et al., 2020), and peers (Batool & Ghayas, 2020; Raabee et al., 2019).

This is consistent with findings reported by Hoag et al. (2017) that students regarded parents, teachers, and peers as significant others who had influence over their choices and decisions. Mohtar et al. (2019) also found that parents, teachers, and peers were students' close contacts who offered support to students, and they could ultimately affect students' career choice in the STEM fields. More specifically, it was found in previous research that parental support was significant for STEM career choice among Indonesian students (Shin et al., 2018). It was also explained in Sheehan et al. (2018) that the family environment provided early exposure to STEM concepts, hence promoting their engagement with STEM from young at home. Besides, Avargil et al. (2020) suggested that teachers could also create awareness about career opportunities and increase students' likelihood to pursue careers in the respective taught subjects. Similar finding is also reported in Batool and Ghayas (2020) that students considered teachers as role models whose encouragement and comments could affect their future career paths. In the same study, students' career choice was influenced by peers (Batool & Ghayas, 2020). Rabee et al. (2019) explained that while the roles of family members were important, the roles of peers also become more crucial as children grow older.

Besides, Gómez-Galán (2020) suggested that technologies and media do not only change how students learn, but also how they develop habits and understanding about their environment. This is because students in primary and secondary education today are digital natives in which ICT has an important part in developing their career paths (Ottestad & Gudmundsdottir, 2018). It was reported that students learnt about STEM from the media (Sheehan et al., 2018), and STEM media consumptions played a vital role in teenage students' career aspirations (Chen et al., 2022). Batool and Ghayas (2020)

also explained in their study that the media could affect students' preferences towards careers as how they were portrayed in media such as television shows, dramas and songs. Li et al.'s (2021) research also found that students believed they would be able to pursue STEM careers when they received positive information regarding STEM careers via social media. Additionally, Kettunen et al. (2020) described the Internet as the first contact point for students to explore career opportunities. This is supported by Hasin and Nasir (2021) that the Internet was important for students not only for learning, but also to seek information about careers. Students are indeed very resourceful that they can also retrieve occupational information from various types of media. These emerging adults use the media, especially the Internet, to gather career information and prepare for future careers (Levine & Aley, 2020). Kricorian et al. (2020) also revealed that students would take the initiative to follow STEM news on social media or websites and watch movies about STEM careers which they perceived useful.

In addition, Badri et al. (2016) suggested that students' willingness to engage in STEM career pathways is associated with their attitude. This is consistent with the Aziz et al.'s (2020) study which found a significant positive relationship between attitude and intention. Attitude is an indication of favourable or unfavourable judgement and individuals were more inclined to choose a career which they had more favourable judgement (Aziz et al., 2020). James et al.'s (2018) reported that students who showed a positive attitude towards the pharmacy career had greater intention for careers related to the field. A similar finding was also found in Solikhah (2014) that attitude had significant influence on students' career choice intention in accounting context.

As reviewed earlier, many past studies confirmed that subjective norms (parents, teachers, and peers) had a direct influence on students' career choice intention. On top of that, it was also noticed that media exposure and attitude could have mediating effects between subjective norms and career choice intention. For example, Batool and Ghayas (2020) pointed out that students did not follow parents' advice uncritically although students perceived them as influential significant others who were wise and reliable. Steinke et al. (2022) also emphasised in their study that information on the media and online resources are seemingly more accessible than in real life for students. Therefore, teachers were recommended to use the media to introduce STEM professions and highlight the achievements of the professionals to reinforce students' sense of STEM belonging (Kricorian et al., 2020). Moreover, Aziz et al.'s (2020) study also confirmed that attitude had a significant mediating effect between subjective norms and intention. This finding is in line with the finding of Al-Swidi et al. (2014) that subjective norms had both direct and indirect effects on intention, and attitude was a significant mediator between subjective norms and intention.

The research model is presented in Figure 1. The hypotheses of this study were established based on the research objectives and review of literature:

H_1: Subjective norms significantly influence media exposure.
H_2: Subjective norms significantly influence students' attitude.
H_3: Subjective norms significantly influence career choice intention.
H_4: Media exposure significantly influences career choice intention.
H_5: Students' attitude significantly influences career choice intention.
H_6: Media exposure mediates the relationship between subjective norms and career choice intention.
H_7: Students' attitude mediates the relationship between subjective norms and career choice intention.

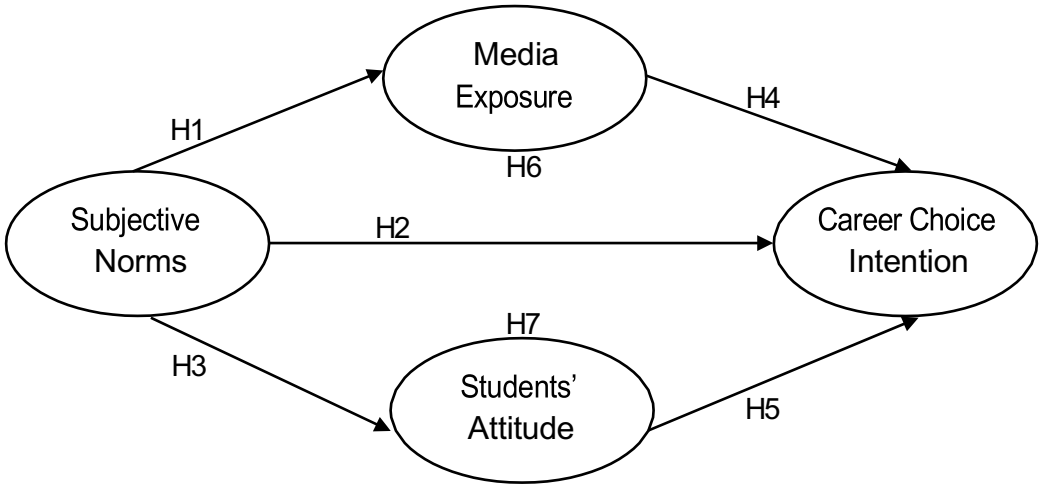


Figure 1: Research model with direct and indirect relationships.

3. Methods

Prior to data collection, permissions were sought from the Malaysian Ministry of Education, and state educational departments, school authorities and students' guardians which were involved in this study. Data was collected using a bilingual online questionnaire which students were given the options to answer in either English or the national language of Malaysia.

Three experts in the research areas were invited to review the questionnaire for its face and content validity. This step helped the researcher to ensure the survey items represent the aspects of examination aligned with the research objective. Besides, another three language experts were involved in the translation and back translation process to ensure translation accuracy as well as to reduce reliability and validity issues related to the questionnaire (Brace, 2018). The questionnaire consisted of 32 adapted and self-developed items to meet the objective of this research. The measures adapted for this study are listed in Table 1. In this questionnaire, respondents indicated their responses on a five-point Likert scale, from strongly disagree to strongly agree.

Table 1. *Referenced Sources for Questionnaire Development*

Constructs	Items	Referenced sources
Subjective Norms	15	Ajzen (2002); Francis et al. (2004)
Media Exposure	7	Hoag et al. (2017); Qader & Zainuddin (2011)
Students' Attitude	6	Ajzen (2002); Francis et al. (2004)
Career Choice Intention	4	Ajzen (2002); Francis et al. (2004)

As a result, a total of 806 complete responses were received from upper secondary school students from 13 states in Peninsular Malaysia. The respondents were all Form Four students (equivalent to Year 10) who were typically 16 years old. The respondents were recruited using proportional stratified sampling approach, hence 44% (f=354) of them were from STEM stream whereas 56% (f=452) were from non-STEM stream. More than half (60.8%) of the respondents were female, while the remaining 39.2% of them were male.

4. Results

Structural Equation Modelling (SEM) was used in this study to examine the relationships of the constructs as proposed in the research model. Hence, a two-stage analytical procedure was conducted using the SmartPLS 3.2.8 software by Ringle et al. (2015). Through this approach, reliability and

validity can be established with a measurement model, whereas hypotheses can be tested with a structural model.

4.1 Assessment of Measurement Model

Construct reliability of the model was assessed with Cronbach's alpha (α), composite reliability (CR) and the average variance extracted (AVE). According to Hair et al. (2017), the recommended threshold values for α and CR indices are 0.70 and 0.50 for AVE respectively. Table 2 shows that the α values of the constructs range from 0.808 to 0.948, while CR values range from 0.858 to 0.962, hence verifying the internal consistency of the constructs. In the same table, it was shown that AVE values of the construct are larger than the recommended threshold of 0.50 except media exposure (CR=0.465). According to Malhotra (2010), AVE is a stringent measure of convergent validity. This is also supported by Fornell and Larcker (1981) that convergent validity of a construct is still adequate when AVE is below 0.50, but it is acceptable when CR is above 0.60. As the CR value of media exposure was above 0.60 (CR=0.858), the convergent validity of the construct was established.

Table 2. *Measurement Model Assessment*

Constructs		α	CR	AVE	HTMT			
					1	2	3	4
1	Subjective Norms	0.948	0.953	0.577				
2	Media Exposure	0.808	0.858	0.465	0.451			
3	Students' Attitude	0.881	0.910	0.627	0.728	0.447		
4	Career Choice Intention	0.947	0.962	0.862	0.722	0.455	0.753	

At this stage, outer loadings of the constructs were also assessed. Hair et al. (2017) suggested that it is ideal to obtain all loadings larger than 0.708, but it is also common to see outer loadings smaller than 0.70 especially in social science research. It was suggested to retain the items for content validity if they load between 0.40 and 0.70, unless the AVE and CR values are increased after removing the items (Hair et al., 2017; Wong et al., 2016). Based on the results, outer loadings of the items ranged from 0.612 to 0.941 (>0.40), and the values of AVE and CR did not increase after removing items that loaded <0.70. Therefore, all items in the measurement model were retained.

Discriminant validity of the constructs was verified with Heterotrait-Monotrait Ratio of Correlations (HTMT) (Henseler et al., 2015). HTMT ratio with values above 0.85 indicates issues of discriminant validity, whereas values below 0.85 suggests the constructs are distinctively different from one another (Hair et al., 2017; Henseler et al. 2015). As presented in Table 2, HTMT values of the constructs range between 0.447 and 0.753 (<0.85). Hence, the constructs were distinctively different from one another, and discriminant validity was not a threat in this study.

4.2 Assessment of Structural Model

In the assessment of the structural model, the Variance Inflation Factor (VIF) was calculated to examine multicollinearity issues. It was found that the highest VIF was 2.238 (VIF<5), indicating that there was no multicollinearity issue among the exogenous constructs as the VIF values were below the recommended cut-off (Hair et al., 2017). In Table 3, the results indicated that all path coefficients were significant because the t-values were above 2.58 ($p<0.01$), and there was no zero straddled between the 95% confidence intervals.

Table 3. *Structural Model Assessment and Hypothesis Testing*

Direct Paths		β	SD	t-value	p-value	BC LL	BC UL	Result
H1	SN➔ME	0.467**	0.036	13.003	0.000	0.387	0.532	Supported
H2	SN➔ SA	0.732**	0.022	32.84	0.000	0.686	0.773	Supported
H3	SN➔CCI	0.447**	0.046	9.727	0.000	0.686	0.773	Supported
H4	ME➔ CCI	0.082*	0.030	2.766	0.006	0.025	0.141	Supported
H5	SA➔ CCI	0.392**	0.047	8.28	0.000	0.298	0.486	Supported

Note: SN: $p<0.01$*, $p<0.001$**. t-value>2.58 for $p<0.01$; t-value>1.96 for $p<0.05$. BC LL: Bias corrected lower level (5%), BC UL: Bias corrected upper level (95%). VIF<5.

PLS-SEM was employed to test the hypotheses and significance of path coefficients (β) was assessed with a bootstrapping setting of 5000 resamples. Table 3 shows that all hypotheses proposed in this study were supported by the results with $p<0.01$. In specific, subjective norms had a positive significant influence on students' attitude (β=0.732, $p< 0.001$), media exposure (β=0.467, $p<0.001$), and career choice intention (β=0.447, $p< 0.001$). The results thus supported H1, H2 and H3. Besides, H4 and H5 were also supported by the results shown in Table 3. Career choice intention was found to be significantly influenced by students' attitude (β=0.392, $p<0.001$) and media exposure (β=0.082, $p<0.01$).

In addition to the assessment of path coefficients, this study also evaluated the coefficient of determination (R^2), predictive relevance (Q^2), and effect size (f^2) of hypothesised causal relationships. Based on the results in Table 4, results suggested that 68.2% (R^2=0.682) of the variance of career choice intention can be explained by the research model of this study. It was also found that the research model explained 53.6% (R^2=0.536) of students' attitude variance and 21.8% (R^2=0.218) of media exposure variance.

Table 4. *Coefficient of Determination (R^2), Predictive Relevance (Q^2), and Effect Size (f^2)*

Constructs	R^2	Q^2	f^2			
			Media Exposure	Students' Attitude	Career Choice Intention	
Media Exposure	0.218	0.068			0.016	
Students' Attitude	0.536	0.263			0.172	
Career Choice Intention	0.682	0.499				
Subjective Norms			0.202	0.324	0.289	

Stone-Geisser's Q^2 index was also used to determine the predictive relevance of the model. Hair et al. (2017) noted that Q^2 values above zero indicate that the exogenous constructs have predictive relevance towards the respective endogenous constructs. Based on the results in Table 4, the Q^2 values ranged from 0.068 to 0.499 (Q^2>0), hence the predictive relevance of the exogenous constructs towards the endogenous constructs (media exposure, students' attitude, and career choice intention) was established. Besides, the effect sizes of the paths were also calculated with f^2. Cohen (1988) recommended that the magnitude effect size is considered small when f^2=0.02, medium when f^2=0.15, and large when f^2=0.35. As presented in Table 4, all paths had medium to large effect (f^2=0.172-0.324) except for media exposure to career choice intention which suggested small effect with f^2=0.016.

4.3 Mediation

In line with the purpose of this study, the mediating relationships of the research model were assessed using bootstrapping procedure on SmartPLS (Zhao et al., 2010). As shown in Table 5, H6 proposed that media exposure is a mediator between subjective norms and career choice intention, whereas H7 hypothesised students' attitude as the mediator between subjective norms and career choice intention. Based on the results in mediation analysis, it was found that both students' attitude (β=0.287, $p<0.001$) and media exposure (β=0.038, $p<0.01$) were significant mediators between subjective norms and career

choice intention. Since subjective norms had significant direct influence on career choice intention with β=0.447, it can be concluded that students' attitude and media exposure had partial mediating effects in the relationship between subjective norms and career choice intention in this model.

Table 5. *Mediation Effect Testing*

Indirect Paths		Indirect β	SD	t-value	p-value	BC LL	BC UL	Direct β	Result
H6	SN➔ME➔CCI	0.038*	0.015	2.629	0.009	0.012	0.069	0.447**	Partial
H7	SN➔SA➔CCI	0.287**	0.036	7.891	0.000	0.217	0.361		Partial

Note: SN: p<0.01*, p<0.001**. t-value>2.58 for p<0.01; t-value>1.96 for p<0.05. BC LL: Bias corrected lower level (5%), BC UL: Bias corrected upper level (95%). VIF<5.

5. Discussion

In general, all proposed hypotheses of this study were supported by the results. The results demonstrated that subjective norms had a positive significant influence on media exposure, students' attitude and career choice intention in STEM, hence supporting H1, H2, and H3. In other words, parents, teachers, and peers were significant others who had direct influence over student's media exposure, attitude, and intention to pursue careers in STEM. This finding is consistent with those from Hoag et al. (2017) and Bergin (2016) that parents, teachers, and peers play the biggest roles in students' career choice intention, and they develop students' exposure and shape their preferences at school and at home. This finding is also aligned with Mohtar et al. (2019) that parents, teachers, and peers' support and encouragement would increase their intention to opt for STEM careers in future.

Results also supported the proposition that students' career choice intention in STEM can also be predicted by media exposure and attitude (H4 and H5). This finding is indicative of the importance of students' media exposure and attitude in their intention to choose a career in STEM. Similar to the findings from Gómez-Galán (2020), this study confirms that the power of media was beyond teaching and learning that students' exposure to media also developed their understanding about the environment. This can be explained by Chen et al.'s (2022) that teenage students' consumptions of STEM media increased their career aspirations in STEM fields. In the context of this study, media exposure in general refers to the opportunities for a secondary school student to attain information from different types of media. This finding implies that students explored information about STEM careers via various forms of media such as the Internet, social media platforms, and television contents. Besides, the finding of this study concurs with Badri et al. (2016) and James et al. (2018) that students' attitude had significant influence on students' STEM career choice intention. This means that students were more likely to choose a STEM career when they showed greater preference and judgement to STEM careers.

Interestingly, this study also confirmed that both media exposure and students' attitude were significant mediators with partial mediating effects in the relationship between subjective norms and career choice intention, thus supporting H6 and H7. Recent studies had suggested media exposure (Steinke et al., 2022) and attitude (Aziz et al., 2020) as the mediators between subjective norms and career choice intention. Similar to these studies, finding of this research suggests that parents, teachers, and peers had direct influence on students' career choice intention and the influence could also take place via media and attitude. This finding could imply that students took significant others' encouragement and advice while considering STEM careers, but they did not follow their advice uncritically. Instead, students also took into considerations the information they obtained from the media and their personal evaluations to decide whether they are likely to choose a career in STEM.

6. Conclusion

Theoretically, this study offers new insights into the influence of subjective norms on STEM career choice intention as well as the mediating roles of media exposure and students' attitude between the

aforementioned variables. Findings of this study could be used as the latest reference for the authorities and policy makers to support the initiatives for STEM workforce in in the Malaysian context by emphasising the importance of subjective norms, media exposure, and students' attitude.

Based on the findings, parents, teachers, and peers should also be included in STEM initiatives to effectively pave students towards STEM careers. This is to ensure students' significant others are also aware of the importance of STEM and be aware of their roles for students' STEM pursuits, as well as the opportunities in STEM industries. The roles of parents, teachers, and peers are important because they can also affect students' evaluation and judgement towards STEM professions and ultimately influence their intention to participate in STEM. Therefore, parents, teachers, and peers should also be involved in the planning and implementation of STEM initiatives.

Indeed, the existing STEM initiatives such as the revamp of the national curriculum could bridge the boundaries between academic contents and vocational implications for students (Ottestad & Gudmundsdottir, 2018). Further enhancements can be considered based on the findings of this study. For example, classroom or non-academic activities can include video interviews of STEM professionals who are able to provide students with accurate information and real-life stories about STEM careers (Wyss et al., 2012). This is because even though today's students are familiar with the use of media, they might not know how to use these resources to obtain accurate information for career planning purposes (Ottestad & Gudmundsdottir, 2018). Hence, when they are exposed to media for STEM-related information under the guidance of teachers whom they perceive to be more knowledgeable and reliable, they are more likely to develop likelihood for STEM careers with accurate understanding of the professions. This finding is potentially applicable to other contexts because students today are digital natives who depend on media for career information acquisition regardless of nationalities.

There are several limitations in this study. The data was collected using a self-report online survey; hence this could have caused issues on common method variance. Besides, findings of this study can only be generalised to Form Four (Year 10) students. This study was planned to involve all upper secondary students, but the scope of study was limited to only Form Four students because Form Five (Year 11) students were preparing for national exams during the research period. In terms of locality issues, the research findings of the present study can only represent students in Peninsular Malaysia because East Malaysia which comprises Sabah, Sarawak, and the Federal Territory of Labuan were excluded from this study due to budget and time restrictions.

Future researchers could expand the scope of study by including both Form Four and Form Five students from Peninsular and East Malaysia for greater generalisation of research findings. It would be meaningful to involve Form Five students in this study because they are generally more mature and have clearer ideas about STEM career choices since they are close to completion of secondary education. Future studies could also apply the research model presented in this study in other developed countries or western cultures to examine if the results remain the same in other contexts. Researchers are also recommended to use a longitudinal approach in future studies. This approach will help researchers to better understand the underlying explanations of the causal and mediating relationships among constructs. A longitudinal approach will also allow researchers to observe changes in students' responses over a longer period because students' knowledge, exposure and experience will develop as they grow older (Shin et al., 2018).

Acknowledgements

We are thankful to Universiti Tunku Abdul Rahman, Malaysia for supporting this research project through the Universiti Tunku Abdul Rahman Research Fund and Universiti Tunku Abdul Rahman Research Scholarship Scheme.

References

Ajzen, I. (2002). *Constructing a TPB questionnaire: Conceptual and methodological considerations*. https://pdfs.semanticscholar.org/0574/b20bd58130dd5a961f1a2db10fd1fcbae95d.pdf

Al-Swidi, A., Huque, S. M. R., Hafeez, M. H., & Shariff, M. N. M. (2014). The role of subjective norms in theory of planned behavior in the context of organic food consumption. *British Food Journal, 116*(10), 1561-1580. https://doi.org/10.1108/BFJ-05-2013-0105

Avargil, S., Kohen, Z., & Dori, Y. J. (2020). Trends and perceptions of choosing chemistry as a major and a career. *Chemistry Education Research and Practice, 21*(2), 668-684. https://doi.org/10.1039/C9RP00158A

Aziz, S., Afaq, Z., Muhammad, L., & Khan, B. (2020). The role of media, word of mouth, and subjective norms in determining attitude and intentions to purchase family takaful schemes. *Journal of Islamic Business and Management, 10*(1), 111-131. https://doi.org/10.26501/jibm/2020.1001-008

Badri, M., Alnuaimi, A., Mohaidat, J., Al Rashedi, A., Yang, G., & Al Mazroui, K. (2016). My science class and expected career choices — A structural equation model of determinants involving Abu Dhabi high school students. *International Journal of STEM Education, 3*(1), 1-21. https://doi.org/10.1186/s40594-016-0045-0

Batool, S. S., & Ghayas, S. (2020). Process of career identity formation among adolescents: Components and factors. *Heliyon, 6*(9), Article e04905. https://doi.org/10.1016/J.HELIYON.2020.E04905

Brace, I. (2018). *Questionnaire design: How to plan, structure and write survey material for effective market research* (4th ed.). Kogan Page Publishers. http://educationinpakistan.org/wp-content/uploads/2015/09/Questionnaire-Design-How-to-Plan-Structure-Write1.pdf

Chen, Y., Chiu, S. W. K., Zhu, J., & So, W. W. M. (2022). Maintaining secondary school students' STEM career aspirations: The role of perceived parental expectations, self-efficacy, and cultural capital. *International Journal of Science Education*, 1-29. https://doi.org/10.1080/09500693.2022.2032463

Chen, T. P., Lee, K. Y., Kabre, P. M., & Hsieh, C. M. (2020). Impacts of educational agritourism on students' future career intentions: Evidence from agricultural exchange programs. *Sustainability, 12*(22), Article 9507. http://dx.doi.org/10.3390/su12229507

Cohen, J (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). https://doi.org/10.4324/9780203771587

Curriculum Development Division. (2016). *Panduan perlaksanaan sains, teknologi, kejuruteraan dan matematik (STEM) dalam pengajaran dan pembelajaran (Implementation guide for teaching and learning in STEM).* Ministry of Education Malaysia. http://bpk.moe.gov.my/index.php/terbitan-bpk/buku-panduan-pelaksanaan-kssr-kssm?start=30

Dockery, A. M., Phillimore, J., & Bawa, S. (2021). Changing demand for STEM skills in Australia and gender implications. *Australian Journal of Labour Economics, 24*(1), 71-110. https://businesslaw.curtin.edu.au/wp-content/uploads/sites/5/2021/05/AJLE-Vol-24-No-1-Paper-4.pdf

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39-50. https://doi.org/10.1177/002224378101800104

Gómez-Galán, J. (2020). Media education in the ICT era: Theoretical structure for innovative teaching styles. *Information, 11*(5), 276. https://doi.org/10.3390/info11050276

Hair, J. F., Hult, G. M., Ringle, C. M., & Sarstedt, M. (2017). *A primer on partial least squares structural equation modeling (PLS-SEM) (2nd ed.).* Sage Publications.

Hasin, I., & Nasir, M. K. M. (2021). The effectiveness of the use of information and communication technology (ICT) in rural secondary schools in Malaysia. *Journal of Education and e-Learning Research, 8*(1), 59-64. https://doi.org/10.20448/JOURNAL.509.2021.81.59.64

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115-135. https://doi.org/10.1007/s11747-014-0403-8

Hoag, A., Grant, A. E., & Carpenter, S. (2017). Impact of media on major choice: Survey of communication undergraduates. *NACADA Journal, 37*(1), 5-14. http://dx.doi.org/10.12930/NACADA-15-040

James, P. B., Batema, M. N. P., Bah, A. J., Brewah, T. S., Kella, A. T., Lahai, M., & Jamshed, S. Q. (2018). Was pharmacy their preferred choice? Assessing pharmacy students' motivation to study pharmacy, attitudes and future career intentions in Sierra Leone. *Health Professions Education, 4*(2), 139-148. http://dx.doi.org/10.1016/j.hpe.2017.06.001

Kettunen, J., Lindberg, M., Nygaard, E., & Kárdal, J. (2020). Enhancing career practitioners' understanding and use of ICT in guidance and counselling. In E. H. Haug, T. Hooley, J. Kettunen, & R.Thomsen (Eds.). *Career and career guidance in the Nordic countries* (pp. 163-175). Brill Sense. https://doi.org/10.1163/9789004428096_011

Kong, Y., Nima Ngapey, J. D., & Qalati, S. A. (2020). Effects of financial rewards, parents and peers, and benefits and costs on choosing accounting career: A global perspective. *The Journal of Asian Finance, Economics, and Business, 7*(11), 157-167. https://doi.org/10.13106/jafeb.2020.vol7.no11.157

Kricorian, K., Seu, M., Lopez, D., Ureta, E., & Equils, O. (2020). Factors influencing participation of underrepresented students in STEM fields: Matched mentors and mindsets. *International Journal of STEM Education, 7*, 1-9. https://doi.org/10.1186/s40594-020-00219-2

Levine, K. J., & Aley, M. (2020). Introducing the sixth source of vocational anticipatory socialization: Using the internet to search for career information. *Journal of Career Developmen*t, Article 0894845320940798. https://doi.org/10.1177/0894845320940798

Li, J., Mau, W. C. J., Chen, S. J., Lin, T. C., & Lin, T. Y. (2021). A qualitative exploration of STEM career development of high school students in Taiwan. *Journal of Career Development, 48*(2), 120-134. https://doi.org/10.1177/0894845319830525

Malhotra, N. K. (2010). *Marketing research-An applied orientation* (6th ed.). Pearson Education, Inc., publishing as Prentice Hall.

Mohtar, L. E., Halim, L., Rahman, N. A., Maat, S. M., Iksan, Z. H., & Osman, K. (2019). A model of interest in STEM careers among secondary school students. *Journal of Baltic Science Education, 18*(3), 404-416. http://dx.doi.org/10.33225/jbse/19.18.404

Mokhtar, R., Zulkifli, M. A., & Zainuddin, Y. (2016). Psychological characteristic and entrepreneurial intention: The mediation effect of TPB. *Proceedings of the 6th National Conference on Education Technical and Vocational in Education and Training (NCiE-TVET) 2016,* 1120-1135. https://scholar.googleusercontent.com/scholar?q=cache:fjlrgHpw5X0J:scholar.google.com/+Psychological+characteristic+and+entrepreneurial+intention:+The+mediation+effect+of+TPB.+&hl=en&as_sdt=0,5

Navy, S. L., Kaya, F., Boone, B., Brewster, C., Calvelage, K., Ferdous, T., Hood, E., Sass, L., & Zimmerman, M. (2021). "Beyond an acronym, STEM is…": Perceptions of STEM. *School Science and Mathematics, 121*(1), 36-45. https://dx.doi.org/10.1111/ssm.12442

Ottestad, G., & Gudmundsdottir, G. B. (2018). Information and communication technology policy in primary and secondary education in *Europe. Second Handbook of Information Technology in Primary and Secondary Education*, 1-21. https://doi.org/10.1007/978-3-319-53803-7_92-1

Qader, I. K. A., & Zainuddin, Y. B. (2011). The impact of media exposure on intention to purchase green electronic products amongst lecturers. *International Journal of Business and Management, 6(*3), 240. http://dx.doi.org/10.5539/ijbm.v6n3p240

Raabe, I. J., Boda, Z., & Stadtfeld, C. (2019). The social pipeline: How friend influence and peer exposure widen the STEM gender gap. *Sociology of Education, 92*(2), 105-123. https://doi.org/10.1177/0038040718824095

Razali, F. (2021). Exploring crucial factors of an interest in STEM career model among secondary school students. *International Journal of Instruction, 14*(2), 385-404. https://files.eric.ed.gov/fulltext/EJ1291108.pdf

Ringle, C. M., Wende, S., & Becker, J. M. (2015). *SmartPLS 3.0*. SmartPLS GmbH http://www.smartpls.com

Sheehan, K. J., Hightower, B., Lauricella, A. R., & Wartella, E. (2018). STEM media in the family context: The effect of STEM career and media use on preschoolers' science and math skills. *European Journal of STEM Education, 3*(3), 17. https://doi.org/10.20897/ejsteme/3877

Shin, S., Rachmatullah, A., Roshayanti, F., Ha, M., & Lee, J. K. (2018). Career motivation of secondary students in STEM: A cross-cultural study between Korea and Indonesia. *International Journal for Educational and Vocational Guidance, 18*(2), 203-231. https://doi.org/10.1007/s10775-017-9355-0

Solikhah, B. (2014). An application of theory of planned behavior towards CPA career in Indonesia. *Procedia-Social and Behavioral Sciences, 164*, 397-402. http://dx.doi.org/10.1016/j.sbspro.2014.11.094

Steinke, J., Applegate, B., Penny, J. R., & Merlino, S. (2022). Effects of diverse STEM role model videos in promoting adolescents' identification. *International Journal of Science and Mathematics Education, 20*(2), 255-276. https://doi.org/10.1007/S10763-021-10168-Z

Warne, R. T., Sonnert, G., & Sadler, P. M. (2019). The relationship between advanced placement mathematics courses and students' STEM career interest. *Educational Researcher, 48*(2), 101-111. https://doi.org/10.3102/0013189X19825811

Wong, K. K. K. (2016). Mediation analysis, categorical moderation analysis, and higher-order constructs modeling in partial least squares structural equation modeling (PLS-SEM): A B2B Example using SmartPLS. *Marketing Bulletin, 26*(1), 1-22.

Wyss, V. L., Heulskamp, D., & Siebert, C. J. (2012). Increasing middle school student interest in STEM careers with videos of scientists. *International Journal of Environmental and Science Education, 7*(4), 501-522, Article ijese.2012.039. http://www.ijese.net/makale/1553.html

Zhao, X., Lynch, J. G., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research, 37*(2), 197–206. http://dx.doi.org/10.1086/651257

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Rural Teachers' Professional Development under the Model of "Internet plus" Volunteer Teaching in Normal College

Jian ZHAO*Wenjing YANG
School of Educational Technology, Northwest Normal University, China
*KjYwj0911@163.com

Abstract: "Internet plus" Volunteer Teaching Service Project in Normal College was carried on by 8 normal colleges and 17 deeply impoverished counties in Gansu Province in China and this project was designed to improve Information Communication Technology(ICT) of normal students and rural teachers through field teaching mixed with remote teaching. The present study analyzed the current situation and difficulties of rural school teachers' professional development, based on an investigation of the adaptability of rural teachers in "Internet plus" Volunteer Teaching Service for Normal College, and proposed a set of long-term mechanisms for rural teachers' professional development to improve the teaching quality of rural weak schools.

Keywords: Mobile Technology, Unit Teaching, Classroom interaction behavior

1. Introduction

The development of rural education is related to the breadth and depth of the implementation of the rural revitalization strategy. The State Council's "Guiding Opinions on Actively Promoting the" Internet Plus "Action" emphasize the requirement to explore the new mechanism of providing education services, expand the coverage of high-quality education resources, and ensure the fairness of the allocation of education resource,(Guiding Opinions on Actively Promoting the "Internet Plus" action,2015). According to The Lisbon Education and Professional Teacher Training Development Strategy, the capability-depend methodology is required to solve the problem of training for new generation professional talent in the education field and the critical point is how to maintain sufficient flexibility during the innovation,(V Valentyna C,2020). Guay F. etc. in the scientific work “Teacher structure as a predictor of students” perceived competence and autonomous motivation: The moderating role of differentiated instruction “notes that through their competencies the teacher to some extent positively affects the formation of students” autonomy(Guay F. ,2017). Studies on influencing factors of rural teachers' professional development include Wang J, Tigelaar D. E. H. etc. all believed that the professional development of rural teachers was limited by rural economic and cultural development and educational cognition level,(Wang J,2021). Xiao Zhengde believes that the development dilemma faced by rural teachers mainly includes five aspects: heavy teaching tasks, great psychological pressure, reduced career satisfaction, high job burnout, and low career happiness,(Xiao,2011). L. s. et al. proposed that the key factors affecting the development of teachers' professional ability include internal factors, namely motivation, activity, emotion, and will, and external factors, including the content, technology, methods, tools, and forms of educational materials,(Guay, F. ,2017).

The researches whose topics are teachers’ development in the rural area, but most of them mainly focus on the teachers themselves and related external environment. However, there are few kinds of literature about how rural teachers can achieve long-term development under the "Internet plus" model of supporting teaching in normal universities(Generally refers to colleges and universities that train various types of teachers). In recent years, China's education reform has focused on improving teachers’ ICT in rural areas. "Internet plus"Volunteer Teaching Service Project in Normal College make it achieved that schools in the rural areas could share higher quality education resources of urban areas. For solving the rural school unable to open national courses completely,(Ministry of Education

issues guidelines on strengthening the application of "three classrooms",2015)" as the main goal, breaking the time-space limitations using the Internet, the project was designed to promote rural small-scale school education teaching quality as the common desire. Achieving multiple breakthroughs and comprehensive innovation has the significance of leading exploration in China. The Study is focused on the following questions: What are the difficulties rural teachers' professional development in the "Internet plus" Volunteer Teaching Service Project in Normal College? How can teachers in rural areas achieve long-term professional development under the model of "Internet plus" Volunteer Teaching in Normal Colleges?

2. Method

This study focuses on the rural teachers' professional development in the "Internet Plus" Volunteer Teaching Service Project in Normal College, which was first implemented in Gansu, China. Based on the literature research method and questionnaire method, this study analyzes the factors influencing the professional ability development of rural school teachers from two aspects: the current situation of professional development of rural teachers in"Internet plus"volunteer teaching and the current situation of dual-teacher classroom teaching structure, and puts forward the professional development path of rural teachers according to the obtained data. The research object is 379 rural school teachers who participated in the project, covering 175 primary and secondary schools and teaching points in 154 administrative villages of 62 towns and townships in 9 cities and 17 counties and districts.

3. Results and Analysis of Data

In this part, we analyzed the collected data, which reflected the difficulties and appeals of 379 rural teachers who participated in the survey during the project implementation. The problems are mainly found in the low professional ability of information technology, the unclear orientation of teacher training, lack of professional knowledge of teachers, and poor adaptability of the teaching model of the "double-teacher classroom".

Before the implementation of the project, most rural teachers believed that limited teaching methods, single teaching evaluation, backward education concepts, and lack of teaching resources hindered the improvement of the learning quality of school students. After the implementation of the project, we found that rural teachers had a great improvement in these aspects. (See Figure 1)

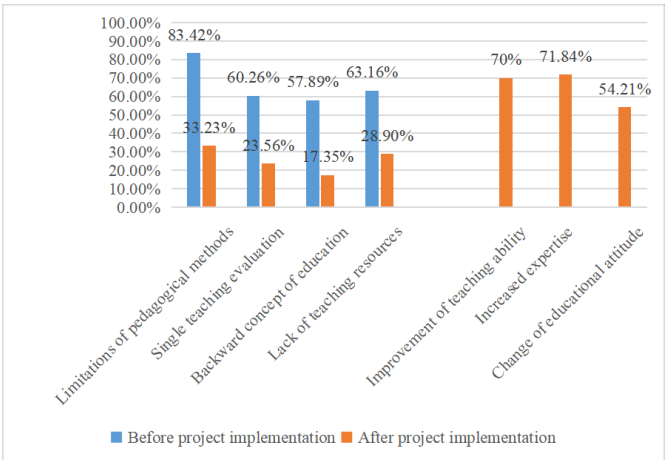


Figure 1. Comparison before and after project implementation.

The survey found that 68.34% of the teachers in the dual-teacher classroom thought that the cooperation between the main teacher and assistant teacher was low, resulting in a poor teaching effect(See Table 1). The results show that 75.73% of rural teachers think their schools are short of teacher resources, while 53.56% think their schools have an unreasonable teacher-team structure(See Table 2).

Table 1. *Problems to be Improved in Dual-teacher Classroom*

	Frequency	Percent	Mean	Standard Deviation
Increase the communication and cooperation between the lecturers and auxiliary teachers	259	68.3%	.68	.466
Strengthen the classroom management of auxiliary teachers	234	61.7%	.62	.487
Increase the interaction between auxiliary teachers and students	200	52.8%	.53	.500
Strengthen the supporting role of network technology	148	39.1%	.39	.489

Table2. *Factors Influencing Curriculum Ofering in Small-scale Rural Schools*

	Frequency	Percent	Mean	Standard Deviation
Shortage of Teachers	287	75.7%	.76	.429
Teachers' Lack of Subject Professional Knowledge	199	52.5%	.53	.500
Teachers Lack Professional Quality	94	24.8%	.25	.432
Structure of the Teaching Staff is not Reasonable	203	53.6%	.54	.499
Subject Teaching Tools are Inadequate	197	52.0%	.52	.500
Other	8	2.1%	.02	.144

4. Conclusions

Based on the "Internet plus" teaching support service project in normal colleges and universities implemented for the first time in China, this research analyzes the surveyed data, tries to use information technology to break down the barriers of time and space, and improve the professional development of rural teachers, to ensure the quality of education and promote the Equity in education, and promote the development of education. The results show that from the perspective of teachers' development, it is necessary to clarify the self-awareness of teachers' professional development, improve the technical practice ability of teachers' professional development, enhance the adaptability of teachers' dual-teacher classroom teaching mode, reduce teachers' remote dependence, and enhance classroom participation. Secondly, from the perspective of external environmental impact, it is necessary for the relevant education departments of the country to jointly ensure the accuracy of teacher training and solve the problem of the structural shortage of rural teachers. Finally, although we have done some research and surveys, summarized and analyzed the current situation and difficulties of rural school teachers, and proposed a long-term development mechanism, the research object is from Gansu, China, which is not enough to represent all rural school teachers. The practice process must conform to the actual situation. The "Internet plus" teaching in normal colleges and universities is a new teaching model, and more exploration and research are needed in the future.

References

Ministry of Education issues guidelines on strengthening the application of "three classrooms"[EB/OL]. http://www. gov.cn/zhengce/zhengceku/2020-03/16/content_5491791. htm,2020-03-03.

The State Council. Guiding Opinions on Actively Promoting the "Internet Plus" action. [EB/OL]. http://www. gov. cn/zhengce/content/2015-07/04/content_10002. htm,2015-07-04.

V Valentyna C,Olha Z,et al.Development of professional competence of present and future teachers under the conditions of transformational processes in education[J]. Journal for Educators, Teachers, and Trainers,2020,11(1):56-67.

Guay,F.,Roy,A.,Valois,P.(2017).Teacher structure as a predictor of students ' perceived competence and autonomous motivation:The moderating role of differentiated instruction. British Journal of Educational Psychology,87(2),224– 240.URL:https://doi.org/10.1111/bjep.12146(date of issue 02. 06. 2020).

Wang,J.,Tigelaar D E H, et al.Rural teachers ' sharing of digital educational resources: From motivation to behavior [J]. Computers & Education,2021,161:1-17.

Xiao,Z.D.,Living Situation of Rural teachers in the Process of Urbanization and Improvement strategies[J]. China Education Journal,2011(08):1-4.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Exploring Predictive Indicators of Reading-Based Online Group Work for Group Formation Teaching Assistance

Changhao LIANG[a*], Izumi HORIKOSHI[a], Rwitajit MAJUMDAR[a], Brendan FLANAGAN[a] & Hiroaki OGATA[a]
[a] *Kyoto University, Japan*
*liang.changhao.84c@st.kyoto-u.ac.jp

Abstract: Using digital systems to group students according to their indicators provides opportunities for better group work implementation. However, how these indicators can affect group work performance remains unclear. Teachers tend to feel confused about which indicators should be considered when creating groups using learning log data. Capitalized on the data-driven environment under GLOBE, we conducted a preliminary study to explore predictive indicators for algorithmic group formation in a reading-based group learning context. This study presented our effort to explore the key factors that correlated to a desirable group work via factor analysis and correlation analysis. We found that reading engagement and previous peer rating scores suggest a higher potential to predict desirable group work performance in the reading-based online group work, which aims to help teachers set appropriately in future student model data-based group formation.

Keywords: data-driven group formation, reading-based group work, CSCL, factor analysis, correlation analysis

1. Introduction

Group learning is progressively adapted in pedagogical practice and prevalent online courses nowadays raise impetus to the demand for such interactive activities. To initial a group work, teachers should align students appropriately according to different learning contexts (Urhahne et al., 2010). With the scaffold of Computer-supported collaborative learning (CSCL) (Stahl et al., 2006) and learning analytics (LA) (Siemens, 2012), tools and enriched data became available to support group work implementation such as algorithmic group formation systems (Liang et al., 2019).

Meanwhile, teachers face difficulties when using digital systems (Austin et al., 2010). As the group formation system provides multiple input variables, teachers may feel confused about what each variable conveys and get overwhelmed by the settings. To enable teachers to smoothly form desirable groups, we conducted this study based on a data-driven environment. Using the log data from reading-based online group work, we made a preliminary analysis to estimate the potential of existing indicators to predict group work performance. This study aims to assist teachers by achieving the following goals:

- Present more explainable group formation input variables for teachers.
- Enrich experience by parameter-setting recommendations for similar group learning contexts.

2. Research background

There exist multiple issues that deserve nuanced consideration when executing collaborative learning activities (Urhahne et al., 2010). Janssen & Kirschner (2020) categorized these indicators into three attributes: antecedent, process and consequence. Since indicators of antecedents can pose an effect on processes and consequences of collaboration (see Figure 1), to deal with antecedent attributes, studies on group formation techniques using these attributes are highlighted in the CSCL field to help teachers set groups appropriately before the group work (Maqtary et al., 2019).

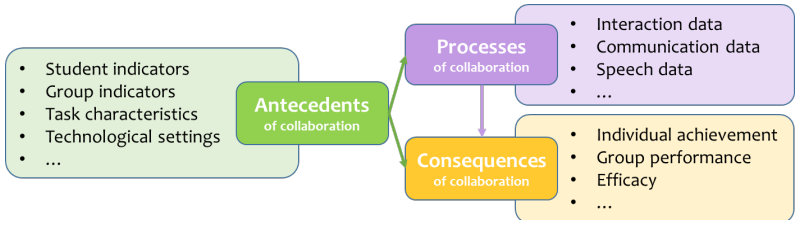


Figure 1. Indicators among Collaborative Process Attributes (Janssen & Kirschner, 2020).

As for antecedent, Janssen & Kirschner (2020) listed several indicators based on what it describes. Student and group indicators are frequently-used (Saqr et al., 2020) but vary from group work tasks. Student indicators can include intellectual abilities, experiences, preferences and personalities (Sánchez et al., 2021). Group indicators depict higher-level aspects such as social relationships and heterogeneity of certain student-level indicators within a group, hence Cress (2008) underscored the necessity to simultaneously take into account student indicators as well. In terms of process attributes, behavior logs such as students' interaction logs (Saqr et al., 2020) as well as utterance (Liang et al., 2021) are highlighted to reflect an overall look of the group work engagement. In terms of group work implementation, process data such as students' interaction logs are highlighted to reflect an overall look of the group work engagement (Saqr et al., 2020). Besides, when it comes to consequences, individual achievement, perceived efficacy and group work performance are used to evaluate the consequence of group work (Janssen & Kirschner, 2020).

Since antecedents tend to affect processes and consequences, researchers have presented the impact of some antecedent indicators on the process and consequence of group work in different scenarios. For example, the previous knowledge and task experience proved to be closely related to the group work performance in the collaborative searching tasks (Jiang et al., 2022). The collaborative tendency can also affect group work outcomes (Zhang et al., 2022). Besides student indicators, groups with different sizes, intimacy levels (Block, 2015) and compositions (Sánchez et al., 2021) make a difference in the consequence. Though former researchers inspected the effect of specific indicators under controlled experiments, less study comprehensively looks at these modeling antecedents, and which input plays a more important role remains unclear.

3. Group work indicators in a data-driven environment

Group Learning Orchestration Based on Evidence (GLOBE) provides a data-driven environment (Liang et al., 2021) to study the indicators of group work mentioned in the previous section. As is illustrated in Figure 2, the data-driven support via GLOBE systems covers four phases: group formation, orchestration, evaluation, and reflection.

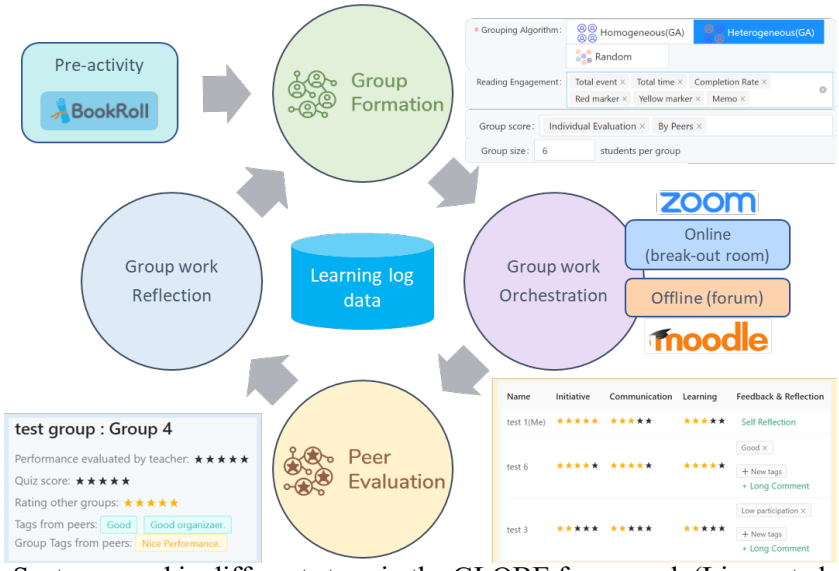


Figure 2. Systems used in different steps in the GLOBE framework (Liang et al., 2021).

3.1 *Conducting group work using GLOBE*

To conduct group work under GLOBE, teachers can select any number of indicators to create either homogeneous or heterogeneous groups via an algorithmic group formation system (Flanagan et al., 2021). These student indicators come from learning log data from the reading attributes computed from BookRoll logs (Flanagan & Ogata, 2018), and external sources such as test scores uploaded to the LAView dashboard (Majumdar et al. 2019). During the group work, the teacher can overview the group work progress and get prompts of intervention for undesirable groups with the help of orchestration data analyzers and dashboards. After the group work, peer and teacher evaluations with ratings and feedback are provided in the evaluation modules.

Throughout the group formation, orchestration and evaluation phases, GLOBE covers multiple indicators presenting antecedent, process and consequences attributes. Such group work-related data are used and generated within different phases, hence ensuring a systematic and continuous usage of data to buttress subsequent group work with the same students.

3.2 *Modeling for key input indicators for prediction*

Table 1 shows the student indicators available in GLOBE systems with their corresponding Collaborative Process Attribute (Janssen & Kirschner, 2020). These indicators convey certain student indicators or skills mentioned in recent studies. Given the set of indicators, while using the system some of the previous indicators can be used as antecedent indicators in subsequent group formation, which shows an advantage of the system that maintains the history of learners' group work activities. However, feature selection remains a problem for teachers, especially those who are not familiar with CSCL tools. We start from the analysis of student indicators that describe individual-level characteristics. A two-step analysis is presented to answer the following questions:

- Q1: What are the principal factors that can prune the antecedent indicators?
- Q2: What are relationships among indicators in different Collaborative Process Attributes?

Table 1. *Student indicators available in GLOBE systems*

Collaborative Process Attribute	Indicators within GLOBE framework	Data source	Proxy for construct (what does it convey)
Antecedents of collaboration	reading time	BookRoll	Academic performance (Juncoet al., 2015)
	operation times	BookRoll	
	completion rate	BookRoll	
	red marker (highlight)	BookRoll	Active reading skills (Toyokawa et al., 2021)
	yellow marker (difficult)	BookRoll	
	*memo	BookRoll	
	*quiz score	LAView	Prior knowledge
Processes of collaboration	forum view	Moodle forum	Engagement and active interactions (Fidalgo-Blanco et al., 2015)
	forum post	Moodle forum	
Consequences of collaboration	*teacher's rating	Peer evaluation	Group work experience and task experience (Janssen & Kirschner, 2020)
	peer rating (individual)	Peer evaluation	
	peer rating (group)	Peer evaluation	
	final course grade	LAView	Academic performance

* not used in this study

4. Dataset and analysis results

The data of a university course in which 30 students were enrolled is selected for this study. Reading-based online group work was conducted several times across the 15-week semester, where students

were required to read and summarize research articles as after-class group work and present their reading outcomes during the online lecture. We used the data from the last group work in week 11 for analysis, since it contains abundant data and students had got familiar with the process. Five groups with six members were formed by a heterogeneous algorithm with checked antecedent indicators in Table 1. Two students were excluded from the dataset due to their absence.

4.1 *Factor analysis for input parameters*

To explore the underlying factors from the antecedent indicators, principal component factor analysis with varimax rotation was implemented. The Kaiser–Meyer–Olkin Measure of Sampling Adequacy (MSA) value for the correlation matrix of the dataset is 0.64, showing medium appropriation for factor analysis (Ferguson & Cox, 1993). The F value of Bartlett's Test is significant as well.

As a result, three factors were identified with initial eigenvalues greater than 1, which can account for 81.85 percent of the indicators. Table 2 shows the percentage variance explained by the antecedent indicators and Table 3 presents the rotated component matrix of factors. The communalities of the input variable range from 0.58 to 0.95, showing that all of the indicators can be expressed by these factors. A cutoff for statistical significance of the factor loadings of 0.5 was used (Muilenburg & Berge, 2005). Each item loaded distinctively on one factor; the highest factor loading was separated from its next nearest loading by at least 0.4.

Table 2. *Total variance explained by factors*

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
Factor 1: engagement	3.364	48.056	48.056
Factor 2: active reading tendency	1.339	19.134	67.191
Factor 3: previous peer ratings for group	1.027	14.667	81.858

Table 3. *Rotated component matrix of factors*

	Component		
	Factor 1: engagement	Factor 2: active reading tendency	Factor 3: previous peer ratings for group
completion rate	0.908		
operation times	0.907		
reading time	0.807		
previous peer rating (individual)	0.693		
red marker		0.909	
yellow marker		0.888	
previous peer rating (group)			0.973

We named factor 1 as “engagement” since it covers the reading engagement indicators. The previous peer ratings scores are loaded into this factor as well. We can infer that the peer rating of online group work is partially based on the engagement which is the most obvious to the groupmates, hence members with high reading engagement tend to show high group work engagement as well. Factor 2 was named as “active reading tendency” according to Toyokawa et al. (2021), and factor 3 as “previous group work output”. These factors provide a more understandable integration of raw input variables from the group formation system, hence also throw light on the optimization of the system.

4.2 *Correlational analysis for predictive antecedents*

Besides dimension reduction, we explored predictive antecedents using correlational analysis to find predictive indicators for novice teachers to choose from in a similar context. Pearson correlational examination was calculated among indicators of three phases and Figure 4 presents the result.

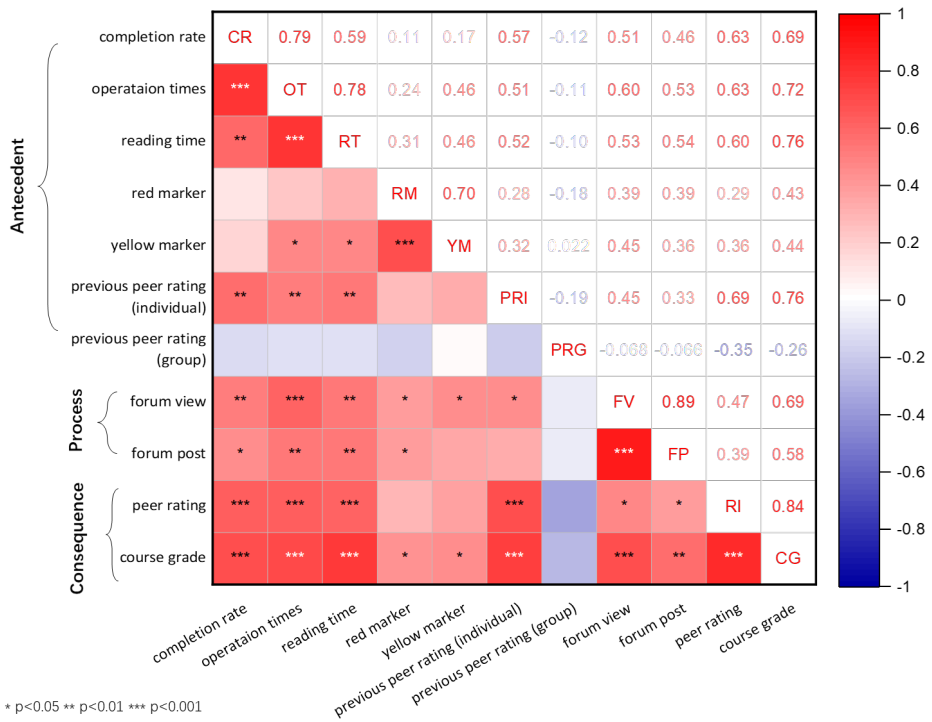


Figure 4. Results of correlation analysis of all indicators in group work

Reading engagement indicators (completion rate, operation times and reading time) suggest significant positive pertinence to all process and consequence attributes. Students with high engagement in reading tasks are inclined to not only get better academic achievements (Juncoet al., 2015) but also take an active part in group discussion and get positive peer feedback. Meanwhile, previous peer rating scores for group members show similar relevance to the reading engagement indicators with high predictive potential, which proves the impact of cyclically accumulated group work data under GLOBE (Liang et al., 2021). The results also agree with the idea that prior knowledge and task experience can influence the group work consequence (Janssen & Kirschner, 2020). Besides, annotation behaviors are positively related to the forum views as well as final course grades, and red markers suggest positive relation to the forum post behaviors which indicates active participation (Fidalgo-Blanco et al., 2015). Such finding indicates a promising connection between acting reading behaviors (Toyokawa et al., 2021) and its corresponding forum-based group work participation. This result is also meaningful to inform and assist the ways teachers create groupings for future reading-based group work.

For peer ratings of previous group work output, there appears no significant relation to any other indicators, indicating a low predictive weight. As this indicator derives from a group-level rating, social loafing and free-riding could decrease its reliability, it should be less considered in following group creations for teachers.

5. Conclusions and Future Work

This study puts forward a foreground to modeling from log-based antecedent indicators to predict group work processes and consequences. Though we focused on reading-based group work context, we disclosed the opportunity for predictive analysis in different tasks under the GLOBE environment.

The indicators listed herein are still limited in the current stage and cannot fully depict student characteristics. We only examined student indicators in a specific group work context, while group-level indicators such as intimacy and heterogeneity were not covered. Since the dataset only covers one group work of the course, the results need validation within a larger sample in subsequent works.

In the next step, more process indicators such as forum interaction logs will be re-used as antecedents, and group-level indicators like heterogeneity will be examined. Capitalized on the integrated data-rich environment, predictive modeling such as regression would be done with a larger sample so that the specific weight of each antecedent indicator can be estimated. This preliminary

analysis attempted to make parameter selection more understandable to teachers and presented parameter-setting recommendations for similar learning contexts, and will lead to system improvement as future work to lower the threshold for teachers.

Acknowledgements

This research was supported by the following grants: JSPS KAKENHI 20K20131, 20H01722, 22H03902, NEDO JPNP20006, JPNP18013, and JPMJSP2110.

References

Austin, R., Smyth, J., Rickard, A., Quirk-Bolt, N., & Metcalfe, N. (2010). Collaborative digital learning in schools: Teacher perceptions of purpose and effectiveness. *Technology, Pedagogy and Education*, *19*(3), 327-343.

Block, P. (2015). Reciprocity, transitivity, and the mysterious three-cycle. *Social Networks*, *40*, 163-173.

Cress, U.(2008).The need for considering multilevel analysis in CSCL research: An appeal for the use of more advanced statistical methods.*International Journal of Computer-Supported Collaborative Learning*,*3*,69-84.

Ferguson, E., & Cox, T. (1993). Exploratory factor analysis: A users' guide. *International journal of selection and assessment*, *1*(2), 84-94.

Fidalgo-Blanco, Á., Sein-Echaluce, M. L., García-Peñalvo, F. J., & Conde, M. Á. (2015). Using Learning Analytics to improve teamwork assessment. *Computers in Human Behavior*, *47*, 149-156.

Flanagan, B., & Ogata, H. (2018). Learning analytics platform in higher education in Japan. Knowledge Management & E-Learning: An International Journal, 10(4), 469-484.

Flanagan, B., Liang, C., Majumdar, R., and Ogata, H. (2021). Towards explainable group formation by knowledge map-based genetic algorithm. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp.370–372).

Janssen, J., & Kirschner, P. A. (2020). Applying collaborative cognitive load theory to computer-supported collaborative learning: towards a research agenda. *Educational Technology Research and Development*, *68*(2), 783-805.

Jiang, X., Li, Y., Zhang, Y., Z. & Liu, C. (2022). Research on the Effect of Prior Knowledge Differences on Collaborative Information Search and Learning. *Library and Information Service*, *66*(8), 32-41.

Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, *27*, 54-63.

Liang, C., Boticki, I., & Ogata, H. (2019). Supporting teachers in group work formation and analytics for in-class group activities. In *the 27th International Conference on Computers in Education (ICCE 2019)*(pp.744-749).

Liang, C., Majumdar, R. & Ogata, H. (2021). Learning log-based automatic group formation: system design and classroom implementation study. *Research and Practice in Technology Enhanced Learning*, *16*(1), 1–22.

Majumdar, R., Akçapınar, A., Akçapınar, G., Flanagan, B., & Ogata, H. (2019). Learning analytics dashboard widgets to author teaching-learning cases for evidence-based education. In *Companion Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (pp.600-607).

Maqtary, N., Mohsen, A., & Bechkoum, K. (2019). Group formation techniques in computer-supported collaborative learning: A systematic literature review. *Technology, Knowledge and Learning*, *24*(2),169-190.

Muilenburg, L. Y., & Berge, Z. L. (2005). Student barriers to online learning: A factor analytic study. *Distance education*, *26*(1), 29-48.

Sánchez, O. R., Ordóñez, C. A. C., Duque, M. Á. R., & Pinto, I. I. B. S. (2021). Homogeneous group formation in collaborative learning scenarios: An approach based on personality traits and genetic algorithms. *IEEE Transactions on Learning Technologies*, *14*(4), 486-499.

Saqr, M., Nouri, J., Vartiainen, H., & Malmberg, J. (2020). What makes an online problem-based group successful? A learning analytics study using social network analysis. *BMC medical education*, *20*(1), 1-11.

Siemens, G. (2012). Learning analytics: envisioning a research discipline and a domain of practice. In *Proceedings of the 2nd international conference on learning analytics and knowledge* (pp. 4–8).

Stahl, G., Koschmann, T., & Suthers, D. (2006). Computer-supported collaborative learning: An historical perspective. In R. K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 409–426).

Urhahne, D., Schanze, S., Bell, T., Mansfield, A., & Holmes, J. (2010). Role of the teacher in computer-supported collaborative inquiry learning. *International Journal of Science Education*, *32*(2), 221-243.

Toyokawa, Y., Majumdar, R., Lecailliez, L., & Ogata, H. (2021). A flipped model of active reading using a learning analytics-enhanced e-book platform. In *the 29th International Conference on Computers in Education (ICCE 2021)* (pp. 510-515).

Zhang, P., Wang, D. & Tang, Z.(2022).Research on the Online Knowledge Collaboration Behaviors: the Influence of Team Member Intimacy and Collaborative Tendencies. *Library and Information Service*, *66*(8), 21-31.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

The Budding Botanist Paradox: Automating Human Inquiry with Immersive Technology

Eric HAWKINSON
Faculty of Global Engagement, Kyoto University of Foreign Studies, Japan
Faculty of Information and Communication Studies, University of The Philippines Open University
https://erichawkinson.com

Abstract: The way humans are using technology to augment our learning processes is evolving rapidly on a fundamental level. This evolution can be seen in the ways mobile computing has given us the ability to seek just-in-time information and have that information be customized and contextualized to fit our needs. Immersive technology is poised to exacerbate some of these trends and further supplement digital information to everything, everywhere. This study attempts to frame this progression in the automation of some of the most fundamental processes of human inquiry from mobile learning and the transition into immersive learning using tools like augmented and virtual reality. A hypothetical scenario of a learner of botany is presented and analyzed to explore continuing trends and concepts of these technologies and their use in learning contexts as they evolve. The budding botanist paradox is a thought experiment that attempts to reason that the more automated the learning process becomes, enhanced, and augmented with immersive technology, the more learners could be dependent on these automated systems for basic learning, more susceptible to second order influences in behavior, and the more vulnerable teachers and learners are if/when the automated systems have problems.

Keywords: Immersive Learning, Augmented Reality in Education, Metaverse Learning, Automation in Education

1. Setting the Stage

A great deal of opportunity for learning lies in the integration of technology into educational contexts. Teachers and students can both benefit from technology by being more efficient and effective, or perhaps in the case of immersive technologies, new viewpoints, and realities from which to understand our world (Eric Hawkinson, Mehran, and Alizadeh 2017). It is possible that in the coming years as these technologies come into ubiquity, we may be forced to rethink some decisions of how we introduce and maintain automated systems in education and other related fields, especially when immersive technology is applied. This trend is already making its way into educational contexts with the introduction of online and mobile learning. Now immersive technologies are being implemented in various ways, bringing with it a new level of possibilities for learning, but also unprecedented risks in privacy, data collection, and other second order effects in the automated processes that are built into our technologies (E. Hawkinson and Klaphake 2020).

1.1 The Paradox of Automation

Human beings will become less accomplished at less challenging or novel work as automation eliminates opportunity to do it. This is the concept of the paradox of automation, the more automated systems are introduced, the harder it is to understand and fix when things go wrong and the fewer the people who understand how to troubleshoot and fix the problem (Bainbridge 1983; Strauch 2018). A prime example is the idea of "Loss of control" due to over automation is the classification for Air France Flight 447 (AF447). Air France Flight 447 crashed with all

lives lost over the Atlantic Ocean as it flew from Rio de Janeiro, Brazil, to Paris, France. On June 1st, 2009. After an investigation, it was found that the pilots stalled the plane when the autopilot malfunctioned, known as a 'loss of control' event to the Federal Aviation Association (FAA). Commercial aviation's greatest cause of casualties is a loss of control. From 2006 to 2015, nearly 1,400 people died in 15 fatal air accidents; between 2010 and 2014, there were 37 fatal air accidents and 1,242 deaths (Palmer 2013). There has been a decline in most types of aviation accidents, but incidents related to loss of control have stuck around. It is interesting to note that reducing accidents has also been credited to some of the same measures that have been attributed to loss of control - like sophisticated cockpit automation. In learning, automation is likely also to be a net positive, but AF447 shows us that we should at least be aware of how automation can take us to unintended places, and how we might design learning to avoid these pitfalls. It could be argued that AF447, and losses of control in general, represent extreme examples of failure of systems that function well within limits normally (Salmon, Walker, and Stanton 2016; Oliver, Calvard, and Potočnik 2017). It is my hope that more teachers and educational technologists grasp this paradox and its consequences for teaching and learning. The events leading to the loss of AF447 can be explained in part in lack of practice of manual control.

1.2 Relating Automation to Educational Contexts

There is a lot of work to be done in better understanding the role of automation in educational contexts such as plagiarism checkers, online testing, and massive open online courses, among others. It is a standard policy that student submissions should be carefully reviewed and graded by teachers to identify problems and provide feedback (Pappano 2012; Adedoyin and Soykan 2020). It is a huge amount of work to grade students' submissions once they are received that is being automated more and more. Examples in online education could be students who must give peer instruction or work together to answer an automated quiz instead or in addition to a weekly lecture. Grading can automated in such a way that students receive feedback in a timely manner through canned responses. This process is observed to be very similar to the aviation industry and autopilot leading up to AF447. There is an organizational need to scale and become more efficient, and systems are built to accommodate those needs, and the cycle continues. This cycle is poised to accelerate with the introduction of immersive technologies, as the amount of data to analyze and the ability to quickly make changes is increased simultaneously. For example, a virtual learning environment can take user biometric data taken from a VR headset and quickly make changes to all manner of aspects from detecting a learning disability to replacing content with user preferences (Cowan, Javornik, and Jiang 2021; Bekele et al. 2013).

2. Thought Experiment: The Budding Botanist Paradox

Frequently, universities automate important processes, such as grading programs, in order to deal with resource constraints (Hayes and Introna 2005; Savage 2004). As we observe possible training/learning failures in pilots when autopilot systems are introduced, there are perhaps lessons to be learned and applied to learning design and technology enhanced interventions. Perhaps forcing educators and policy makers to think beyond the obvious questions when looking to design and implement automated systems for teaching and learning. Cost and convenience are sometimes measured against risk, or whether the proposed automation supports or deters student learning. Automating instruction is a way to improve student achievement while reducing the demand on scarce resources for classroom instruction. But the paradox of automation, paired with the iterative cycle and mass scaling may bring into question some deeper questions and design decisions. To provide a better understanding of this concept

in educational contexts, especially when paired with immersive technologies, I have created a thought experiment, a scenario from which to take a specific example of learning with automated systems and immersive technology to explore.

2.1 Learning Botany in Nature

Let us imagine someone learning botany and explore how automated systems contrast the learning of botany with and without automated and immersive technologies. As a budding botanist studying out in nature, there is a process, a method of inquiry that one must become skilled at to accurately identify the species of plants correctly. This process, perhaps most importantly, requires the learner to ask a series of questions about the environment, the time, the place, the season, and a host of other questions that would help inform perhaps other questions that lead to a category of plants and the process continues until one can name the individual species of plant with any sort of confidence.

By contrast, using online services or automated tools to identify plants or parts of plants that are becoming more common (White, Feiner, and Kopylec 2006). Our devices are increasingly collecting and curating information about where we are, what we do, what we say, who we are with, and much more. With this a pattern of behavior is formed and predictions can be made about what questions might need answering next. Smartphones have GPS and other radios that track our movements, and cameras, microphones and more that take in data to analyze a variety of things. This data can now be the foundation of automating the process of inquiry that traditionally happened for the budding botanist out in nature. Knowing information like our location, recent weather, climate trends, and the mountains of pictures, video, and other data generated by other users, the answers to the questions one would ask to come down on a plant species can be automated, and perhaps be prompted to us before we ask them. These systems have already been tried in botanical training in higher education to some success (Chien et al. 2019).

With immersive technologies, in this case augmented reality powered by computer vision and image recognition technologies, a picture, video, or 3D scan of the plant can be taken, uploaded, and compared with mountains of other similar pictures, videos, or scans. Paired with the other data, such as location, search history, weather, and other bits of information, an answer to the question, “What species of flower is this” is answered in an instant, along with the process and questions leading up to it. This could be analogous to modern-day equivalent to looking in the back of the book for the answers.

It is not clear if this is also a net plus for learners. It likely depends on the context of learning and the context of automating in practice, much like pilots and autopilot. In one case of young learners, a greater comprehension of plants studied was found by employing an AR app in the learning process (Chien et al. 2019). If one intends not to become a botanist, perhaps studying climate change having a tool to track plants to correlate to climate patterns might be a great learning boost to tangential learners to botany. But just as the paradox of automation suggests, we are outdating the underlying skills of inquiry. Thus, people are less able to cope with or replace failed or misleading automation systems. Learners of botany are now leaning on the automated system. If there are errors in this system, they may scale easily to many and harder to correct.

2.2 Contrast Learning about Plants with AR and without AR

It can be very helpful to learn about plants using AR. It appears that a general written query to Google will only provide us with a small amount of accuracy and a large amount of other relevant information. However, the situation can be reversed by using AR. In the search results, it can find exactly what you just imagined, so you will get a lot of accurate information and

very little irrelevant information. This is due to the amount of data used in the curation of the response to a query. Augmented reality employs data from tools much later in the chain of telemetry, and therefore more robust in its ability to analyze a question, and by extension, more empathic.

3. 3 Stages in Automated Learning Systems

To better illustrate this trend in the automation of human inquiry towards immersive technology, let's examine three steps/stages in the most fundamental forms of the simple task of answering a question with the aid of technology. These three stages come in the form of three technology developments from the internet connected personal computer, to smartphones, and then to augmented reality enhanced smart glasses. Each of these technologies have popular methods for seeking information to answer simple questions. It is my hope that an examination of these stages will both illustrate the trend around automated learning systems and build context to the budding botanist paradox.

3.1 Personal Computers and Google Search

Internet search was a boon soon after the world wide web layer was developed and popularized in the 1990s. Google determines the order of its search results by using a ranking system called PageRank. Especially in its earliest iterations, Google search didn't use much more than a string of text for input data to give a very large set of data as results. This many times forced users to either reword or refine their search or to learn the syntax of Google to specify, exclude, or exclude specific criteria. This also gave the chance for the accidental discovery of pertinent tangential information along the way.
This could be considered analogous to the budding botanist in nature without the aid of technology. Giving incentive to the learner to ask different questions to find the answer needed, and in the process offering the chance of discovery, but also being presented with information to curtate and integrate on your own (Fisher, Smiley, and Grillo 2021).

3.2 Smartphones and Virtual Assistants

Users can now search for spoken words instead of typed ones by using a voice assistant like Siri (iOS) and Bixby (Samsung). The voice assistant also takes in audio data along with data from other sensors like GPS and motion, which is increasingly used to verify the user and their state of mind using a voiceprint. All this extra data allows algorithms to curate your question and give one answer, rather than to 100,000s of answers given in a text search from the previous stage. More data in, less data out, and less of a chance for an unexpected or nuanced finding without further inquiry from the user.

3.3 Smart Glasses and Computer Vision (CV)

Using computer vision (CV), the computer can acquire, process, analyze, and understand digital videos and images. It identifies an object by observing its appearance, location, and settings. This takes us to the third stage in this trend towards the budding botanist paradox.
A pair of smart glasses can overlay virtual objects on the real world using augmented reality (Polvi et al. 2016). This again ups the ante for collected data for the purpose of curating user queries. In addition to all the data in past stages, a larger amount of data is collected passively about a user's surroundings. In the smartphone and virtual assistant stage, we were starting to get prompted with information that was not requested, perhaps with automated notifications related to our past use of our devices. This is another component of taking the inquiry process

out of using these systems. Automation when combined with engagement algorithms, attempt to predict, curb, or nudge our behavior to revisit information, and perhaps to start anticipating our questions and feeding us the appropriate information. There is a possibility for immersive technology, with the aid of computer vision and a heads-up display to start exacerbating this trend (Amin and Govilkar 2015). The nature of this technology would now serve as a passive input device and allow for much more unprompted information in the form of virtual objects and other visual information displays. Because the camera and microphone are passively scanning and always on, the amount of data to be curated again jumps tremendously. In the budding botanist scenario, the heads-up display might start displaying the names of plants in view without asking after asking the device to identify some plants a couple of times. The automated system is now anticipating what questions might be coming up next and displaying them in our view. Other examples are pairing facial recognition to display people's names above their heads, helping us not to remember names just as we don't need to remember phone numbers.

In this stage of automation, the basic inquiry process has become again more dependent on the supporting algorithms that curate and prompt information, even before the user has formed a question.

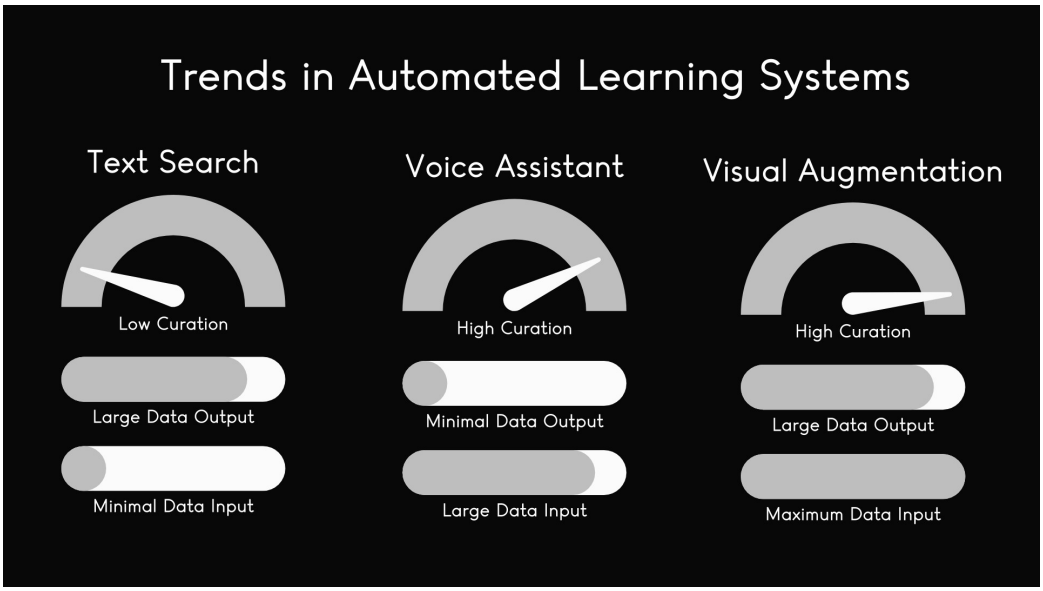


Figure 1. Comparing 3 Stages of Input Methods for Learning

4. Discussion

Covid-19 and the global pandemic accelerated distance learning and the need for developing digital competencies related to automation for teachers and students. To be effective in the classroom, technology must also be incorporated only when it is appropriate for the task at hand. There are some approaches that immersive technologies can offer in an educational context that other media cannot, such as highly accurate simulations of expensive or dangerous or novel environments. It takes an educator experienced in using these digital tools to best implement them. And in the absence of training, may rely more on the general automated or prescribed use and could contribute to the dark patterns and other designs that lie in the business models that shape the automation of many of these newer tools.

5. Conclusion

One driving force in the growth of automation in human inquiry is the evolution of human-computer interface and the move to more empathic computing. A precise and responsive input interface with just enough digital intervention to nudge behaviors. Glasses are challenging the smartphone as the dominant form factor. A simple, accurate interface is required for the augmentation domain. We will adopt new input methods like brainwaves, gaze detection, and muscle signals in the future.

The budding botanist paradox is a thought experiment to attempt to reason that the more automated the learning process becomes, enhanced, and augmented with immersive technology, the more learners could be dependent on these automated systems for basic learning, more susceptible to second order influences in behavior, and the more vulnerable teachers and learners are if/when the automated systems have problems.

References

Adedoyin, Olasile Babatunde, and Emrah Soykan. 2020. "Covid-19 Pandemic and Online Learning: The Challenges and Opportunities." Interactive Learning Environments, September, 1–13.

Amin, and Govilkar. 2015. "Comparative Study of Augmented Reality SDKs." International Journal of Computational Intelligence and Applications.

Bainbridge, L. 1983. "IRONIES OF AUTOMATION." In Analysis, Design and Evaluation of Man–Machine Systems, edited by G. Johannsen and J. E. Rijnsdorp, 129–35. Pergamon.

Bekele, Esubalew, Zhi Zheng, Amy Swanson, Julie Crittendon, Zachary Warren, and Nilanjan Sarkar. 2013. "Understanding How Adolescents with Autism Respond to Facial Expressions in Virtual Reality Environments." IEEE Transactions on Visualization and Computer Graphics 19 (4): 711–20.

Chien, Yu-Cheng, Yen-Ning Su, Ting-Ting Wu, and Yueh-Min Huang. 2019. "Enhancing Students' Botanical Learning by Using Augmented Reality." Universal Access in the Information Society 18 (2): 231–41.

Cowan, Kirsten, Ana Javornik, and Peilin Jiang. 2021. "Privacy Concerns When Using Augmented Reality Face Filters? Explaining Why and When Use Avoidance Occurs." Psychology & Marketing 38 (10): 1799–1813.

Fisher, Matthew, Adam H. Smiley, and Tito L. H. Grillo. 2021. "Information without Knowledge: The Effects of Internet Search on Learning." Memory , February, 1–13.

Hawkinson, E., and J. Klaphake. 2020. "Work-in-Progress—Legal and Ethical Issues in Immersive Education." In 2020 6th International Conference of the Immersive Learning Research Network (iLRN), 305–7.

Hawkinson, Eric, Parisa Mehran, and Mehrasa Alizadeh. 2017. "Using MAVR to Bring New Dimensions to the Classroom." LANGUAGE TEACHER 41: 31.

Hayes, Niall, and Lucas Introna. 2005. "Systems for the Production of Plagiarists? The Implications Arising from the Use of Plagiarism Detection Systems in UK Universities for Asian Learners." Journal of Academic Ethics 3 (1): 55–73.

Oliver, Nick, Thomas Calvard, and Kristina Potočnik. 2017. "Cognition, Technology, and Organizational Limits: Lessons from the Air France 447 Disaster." Organization Science 28 (4): 729–43.

Palmer, Bill. 2013. Understanding Air France 447. William Palmer.

Pappano. 2012. "The Year of the MOOC." The New York Times.

Polvi, Jarkko, Takafumi Taketomi, Goshiro Yamamoto, Arindam Dey, Christian Sandor, and Hirokazu Kato. 2016. "SlidAR: A 3D Positioning Method for SLAM-Based Handheld Augmented Reality." Computers & Graphics 55 (April): 33–43.

Salmon, Paul M., Guy H. Walker, and Neville A. Stanton. 2016. "Pilot Error versus Sociotechnical Systems Failure: A Distributed Situation Awareness Analysis of Air France 447." Theoretical Issues in Ergonomics Science 17 (1): 64–79.

Savage, S. 2004. "Staff and Student Responses to a Trial of Turnitin Plagiarism Detection Software." Proceedings of the Australian Universities Quality. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.187.5717

Shan, Caifeng. 2010. "Gesture Control for Consumer Electronics." In Multimedia Interaction and Intelligent User Interfaces: Principles, Methods and Applications, edited by Ling Shao, Caifeng Shan, Jiebo Luo, and Minoru Etoh, 107–28. London: Springer London.

Strauch, Barry. 2018. "Ironies of Automation: Still Unresolved After All These Years." IEEE Transactions on Human-Machine Systems 48 (5): 419–33.

Tan, Desney, and Anton Nijholt. 2010. "Brain-Computer Interfaces and Human-Computer Interaction." In Brain-Computer Interfaces: Applying Our Minds to Human-Computer Interaction, edited by Desney S. Tan and Anton Nijholt, 3–19. London: Springer London.

White, S., S. Feiner, and J. Kopylec. 2006. "Virtual Vouchers: Prototyping a Mobile Augmented Reality User Interface for Botanical Species Identification." In 3D User Interfaces (3DUI'06), 119–26. ieeexplore.ieee.org.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th Int6er6n0ational Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Coding Initiative Provides Different Approaches to Inspire Girls for Programming

Sarina GURSCH
Institute of Software Technology, Graz University of Technology, Austria
sarina.gursch@ist.tugraz.at

Abstract: Attracting more women into science, technology, engineering, and mathematics continues to be a significant concern in the scientific research community. Especially the demand for specialists in computer science is increasing. However, interest drops significantly among teenage girls. Many coding initiatives aim to increase motivation and enthusiasm and provide young women with a bright future in science, technology, engineering, and mathematics. This paper demonstrates a one-week coding initiative with different approaches to encourage 21 teenage girls to program. The different approaches, game design, embroidery programming, and robotics provide a suitable setting to answer the research questions: 1) Are the girls inspired to program through the coding initiative? 2) Has the girl's self-perception of programming skills improved? 3) Do these approaches arouse the girl's interest? The results show that the girls' enthusiasm and interest are successfully raised. We proved that the girls obtained a fundamental understanding of programming skills and computational thinking. Their understanding of loops, variables, data, if-else statements, objects, and local and global constructs confirms the results. The most interesting approaches are embroidery programming closely followed by game design. These validated and purposeful approaches teach (girls) programming and basic computational thinking skills. Researchers and practitioners will benefit from these findings. The results will advance the field of computing education.

Keywords: girls, programming, coding initiative, STEM, gender

1. Introduction

A gender gap exists in science, technology, engineering, and mathematics (STEM) fields, with fewer women than men studying and working in these areas. Worldwide less than 30% of researchers are female. Only 30% of female students choose STEM studies (UN, 2021). At the same time, these disciplines include a broad spectrum of potential applications in a professional field. In Addition, qualification in science, technology, engineering, and mathematics offer graduates a wide range of careers, which can be an asset both professionally and personally (Gough, 2015). Furthermore, the demand for occupations with skills in the technical field is increasing. The European Union intends to increase the participation of women in all areas of STEM education and its employment market (Garcia-Holgado et al., 2020). The field of computing education is well suited for this purpose. Therefore skills in science, technology, engineering, and mathematics are essential, and women and girls should get the chance to shape their own future in the field of technology. Inspiring young girls to enter these fields remains an essential topic in the scientific community. One way of reducing this gap is by encouraging and supporting women in STEM (Beede et al., 2011). Many initiatives and projects try to encourage young women to pursue a career in science, technology, engineering, and mathematics.

The research paper aims to identify appropriate approaches to encourage (novice) girls to program. To compare these different approaches, the same device (smartphone) and the same programming environment (Pocket Code App) were selected for all three approaches in the framework of the one-week coding initiative.

1) Are the girls inspired to program through the coding initiative?
2) Has the girl's self-perception of programming skills improved?
3) Do these approaches arouse the girl's interest?

2. Related Work

Closing the gender gap is a social and global challenge in computing education. In the case of girls, interest drops when they reach their teens (Stein, 2019), and this does not happen in the case of boys. However, coding initiatives inspire young women to pursue a career in computer science and STEM-related fields. These would be, among others, for example, Webgrrls, Women Who Code, PyLadies, Black Girls Code, Girls Who Code, czechitas, and CODING4GIRLS (De Vuyst, 2020), (czechitas, 2021), (Hoic-Bozic et al.,), (Gursch et al., 2020), (Spieler et al., 2019). A particularly successful approach among these is that of the code clubs for girls only (Zagami et al., 2015). Based on these successful and validated coding initiatives for girls, a coding initiative for girls is best suited to test different approaches to give girls first programming experiences. The teaching approaches vary from Makerspaces (Makerspaces, 2022), robots (Sapounidis and Alimisis, 2020), wearables (Merkouris and Chorianopoulos, 2015), embroidery machines (Wolz et al., 2019), game design (Rugelj and Lapina, 2019), to storytelling (Vinayakumar et al., 2018). Compared to the previous work, this research focuses on different approaches that can be used with one device and the same programming language. Furthermore, it should be a device that the participants are familiar with to reduce possible entry barriers. Therefore, a smartphone is the tool of choice. In addition, most participants already have their own smartphones (smartphoneTeens, 2022), which can be used for the course. The coding initiative uses the Pocket Code app to teach programming. In particular, the programming environment of the Pocket Code app is based on a block structure similar to Scratch (Slany, 2014). Block-based programming languages provide a simple method for acquiring first programming skills at a young age. This reinforces concepts such as loops, if-else statements, data, and others (Armoni et al., 2015). Learning programming fosters computational thinking skills (Plaza et al., 2019). This could yield a career in computer science or STEM-related fields.

3. Methodology

The coding week was from July 12 – 16, 2021, at Graz University of Technology. The five days were held all day, with the adequate course time amounting to three hours in the morning and three hours in the afternoon. The coding initiative was aimed at female teenagers who want to learn to program. However, the course was not limited to beginners. This course was promoted on summer course platforms and flyers. Therefore different approaches (for novices) were prepared that provide the first basics for learning programming in an easy, understandable, and fun way.

The Pocket Code app supports many different approaches. It is possible to create a program, like a game, an embroidery pattern, a control for drones, a control for Lego NXT robots, or more. The Pocket Code app was deliberately selected as the tool for this course. The visual block-based programming environment is a successful way to teach programming to novices (Price and Barnes, 2015). Furthermore, additional hardware such as Lego NXT or embroidery machines were used depending on the approach. The different approaches were assigned to individual stations. The station on game design offers the opportunity to program their own games. The approach embroidery programming deals with the area of producing embroidery patterns. The participants could design patterns and program a file for an embroidery machine. These self-designed and programmed patterns are stitched on a bag or t-shirt with the help of an embroidery machine. In the robotics station, Lego NXT robots were programmed and controlled. Thus, the selected different approaches shape three stations in the coding initiative: game design, embroidery programming, and Lego NXT (robotics), Figure 1.



Figure 1. Game design, embroidery programming, and Lego NXT (robotics).

Each station had an introductory example to give the participants a basic understanding. After completing the first basics, there is enough time to program their games, embroidery patterns, or robots. The stations were under the supervision of trainers. Each trainer created the material for their station based on predefined approaches and previous work. The previous work builds on gender-sensitive materials and research experiences. The focus here is to guide the girls in the course through a gender-sensitive approach and with exclusively female trainers (Guenaga et al., 2022). They act as role models and intermediaries, conveying their passion for technology to the girls. The trainers are students in a STEM-related field at the university and have already achieved excellent programming skills.

3.1 Participants

Some girls signed up for the coding initiative, some through legal representatives. Both the girls themselves and their legal guardians gave their consent for the study. A total of 21 girls participated in the course. The participants ranged from 12 to 18 years, eleven girls were 13 years old. The extent of prior programming knowledge was varied; 25% had never programmed before the course, 50% had learned to program in school, and the remaining 25% had already programmed at home or with friends.

3.2 Evaluation Method

At the beginning and end of the coding initiative, a questionnaire was distributed and filled out by the participants. Comparing the questionnaire data before and after the course, conclusions about the method were drawn. The questionnaire asked, among other things, whether the girls had already programmed. Furthermore, they were asked to rate their interest, the importance of learning programming, or how familiar they are with concepts (loops,...) with a 4-level scale (true, rather true, hardly true, and not true). This information was compared with the results of the quiz. Each participant completed a quiz after finishing each station, once for each different approach. The results served as a control for the participants' self-assessment of the questionnaire's answers. In Addition, individual interviews with the girls and their observations helped verify the questionnaire.

4. Results

The evaluation of the questionnaires, quizzes, and observations address the three research questions:

1) Are the girls inspired to program through the coding initiative?
The evaluations show that the interest in learning programming has increased from 36.8% to 68.4%. Surprisingly, the importance of learning to program decreases, and the initial confidence of the participants to be able to program decreases very slightly. However, a significant increase from 36.8% to 63.2% in programming is fun for the participants. Figure 2 shows more details of the results after the coding initiative.

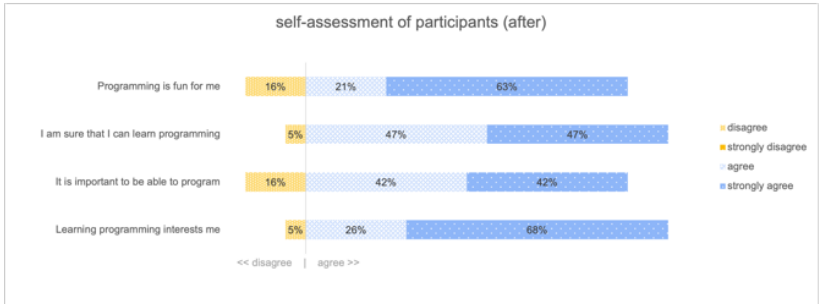


Figure 2. Answers were given in the questionnaire about programming after the coding initiative.

2) Has the girl's self-perception of programming skills improved?

The evaluation of the questionnaires shows that some of the girls had already gained their first experience in programming and computational thinking before the coding initiative. 15.8% of the participants are unfamiliar with loops, whereas 26.3% know what a loop is and can use it in programming. The use of variables or data is unknown for 26.3% of the participants, but an equal number of participants know and use this term. The gap is even more significant for if-else statements. While 36.8% state that they are well aware of this construct and can use if-else statements professionally in programming, 31.6% are not familiar with this term. The area of global and local constructs gets the weakest self-assessment by the participants. Only 5.3% say that they understand these, while a total of 31.6% are not familiar with them. When dealing with objects, quite 42.1% state being able to use them, and only 10.5% do not understand what a construct it is, Figure 3.

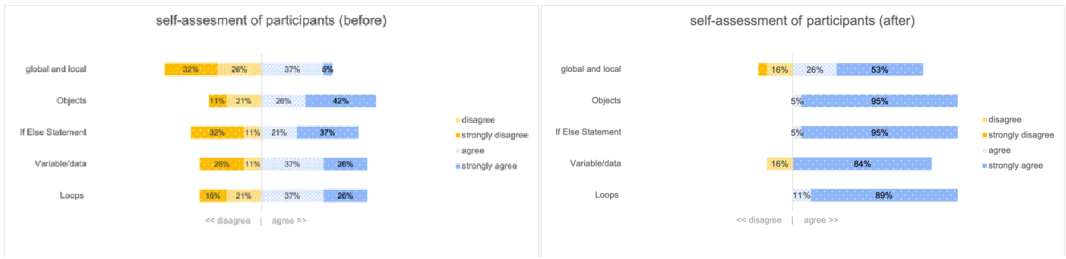


Figure 3. Programming skills before (left) and after (right) the coding initiative.

After the coding initiative, the evaluation of the questionnaire about the skills for loops, variables, if-else statements, objects, and local/global constructs is significantly higher. The lowest percentage in programming skills achieves the participant's understanding of local and global constructs with 52.6%. The programming skills for loops, variables, if-else statements, and objects are generally better assessed, with all participants rating around the 90% mark (see Figure 3). The results show that the participant's self-assessment of their programming and computational thinking skills has risen.

The results of the quizzes are consistent with the participant's self-assessment in the questionnaire about their programming skills. The quizzes reflect a test in which questions are judged based on answers. The questions dealt with programming in the Pocket Code app and comprehension questions about the different concepts in programming. Furthermore, there was a big focus on loops, objects, or how to implement specific shapes in the code. Overall the quiz results confirm that the participants learned to program, and in addition, they also acquired computational thinking skills.

3) Do these approaches arouse the girl's interest?

Before and after the coding initiative, the participants were asked in the questionnaire about their interest in the presented approaches. The interest in programming games decreased slightly. While one station in the coding initiative dealt with robotics (Lego NXT), interest in programming robots fell from 63.2% to 57.9%. A substantial increase can be seen in the programming of embroidery machines. At the end of the course, 68.4% of the girls stated they were interested in embroidery programming, whereas, before the course, the proportion was 31.6%. In addition, the interviews held with the participants after and between the courses confirm that they were very enthusiastic about embroidery programming. The girls created complex programs to embroider elaborate designs on their t-shirts and bags. The questionnaire evaluation matches the observations on whether the participants would have liked to have spent more time in the individual stations.

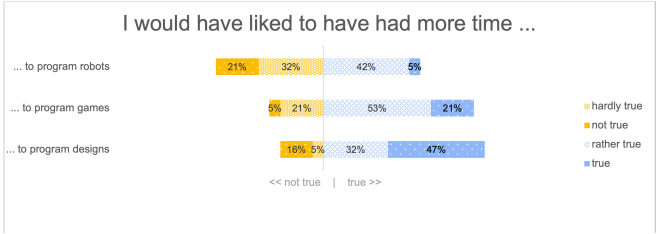


Figure 4: More time for different approaches

The results show that interest was the highest in spending more time on embroidery programming, followed by game design, the station with the least appeal was programming robots, Figure 4. In addition, participants stated that they were proud of their games and the embroidery work they had created during the course. Overall, we found that embroidery programming is a very appealing approach, closely followed by game design.

5. Conclusion

The one-week-long coding initiative was successful in terms of both the results achieved and the efficiency of the applied methodology. The results show that the different approaches, ranging from game design to embroidery programming to robotics (Lego NXT), have aroused the girls' engagement in programming. A total of 21 (primarily novices) girls participated in the coding initiative. Although the sample size is too small to make a significant statement, it became clear that game design and embroidery programming fascinate girls. The approach embroidery programming was top-rated. On the one hand, actual tasks, topics, and problems increase motivation and enthusiasm (Topalli and Cagiltay, 2018). On the other hand, taking home self-created work can arouse excitement because creative projects arouse girls' interest in computer science (Sax et al., 2017). We speculate that embroidery programming might be a particularly high-yield approach to awaken girls' interest in programming. Future studies may repeat this process for a larger sample to broaden our findings. The following steps are enlargement and enrichment. Another dimension by highlighting the role of the different approaches could be performed by enlarging the content by using other different approaches. Future research will enrich this study by deepening the skills in programming. Therefore, teaching even in-depth programming knowledge in an inspiring and fun way.

These findings advance the field of computing education. Educators, practitioners, and researchers will benefit from these results. The different approaches can help to design curriculums or development professional programs. New programmable materials, interactive activities, and traditional and cultural communities of girls significantly motivate them to participate in computing education (Kafai and Burke, 2014). Game design, embroidery programming, and robotics offer excellent opportunities for girls to get (first) experiences with programming. Depending on the availability of hardware like Lego NXT robots or an embroidery machine, a researcher or educator could also choose just one of these approaches. The use of the (own) smartphone offers a particularly outstanding advantage. Mostly the participants use their smartphones, which reduces entry barriers. In addition, this could lead to more opportunities in computing education, where access to a computer or tablet is limited. Although the coding initiative was held exclusively for girls, this does not limit these approaches to heterogeneous groups.

Acknowledgments

This research was supported by Office for Gender Equality and Equal Opportunity of Graz University of Technology and by the European Union as part of the Horizon 2020 research and innovation program - Gender STI project (H2020-SwafS-2019-1, grant agreement No. 872427). I thank my colleagues, Katja Urak and Vesna Krnjic, from Institute of Software Technology (Graz University of Technology), who provided insight and expertise that greatly assisted the research. I would also like to thank the involved students of Graz University of Technology for their support in carrying out the coding initiative.

References

Armoni, M., Meerbaum-Salant, O., and Ben-Ari, M. (2015). From scratch to "real" programming. ACM *Transactions on Computing Education* (TOCE), 14(4):1–15.

Beede, D. N., Julian, T. A., Langdon, D., McKittrick, G., Khan, B., and Doms, M. E. (2011). Women in stem: A gender gap to innovation. *Economics and Statistics Administration Issue Brief*, (04-11).

czechitas (2021). https://www.czechitas.cz/en.

De Vuyst, S. (2020). *Hacking gender and technology in journalism*. Routledge.

Garcıa-Holgado, A., Verdugo-Castro, S., Gonzalez, C., Sanchez-Gomez, M. C., and Garcıa-Penalvo, F. J. (2020). European proposals to work in the gen- der gap in stem: A systematic analysis. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 15(3):215–224.

Gough, A. (2015). Stem policy and science education: Scientistic curriculum and sociopolitical silences. *Cultural Studies of Science Education*, 10(2):445– 458.

Guenaga, M., Eguıluz, A., Garaizar, P., and Mimenza, A. (2022). The impact of female role models leading a group mentoring program to promote stem vocations among young girls. *Sustainability*, 14(3):1420.

Gursch, S., Krnjic, V., Urak, K., Herold, M., and Slany, W. (2020). How to encourage girls to code through embroidery patterns? In *Proceedings of the 4th International Conference on Gender Research.*

Hoic-Bozic, N., Dlab, M. H., Frankovic, I., and Ivasic-Kos, M. Teaching pro- gramming skills to girls.

Jayathirtha, G. and Kafai, Y. B. (2021). Program comprehension with physical computing: A structure, function, and behavior analysis of think-alouds with high school students. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V*. 1, pages 143–149.

Kafai, Y. B. and Burke, Q. (2014). Beyond game design for broadening participation: Building new clubhouses of computing for girls. *GenderIT 2014*.

Makerspaces (2022). How makerspaces in schools help students learn to code. https://edtechmagazine.com/k12/article/2018/09/how-makerspaces-schools-help-students-learn-code-perfcon.

Mathrani, A., Christian, S., and Ponder-Sutton, A. (2016). Playit: Game based learning approach for teaching programming concepts. *Journal of Educational Technology & Society*, 19(2):5–17.

Merkouris, A. and Chorianopoulos, K. (2015). Introducing computer programming to children through robotic and wearable devices. In *Proceedings of the Workshop in Primary and Secondary Computing Education*, pages 69–72.

Plaza, P., Sancristobal, E., Carro, G., Blazquez-Merino, M., Garcıa-Loro, F., Munoz, M., Albert, M. J., Morinigo, B., and Castro, M. (2019). Scratch as driver to foster interests for stem and educational robotics. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, 14(4):117–126.

Price, T. W. and Barnes, T. (2015). Comparing textual and block interfaces in a novice programming environment. In *Proceedings of the eleventh annual international conference on international computing education research*, pages 91–99.

Rugelj, J. and Lapina, M. (2019). Game design based learning of programming. In *Proceedings of SLET-2019–International Scientific Conference Innovative Approaches to the Application of Digital Technologies in Education and* Research, Stavropol–Dombay, Russia, CEUR Workshop Proceedings, volume 2494.

Sapounidis, T. and Alimisis, D. (2020). Educational robotics for stem: A review of technologies and some educational considerations. In *Science and mathematics education for 21st-century citizens: Challenges and ways for- ward, number September, 2020*, pages 167–190. Nova Science Publishers Hauppauge, NY, USA.

Sax, L. J., Lehman, K. J., Jacobs, J. A., Kanny, M. A., Lim, G., Monje-Paulson, L., and Zimmerman, H. B. (2017). Anatomy of an enduring gender gap: The evolution of women's participation in computer science. *The Journal of Higher Education*, 88(2):258–293.

Slany, W. (2014). Tinkering with pocket code, a scratch-like programming app for your smartphone. *Proceedings of Constructionism*.

smartphoneTeens (2022). New research finds 95% of teens have access to a smartphone; 45% online 'almost constantly'. https://www.geekwire.com/2018/new-research-finds-95-teens-access-smartphone-45-online-almost-constantly/.

Spieler, B., Krnjic, V., and Slany, W. (2019). Girls create games: Lessons learned—arXiv *preprint arXiv:1907.05811.*

Stein, B. E. (2019). Stem and adolescent girls.

Topalli, D. and Cagiltay, N. E. (2018). Improving programming skills in engineering education through problem-based game projects with scratch. *Computers & Education*, 120:64–74.

UN (2021). 2021 theme: Women scientists at the forefront of the fight against covid-19. https://www.un.org/en/observances/women-and-girls-in-science- day.

Vinayakumar, R., Soman, K., and Menon, P. (2018). Digital storytelling using scratch: engaging children towards digital storytelling. In *2018 9th International Conference on Computing, Communication and Networking Technologies* (ICCCNT), pages 1–6. IEEE.

Wolz, U., Auschauer, M., and Mayr-Stalder, A. (2019). Code crafting with turtlestitch. In *ACM SIGGRAPH 2019 Studio*, pages 1–2.

Zagami, J., Boden, M., Keane, T., Moreton, B., and Schulz, K. (2015). Girls and computing: Female participation in computing in schools. Australian Educational Computing, 30(2).

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

An Integrative Evaluation of STEM Policy in Secondary Education: Empirical Reflections, Conceptual Discussions and Policy Implications

Percy Lai-yin KWOK [a*] & Heidy Jia-hui WANG [b]
[a] *The Education University of Hong Kong*
[b] *The Education University of Hong Kong*
*lykwok@eduhk.hk

Abstract: Current academic research bodies on STEM education focus on five perspectives of student ability enhancement, teacher training, curriculum integration, pedagogy and curricular implementation in intra-national / cross-national comparative studies. Noteworthy, there has been no comprehensive policy analysis of STEM education in secondary school curricula in Hong Kong. For STEM is not formally documented into school curricula and students mostly join inter-school and international STEM project competitions. International literature reveals that conceptions of STEM education are still fuzzy. In this paper, a 2 X 3 matrix is utilized to evaluate the extent of implementing STEM education in mainstream secondary schools in Hong Kong, and to envisage contextual difficulties and obstacles against STEM education. This paper then leads conceptual discussions on integrated approaches, draw policy implications and recommend further research agendas.

Keywords: STEM education, policy analysis, ICT integration, teacher training, curriculum

1. Introduction

The concept of STEM education emerged in the United States in the 1990s in the hope of nurturing the talent in science and technology through STEM education, enhancing the scientific literacy of the nation, and maintaining a leading position in technology and the economy worldwide. Afterwards some Asian, Australasian and European countries followed the suit, STEM encompasses a multi-disciplinary integration of **S**cience, **T**echnology, **E**ngineering, and **M**athematics. However, such four individual disciplines are not simply superimposed, but are united to form an organic whole to better cultivate students' innovation and exploration and practical skills (Curriculum Development Council, 2015). In primary and secondary school curricula, STEM education is not a stand-alone subject in Hong Kong, but is implemented in cross-learning areas/subjects through the existing science, technology and mathematics education learning areas (The Government of Hong Kong Special Administrative Region, 2020). STEM education fosters student creativity and divergent thinking along with the foundational disciplines, inspires young people to generate new technologies and ideas, and encourages application of knowledge which are the needs of 21st century skills (Norris, 2021).

2. Literature Review

Heated research issues in STEM education include availability of various student scaffolding supports in STEM education (Qori, 2020), teachers' self-efficacy and concerns about STEM education (Geng, Jong & Chai, 2018), the design-thinking professional development of STEM teachers (Chiu, *et al.*, 2021; Wu, *H*u & *W*ang, 2019), students' interdisciplinary (withing the boundary or scope of existing subject disciplines) and transdisciplinary (transcending the boundary or scope of existing subject disciplines) skills in STEM education in school contexts (Drake *et al*., 2015; Wu, Cheng & Koszalka, 2021), parenting influence in students' career aspirations related to STEM (Lloyd *et al.*, 2022), how STEM cultural capital shapes students' STEM careers aspirations (Chen *et al*., 2022) and the emergence of creative problem-solving and systems-thinking skills in and through STEM programs in international and global perspectives (Wang, Shen & Chao, 2021).

Based on review of local and international research literature, the definition and conceptions of STEM education are still fuzzy. Existing forms of implementing STEM education in primary and secondary schools in worldwide can be divided into the three following domains:

1. *Formal curricula:* Integrating STEM elements or related e-learning platform into formal instructions in cross-curricular subjects or new STEM subjects *in schools* (e.g. Li *et al.*, 2019; Mystakidis & Christopoulos, 2022)
2. *Informal curricula:* Integrating STEM elements into school-based curriculum through informal project-based learning or other after-school activities *in schools* (e.g. Kelley & Knowles, 2016; Thingwiangthong *et al*., 2021)
3. *Extensional / instrumental uses (without curriculum development):* Implementing STEM education as a path or means to enhance students' interests, abilities and skills since STEM education is not formally documented into school curricula, such as merely recommending students to participate into inter-school or international competitions *outside schools* (e.g. Ziu & Lou, 2022)

At the same time, there are two senses of integration approaches during / beyond school hours:

A. *New / expandable learning contents:* Incorporating elements or ideas such as AI/Robotics, coding and computational thinking into STEM education or curriculum (e.g.Wang, Shen & Chao, 2021).
B. *Cross-level knowledge transfer:* Transferring advanced and high-level STEM knowledge at tertiary level into the existing curriculum in primary and secondary levels through simplification models or data visualization in using computer apps or through AR/MR/VR/XR environments (e.g. Krajcik & İbrahim, 2017; Ortiz-Revilla, Greca & Adúriz-Bravo, 2019).

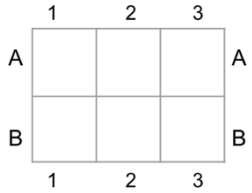


Figure 1. A 2 X 3 matrix of integrative evaluation of STEM curriculum

3. Research Design

3.1 Research Questions

In Hong Kong, previous policy analysis was limited to mere examination of policy documents, instructional design of teacher education programmes run by universities and annual teacher professional development programmes run by EDB (So, *et al.*, 2022). So far there has been no comprehensive analysis of how STEM or STEM-related policies implemented in secondary schools. Four research questions in this paper focus on:

1. How has the Hong Kong SAR government supported STEM education and how has she shaped STEM curriculum development in mainstream secondary schools?

2. How have STEM education policies been implemented in mainstream secondary schools in Hong Kong?

3. How has the implementation of STEM education in mainstream secondary schools in Hong Kong improved students' personal skills in teacher perceptions?

4. What are the contextual barriers against the development of STEM education when it has been implemented in mainstream secondary schools in Hong Kong?

3.2 Research Methodology and Methods

The research methods utilized in this study included an experienced mainstream secondary school teachers' e-survey (N=41), in-depth semi-structured interviews with some key informants who are educationalists and pioneers to promote and lead STEM education in Hong Kong and a documentary analysis of related important policy documents, school circulars and updated website information about STEM education in Hong Kong. In order to enhance its credibility, this study adopts cross-data, cross-method and cross-perspective triangulation (Denzin, 2009) to find some convergent data patterns.

4. Research Results

4.1 How the HKSAR Government Has Supported / Facilitated STEM education

A secondary data analysis of policy documents and EDB (2016, 2020) reports finds that the HKSAR government decentralizes her policy in STEM education by asking mainstream schools to apply for lifelong learning grants and other (non)recurrent school grants to buy STEM resources and encouraging some proactive schools to bid for school-based Quality Education Fund (QEF) projects to testify feasible directions of STEM education. For EDB must provide more flexible educational resources to cater for schools' diversified needs through school-based funding and granting supports, reflected in some interviews with the experienced ex-staff members of Curriculum Development Institute (CDI) of EDB. Noteworthy, the number of STEM projects in QEF was the greatest and the financial expenditure of QEF reached the peak in 2018-19, four times that in 2017-18 [without the public release of official EDB data in 2019-2022].

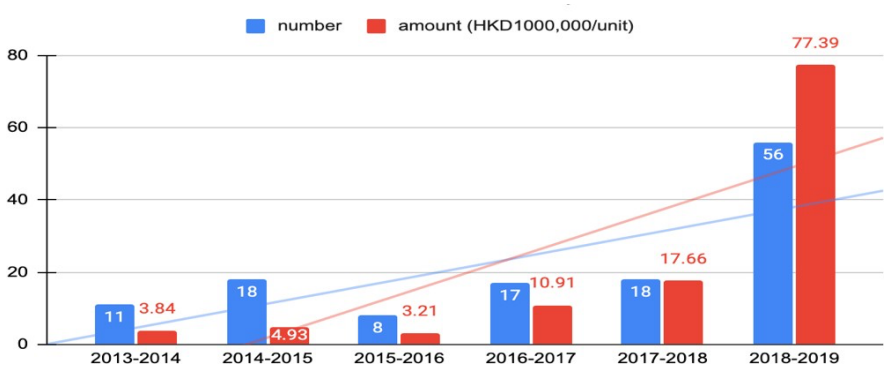


Figure 2. Number and Amount of STEM Projects of Secondary Schools Funded by QEF in 2013-2019

Figure 2 reveals that both sets of number of QE projects and funding data showed an overall upward trend, with the project amount increasing gradually at a faster rate starting from 2016-17 onwards. In contrast with a stable number of QEF projects in 2013-2018, the number of QEF projects increased rapidly to 56 in 2018-2019. It was only after 2015-2016 that QEF projects with STEM-related names and project descriptions gradually emerged due to the publicity of policy addresses. A content analysis of several large-scale yearly education expos and festivals also reveals that the number of demonstration booths and thematic talks related to STEM education or resources has had sharp rises from 2017-18 onwards. Figure 2 also shows vividly the overall increasing trend of the Hong Kong government's funding support for STEM-related projects through QEF from 2013 to 2019. This trend was in line with perceptions with those interviewees who have promoted STEM education in government, school and NGO sectors. In other words, the Hong Kong SAR government and some secondary schools in Hong Kong are paying more attention to STEM education in general. In particular, the rapid increase in STEM-supported fund following the inclusion of STEM education policies in the government policy addresses in 2015, 2016 and 2017 has had a positive effect and helped some secondary schools strengthen Mathematics and Science curricula through STEM education.

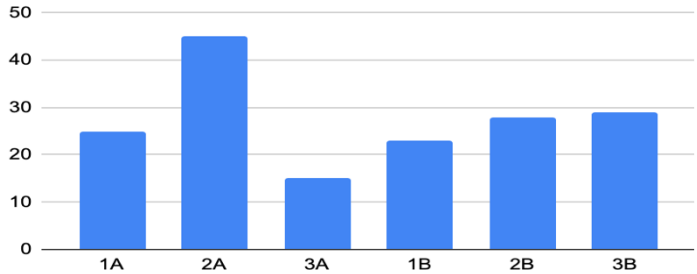


Figure 3. Number of QEF-funded STEM programs by types in secondary schools in 2013-2019

According to the data of the programs sponsored by QEF in the academic years 2013-2019, the number of STEM programs in those participating secondary schools corresponded closely to the curriculum integration approach in the matrix in figure 1 (except for some programs without detailed project descriptions). Notably, figure 3 reflects that QEF-funded projects focused more on 1 and 2 curriculum integration approaches (by comparing 1A + 1B, 2A + 2B with 3 A+ 3B figures) whereas

most secondary schools used 2A as the main curriculum integration approach of STEM education through informal project-based learning or other after-school activities with some new / expandable learning topics like coding, computational thinking, AI & robotics, reported by those project descriptions.

4.2 Implementation of STEM Education in some mainstream secondary schools

According to the EDB's school inspection annual reports, quite a number of schools set up special committees to promote STEM education by coordinating tasks relating to the organization of activities, allocation of resources and professional development of teachers. Examples of these tasks included coordinating relevant subject panels to organize STEM education-related activities and deploying resources for the purchase of equipment to encourage students to participate in STEM activities. Schools with more comprehensive planning even strengthened cross-curricular or cross-subject collaboration, and systematically incorporated the elements of STEM education into the school curriculum and related activities. They also introduced external professional support and professional exchanges across schools to promote a collaborative culture of sharing among teachers in STEM education. Schools tended to design experiential learning STEM activities based on the learning of one subject, and then infiltrate the learning elements of other subjects.

In the 2019-20, more schools were implementing STEM education in the form of cross-curricular project studies. Many schools have also incorporated coding or programming elements into their curricula so that their students can apply programming to their home or resolve daily life problems. Besides, the EDB inspected schools have organized a variety of life-wide learning activities to promote STEM education, such as STEM Education Day / Week, robotics production and exhibition, etc. On the other hand, to enrich students' other learning experience, the schools also arranged advanced training programs for students who were more interested or capable of doing STEM education and recommended them to participate in some external local inter-school and international competitions. However, a transdisciplinary STEM subject or curricula was not commonly set up in those inspected schools through close interviews with some ex-chairmen of professional teacher associations in Mathematics and Computer Education and some experienced school teaches and ICT dept heads in the study. As a whole, STEM education in those primary and secondary schools was implemented through cross-learning areas or cross-/inter-disciplinary collaborations, mainly in the planning of life-wide learning activities, with lesser coordination and collaborations in the curriculum level, which mainly lie in the items 2A, 3A, 2B and 3B in the matrix.

In the e-survey, 63.4% of 41 teacher respondents in a purposive sample of mainstream secondary education sector (4.9% from government secondary, 80.5% from aided secondary and 14.6 % from Direct Subsidy Scheme schools) have been teaching for more than 16 years. A majority (66.70-83.30%) generally agreed or strongly agreed that the current STEM education policy in their secondary schools not only enhances students' interests in learning Mathematics and Science subjects, but also helps them foster their creativity and imagination, problem-solving skills, teamwork, leadership and communication skills, promoting their critical thinking, and advancing their future personal development in figure 4.

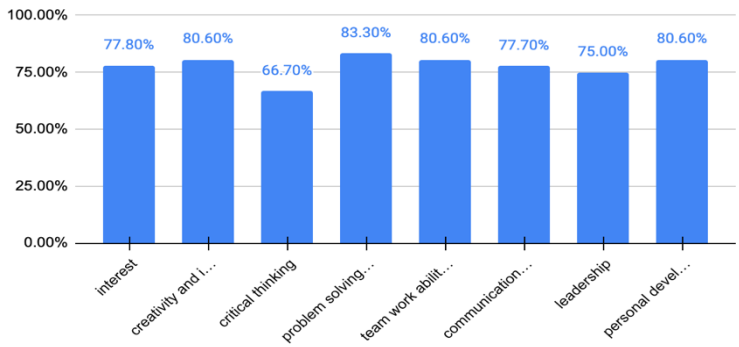


Figure 4. Teachers' opinions about how STEM education policies help improve student abilities

Meantime, 72.9% of survey respondents indicated that their working schools have increased funding in the related subjects after the STEM education policy was implemented. 77.5 % respondents

showed that extracurricular STEM activities in their schools have enriched student learning in related subjects and 75 % stated that their teaching times have been extended in the related subjects after the STEM education policy was implemented. In those mainstream secondary schools, they usually use extracurricular activities to promote STEM education.

4.3 How implementation of STEM education improves students' personal skills in teacher perceptions

Most e-survey and interview respondents described their rich supervision experiences in helping their students join inter-school STEM competitions or in preparation for local and international competitions. This phenomenon often reflects student enthusiasm for participation, creativity, collaboration, communication, and their hands-on skills and leadership, and even data visualization / simplification process and visual or visual presentation skills. This suffices to show that many secondary school students are improving STEM skills by engaging themselves in internal or external competitions beyond school hours. In addition, the rest of the activities mentioned by the respondents such as making robots, manipulating robots, designing rocket cars, smart city design, and so forth indicate that some secondary schools in Hong Kong have incorporated high-tech elements such as AI/Robotics into STEM teaching, which also embodies a form of reinvention of STEM curricula in 2A, 2B and 3A, 3B integration approaches.

4.4 Contextual Barriers and Obstacles Against the Development of STEM education

In-depth semi-structured interviews with those key informants found that besides local and international inter-school competitions, STEM education has been implemented mostly into junior secondary curricula using school-based afterschool activities or some STEM elements have been integrated into computer literacy or computer studies subjects, facing open-examination-driven upper secondary curricula. Finally, those surveyed frontline experienced secondary school teacher respondents revealed lack of instep teacher training programs in enhancing students' STEM skills (which was converged by some interview data), teachers' urge for increasing the publicity of STEM education to gain societal recognition, and their need for development of more school-based STEM curricula in some pilot schools as STEM education has not been fully integrated into their current curricula in their perceptions during the research time. Besides the need for conceptualizing STEM learning in terms of scientific inquiry, technological application, engineering design, and mathematical processing, there is also a need to conceptualize ingredients of teacher training program in STEM education.

5. Conclusions and Implications

5.1 Conclusions

This paper provides a data-supported integrative 2 X 3 matrix using parameters of formal curricula, informal curricula, extensional / instrumental uses, new / expandable learning contents and cross-level knowledge transfer. In fact, such evaluative matrix is a measurement indicator for STEM education policy to further strengthen or improve secondary (especially science and mathematics) curricula in Hong Kong and cross-national perspectives. Within this matrix, the HKSAR government supports schools to explore their own development paths through the bottom-up approach in financial support, such as the Quality Education Fund and other (non-)recurrent school grants.

5.2 Implications

Further research agendas in STEM education in Hong Kong should focus on:

- Learning models or theory for students or student agency to adopt deeply integrated (e.g. transdisciplinary) approaches in STEM education, instead of teacher or school facilitation
- Teacher professional development and pedagogical models for taking deeply integrated approaches
- School or teacher leadership models for initiating and sustaining reforms for adopting deeply integrated approaches (or transforming from interdisciplinary to transdisciplinary approaches in particular) in STEM education (c. f. Liao, 2016)
- Nourishing collaborative learning communities-of-practices among schools, teachers and students for sustainable development or growth of deeply integrated approaches in STEM education

References

Chen, Y., Chiu, S. W. K., Zhu, J. X., & So, W. W. M. (2022). Maintaining secondary school students' STEM career aspirations: the role of perceived parental expectations, self-efficacy, and cultural capital. *International Journal of Science Education*. DOI:10.1080/09500693.2022.2032463

Chiu, T. K. F., Williams, P. J., Chai, C. S., & Lin, T. J. (2021). Teacher professional development on self-determination theory-based design thinking in STEM education. *Educational Technology & Society*, *24*(4):153–165.

Curriculum Development Council (2015). *Brief on stem (overview) CHI 20151105 R - Education Bureau.* https://www.edb.gov.hk/attachment/tc/curriculum-development/renewal/STEM/STEM%20Overview_c.pdf

Denzin, N. K. (2009). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Drake, S. M., Savage, M. J., Reid, J. L., Bernard, M.L., & Beres, J. (2015). *Research summary: An exploration of the policy and practice of transdisciplinarity in the International Baccalaureate Primary Years Programme.* Brock University. https://www.ibo.org/globalassets/publications/ib-research/pyp/pyp-transdisciplinarity-summary.pdf

Education Bureau (EDB) (2016). *Report on promotion of STEM Education - unleashing Potential in Innovation*. https://www.edb.gov.hk/attachment/en/curriculum-development/renewal/STEM%20Education%20Report_Executive%20Summary_Eng.pdf

Education Bureau (EDB) (2020). *Annual Report on Optometry.* https://www.edb.gov.hk/attachment/tc/sch-admin/sch-quality-assurance/reports/insp-annual-reports/C-Ch3_pri.pdf [Chinese]

Geng, J., Jong, M.S., & Chai, C.S. (2018). Hong Kong teachers' self-efficacy and concerns about STEM education. *The Asia-Pacific Education Researcher*, *28*, 35-45.

Kelley, T. R., and Knowles, J. G. (2016). A conceptual framework for integrated STEM education. *International Journal of STEM Education*, 3(1), 11. https://doi-org.ezproxy.lib.purdue.edu/10.1186/s40594-016-0046-z

Krajcik, J., & İbrahim, D. (2017). Õpilaste kaasamine STEM-haridusse. Eesti Haridusteaduste Ajakiri. *Estonian Journal of Education, 5*(1), 35-58. https://doi.org/10.12697/eha.2017.5.1.02b

Lietaert, S., Roorda, D., Laevers, F., Verschueren, K. and De Fraine, B. (2015). The gender gap in student engagement: The role of teachers' autonomy support, structure, and involvement. *British Journal of Educational Psychology, 85*, 498-518. https://doi.org/10.1111/bjep.12095

Liao, C. (2016). From interdisciplinary to transdisciplinary: An arts-integrated approach to STEAM education. *Art Education*, 44-49. https://people.potsdam.edu/betrusak/521/Transdisciplinary%20STEAM.pdf

Li, Y., Schoenfeld, A. H., diSessa, A. A., Graesser, A. C., Benson, L. C., English, L. D., & Duschl, R. A. (2019). *On thinking and STEM education. Journal for STEM Education Research, 2*(1), 1–13. https://doi.org/10.1007/s41979-019-00014-x

Lloyd, A., Gore, J., Holmes, K., Smith, M., & Fray, L. (2022). Parental influences on those seeking a career in STEM: The primacy of gender. *International Journal of Gender, Science and Technology,* 10 (2), 308-328. http://genderandset.open.ac.uk/index.php/genderandset/article/viewFile/510/959

Mystakidis, S., & Christopoulos, A. (2022). Teacher Perceptions on Virtual Reality Escape Rooms for STEM Education. *Information, 13* (3), 136. https://doi.org/10.3390/info13030136

Norris, L. (2021). *Promoting 21st century skills - teaching English*. Promoting 21st century skills. https://www.teachingenglish.org.uk/sites/teacheng/files/J105_12_Promoting_21_Skills_web_0.pdf

Qori, P. H., Sudarmin, S., Sumarni, W., Subali, B., & Saptono, S. (2020). Implementation of STEM integrated ethnoscience-based vocational science learning in fostering students' higher order thinking skills (HOTs). *International Journal of Active Learning, 5*(2), 53-61.

Ortiz-Revilla, J., Greca, I. M., & Adúriz-Bravo, A. (2019). The philosophy in/of integrated STEM education. *Re-introducing science Sculpting the image of science*, 509.

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*(4), 493–529. https://doi.org/10.3102/0034654311421793

So,W.W.M., Lee, C. K. J., Chen, Y., Luo, T., He, Q., Ching, F., N. Y., Li, J.,& Xie, Q. (2022). Teacher education and professional development in STEM education: Experiences in mainland China and Hong Kong. In C. K., J. Lee, & Ehmke, T. (Eds.), *Quality in teacher education and professional development: Chinese and German perspectives.* (pp. 124-138). Abingdon, Oxon: Routledge.

The Government of the Hong Kong Special Administrative Region. (2020). *LCQ17: Promotion of STEM education.* https://www.info.gov.hk/gia/general/202010/21/P2020102100416.htm?fontSize=1

Thingwiangthong, P., Termtachatipongsa, P., & Yuenyong, C. (2021). Status quo and needs of STEM Education curriculum to enhance creative problem solving competency. *In Journal of Physics: Conference Series,* 1835 (1), 012089. IOP Publishing.

Wang, C., Shen, J., & Chao, J. (2021). Integrating computational thinking in STEM education: A literature Review. *International Journal of Science and Mathematics Education,* 1-24.

Wu, B., Hu, Y. & Wang, M. (2019). Scaffolding design thinking in online STEM preservice teacher training. *British Journal of Educational Technology, 50* (5), 2271-2287. https://doi.org/10.1111/bjet.12873

Wu, Y., Cheng, J., & Koszalka, T. A. (2021). Transdisciplinary approach in middle school: A case study of co-teaching practices in STEAM teams. *International Journal of Education in Mathematics, Science, and Technology (IJEMST), 9*(1), 138-162. https://doi.org/10.46328/ijemst.1017

Ziu, G., & Lou, G. (April 1, 2022). Cultivating Stem talents from K12 basic education. *Hong Kong Economic Journal*. [Chinese]

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

A Pedagogy in STEM Classrooms for Primary Students to Develop Knowledge of Electric Circuits and Problem-Solving Skills

Siu Cheung KONG [a*] & Wai Ying KWOK [b]
[a]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong*
[b]*Centre for Learning, Teaching and Technology, The Education University of Hong Kong, Hong Kong*
*sckong@eduhk.hk

Abstract: This study trialed a pedagogy of "To Play and Learn, To Inquire, To STEM and Code". Supplementary with seven hands-on activity worksheets and 19 coding product templates and samples, the pedagogy was designed to expose students to the construction of electric circuits with Micro:bit in STEM classrooms for developing knowledge of electric circuits and problem-solving skills. The pedagogy was implemented as a 280-minute teaching in eight lessons for 23 selected Grade 5 classes, which involved 369 students from nine primary schools at Hong Kong. The results of pre-post-test validate the success of the pedagogy in students' statistically significant growth in all three topic-specific knowledge points. The results of questionnaire survey demonstrate the impact of the pedagogy on students' statistically significant growth in their awareness of and their confidence in applying the target problem-solving skills. The results of focus group interviews confirm that students were satisfied with the experiences and outcomes of the pedagogy in developing topic-specific knowledge and problem-solving skills. This study reveals that the pedagogy is well received by students; noticeably effective to enhance students' knowledge of electric circuits; and potential to promote their development of problem-solving skills. Future directions to extend the pedagogy for senior primary curriculum are discussed.

Keywords: coding, electric circuits, Micro:bit, pedagogy, primary schools, problem-solving, STEM education

1. Introduction and Background of Study

Science, Technology, Engineering and Mathematics (STEM) education and coding education are the two growing curriculum trends in school education in the digital era to nurture the young generation to be problem-solvers with a computer scientist mindset for the success in the digitalized society (Lee, Grover, Martin, Pillai, & Malyn-Smith, 2020; Priemer et al., 2019). Coding is advocated to be a pedagogy potential to foster students to develop capability to solve interdisciplinary problems using techniques from computing (García-Carrillo, Greca, & Fernández-Hawrylak, 2021; Li et al., 2020). This study addressed these trends to innovate a pedagogy which supported senior primary students to develop knowledge of electric circuits and problem-solving skills through STEM classroom activities on constructing electric circuits with Micro:bit.

"Electric Circuits" is a common learning topic in science-related subjects in primary school curriculum around the world; and in Hong Kong this learning topic is structured under the General Studies subject curriculum at senior primary grades (Curriculum Development Council, 2017; Fokides & Papoutsi, 2019). The scope of this learning topic covers knowledge points in three main areas: (i) about electronic components, electronic symbols, and closed circuit; (ii) about conductors, insulators, resistance, series circuit and parallel circuit; and (iii) about the application of electric circuits in daily life (Fokides & Papoutsi, 2019; Preston, Hubber, Bondurant-Scott, & Gunesekere, 2020). The prevailing way to teach this learning topic is to engage students in hands-on activities of constructing basic electric circuits, to enable students to experience and explore the use of electronic components and so to observe and understand the happenings and reasons of different phenomena of electric circuits (Cederqvist, 2022; Preston et al., 2020). STEM activities – which emphasize students' work on

applying knowledge of the four disciplines in STEM domain for designing and making tangible artifacts to solve problems – become a popular lesson component for teaching this topic.

Problem-solving is the prime goal in the process of designing and making tangible artifacts in STEM activities; and therefore, researchers such as García-Carrillo et al. (2021) and Morado, Melo, and Jarman (2021) recognize that it is a natural fit for students from an early age as in primary schools to develop problem-solving skills through the engagement in STEM activities. When students design and make tangible artifacts in STEM domain, they are commonly required to demonstrate the practices of *sequencing* to order steps in a logical manner, *causal reasoning* to identify the relationship between a cause and its effect, *conditional reasoning* to identify the relationship between a decision and its condition, and *engineering systems thinking* to perceive how a component functions as part of a system (Cederqvist, 2022; Lee et al., 2020).

The integration of coding elements into STEM activities is suggested potential to support primary students to develop both domain-specific knowledge and problem-solving skills for their everyday success in the digital era (Fokides & Papoutsi, 2019; Lee et al., 2020). The pedagogy should address the component-related concern on giving students a hands-on experience in using and assembling the actual pieces of artifact components to construct tangible artifacts to solve real-life problems (Blackley & Howell, 2019; García-Carrillo et al., 2021); and the coding-related concern on allowing students to code microcontrollers in visual programming environments to control digital devices to solve real-life problems (Blackley & Howell, 2019; Li et al., 2020).

Micro:bit is a type of single-board microcontroller developed to promote coding among young generation (Cederqvist, 2022; Trilles & Granell, 2020). Micro:Bit is a pocket-sized computer composed of an LED light display, buttons, sensors and a series of input-output features. This tangible programming material is easy for young children to use; and becomes popular for the educational use in primary schools to foster students to develop competencies emphasized in the STEM domain in the digital era (Cederqvist, 2022; Morado et al., 2021). Micro:bit is recognized to be helpful for boosting young students' computational actions in the classroom activities which design and code tangible artifacts as programmable solutions for solving real-life problems, in which students are engaged in analyzing phenomena and contextualizing problems which need programmable solutions; as well as planning and sequencing the construction and coding of tangible artifacts in the programmable solutions (Cederqvist, 2022; Trilles & Granell, 2020).

For the design and implementation of STEM activities which engage primary students in building programmable artifacts for solving real-life problems, researchers suggest three main criteria for the pedagogical designs. First, students should be provided with enough chances to work on the selected programmable artifacts to explore domain-specific knowledge and stimulate their interest in coding (Blackley & Howell, 2019; Fokides & Papoutsi, 2019). Second, students should be provided with enough chances to analyze the problems and the conditions behind the solutions addressed by the selected programmable artifacts for a further thinking about technological solutions with programming materials for solving domain-specific problems in daily life contexts (Cederqvist, 2022; Preston et al., 2020). Third, students should be provided with enough chances to apply and consolidate domain-specific knowledge through building and coding tangible artifacts as programmable solutions for solving real-life problems (Blackley & Howell, 2019; Morado et al., 2021).

2. The Study

2.1 Research Design

This study trialed a pedagogy which supported students to develop both knowledge of electric circuits and problem-solving skills via a three-step pedagogy "To Play and Learn, To Inquire, To STEM and Code" together with seven hands-on activity worksheets and 19 coding product templates and samples (see Figure 1) in eight lessons lasting 280 minutes.

"To Play and Learn" step	**"To Inquire" step**	**"To STEM and Code" step**
(Guiding students to try and learn the connection of electronic components as well	(Guiding students to explore the conditions for controlling the smart light and think about the	(Guiding students to build an artifact of a traffic light system, involving coding with Micro:bit to reuse and remix codes from

as observe the operation of the smart light prototype)

reasons of different phenomena in controlling the smart light)

the smart light into the traffic light system for enabling automation features)

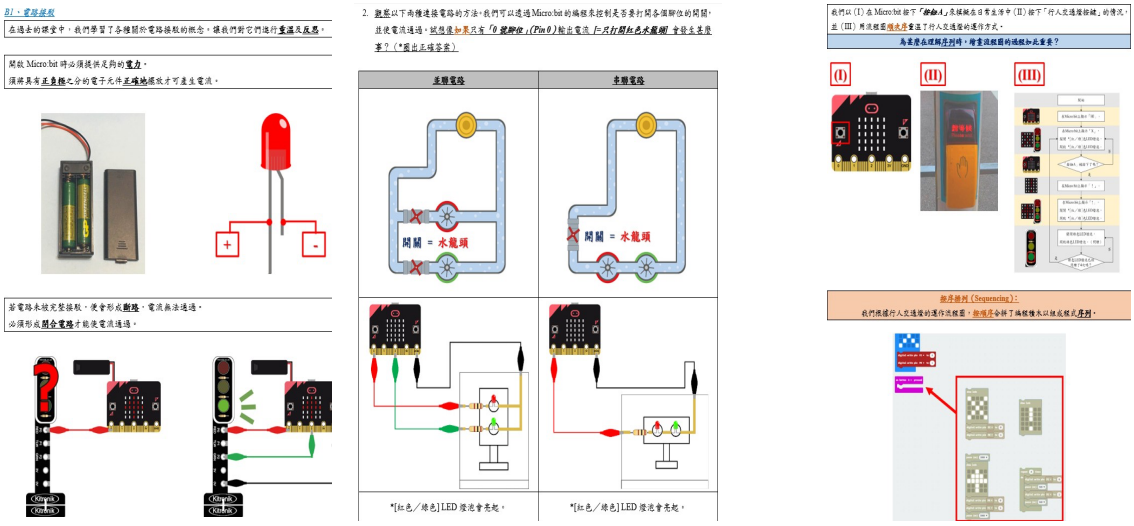


Figure 1. Samples of hands-on activity worksheets in the "To Play and Learn, To Inquire, To STEM and Code" steps in STEM classrooms.

Students coded with Micro:bit in topic-specific hands-on activities for learning the fundamentals for constructing basic electric circuits (i.e., knowledge of electronic components, electronic symbols, and closed circuit); the elements for altering closed circuits construction (i.e., knowledge of conductors, insulators, resistance, series circuit and parallel circuit); and so the application of knowledge of electric circuits for designing and making STEM artifacts. The three-step pedagogy first exposed students to "play" a smart light prototype and "learn" to distinguish between closed and open circuits; differentiate conductors and insulators; and connect electronic components for constructing basic electric circuits. It then arranged students to try to adjust the smart light to "inquire" about the concepts of electric current in electric circuits – specifically, to think about and generalize the conditions of controlling the smart light, and so to abstract the computational ways of controlling smart electrical devices. It finally exposed students to apply the learned knowledge to build a "STEM" artifact of a traffic light system, of which students were arranged to "code" with Micro:bit to reuse and remix codes from the smart light into a traffic light system for enabling features of automating the smart electrical device.

The pedagogy covered four elements of problem-solving skills – namely, sequencing, causal reasoning, conditional reasoning, and engineering systems thinking. Figure 2 illustrates how the hands-on activities under the designed pedagogy supported students to develop knowledge of electric circuits and problem-solving skills. Students first developed an understanding that the proper *sequencing* of *constructing a closed circuit* must start with turning off all switches of the *electronic components* to avoid the risk of electrical injury caused by *electric current* passing through the body. Students then tried to connect various objects of *conductors and insulators* in the smart light prototype; and observed the status of bulb-lighting in the electric circuits with different ways of connecting the electronic components, for their *causal reasoning* of the relationship between *the choice of electronic components and the status of bulb-lighting* when operating the smart light prototype. Students then tried to connect a resistor in the smart light prototype; and observed the change of brightness of bulb-lighting in the electric circuits when adjusting the smart light prototype, for their *conditional reasoning* of the relationship between *the use of resistor and the flow of electric current* when operating the smart light prototype. Students finally add the *microcontroller* Micro:bit as a *part of the system* of the smart light prototype for demonstrating the skill of *engineering systems thinking*.

[1] Students first developed the knowledge of *basic electronic components* and the skill of *sequencing*; then the knowledge of *conductors and insulators* and the skill of *causal reasoning*.

[2] Students then developed the knowledge of *resistance in series circuit and parallel circuit* and the skill of *conditional reasoning*.

[3] Students finally applied the knowledge of *electronic components and circuit connections* and the skill of *engineering systems thinking*.

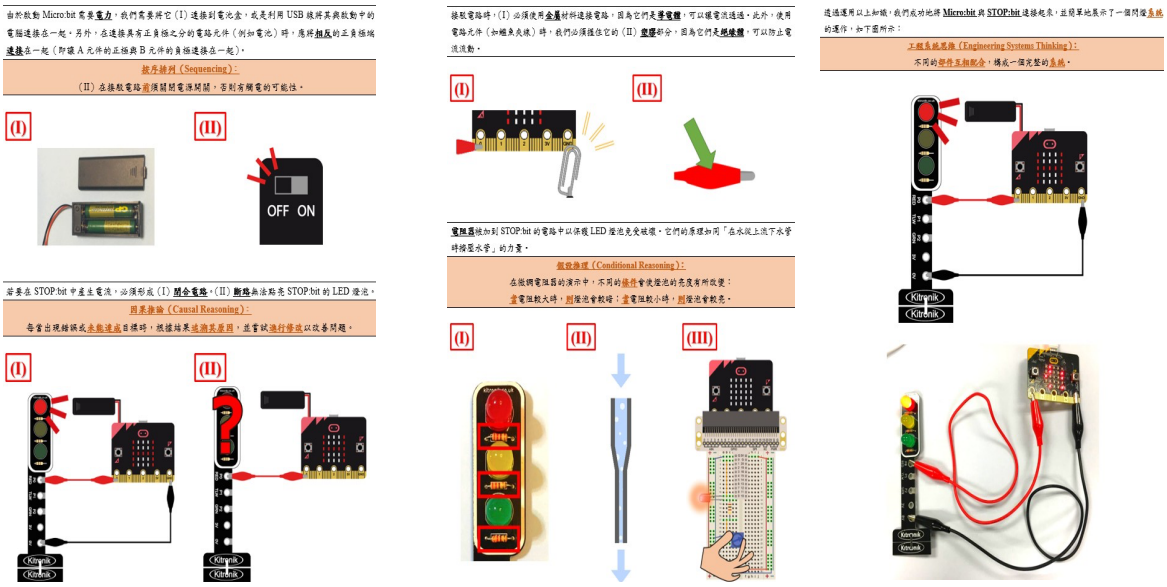


Figure 2. A sample task of hands-on activities in STEM classrooms for developing students' knowledge of electric circuits and problem-solving skills.

2.2 Research Methodology

There was a total of 369 students from 23 Grade 5 classes in nine primary schools at Hong Kong participated in this study (see Table 1). The General Studies teachers of the nine participating classes implemented the pedagogy on a class-specific manner. A three-hour training workshop was arranged for the participating teachers before the trial teaching for preparing them to integrate the rationale of learning through coding in local General Studies curriculum; implement the "To Play and Learn, To Inquire, To STEM and Code" pedagogy in STEM classrooms; and use the seven hands-on activity worksheets and the 19 coding product templates and samples in topic-specific lessons. This study evaluated the pedagogy via three methods for addressing two research questions: (1) How did the students achieve in developing knowledge of electric circuits under the pedagogy? (2) How did the students perceive the pedagogy for developing knowledge of electric circuits and problem-solving skills?

Table 1. *Profile of Students Participated in This Study*

	School 1	School 2	School 3	School 4	School 5	School 6	School 7	School 8	School 9
No. of students	61	62	33	48	18	32	31	25	59
No. of classes	4	4	2	2	1	2	2	1	5
Boys : Girls	30:31	31:31	18:15	26:22	6:12	19:13	16:15	16:9	36:23
Mean age (years)	9.98	9.98	9.71	9.81	10.65	10.20	9.81	9.91	10.12

First, pre-post-tests were conducted at the beginning and the end of the pedagogy. It investigated students' achievement in knowledge of electric circuits. The test papers contained eight questions: two on electronic components, electronic symbols, and closed circuit; five on conductors, insulators, resistance, series circuit and parallel circuit; and one on design of an electronic alarm with the application of knowledge about electric circuits. The pre- and post-test scores of students were compared statistically with the assistance of SPSS software. The Cronbach's alpha reliability coefficients for the pre-test and post-test are 0.74 and 0.75 respectively.

Second, pre-post-surveys were conducted at the beginning and the end of the pedagogy. It investigated students' perception of the pedagogy for developing knowledge of electric circuits and

problem-solving skills. The questionnaire contained seven 5-point Likert scale questions on problem-solving. The mean rating for each question and the corresponding standard deviation were then calculated. The Cronbach's alpha reliability coefficients for the pre-survey and post-survey are 0.93 and 0.95 respectively.

Third, focus group interviews were conducted at the end of the pedagogy. It investigated students' perception of the pedagogy. There were 27 students randomly selected from the nine participating schools, with each focus group consisting of two to four students. The student respondents were asked to discuss the support from the pedagogy for developing knowledge of electric circuits and problem-solving skills, the level of enjoyment in and satisfaction with the pedagogy in STEM classrooms, and the challenges in and suggestions on the pedagogy in STEM classrooms. All the interview content was transcribed and systematically summarized.

3. Results and Discussion

3.1 Students' Achievement in Developing Knowledge of Electric Circuits under the Pedagogy

From the pre-post-tests results as shown in Table 2, the pedagogy effectively supported students to attain a noticeable gain in the knowledge specific for the topic of "Electric Circuits" in Grade 5 General Studies curriculum. The students had a statistically significant increase in the post-test scores for the question items on all three types of topic-specific knowledge points. The students were found to greatly enhance their knowledge about electronic components, electronic symbols, and closed circuit; and their capability of designing an electronic alarm with the application of knowledge about electric circuits.

Table 2. *Students' Achievement Before and After the Pedagogy on "Electric Circuits" (N =369)*

Question items			**Pre-test scores**	**Post-test scores**	**t-test**
Areas	**No. of items**	**Max. scores**	**Mean (SD)**	**Mean (SD)**	
(1) Electronic Components, Electronic Symbols, and Closed Circuit	2	5	3.13 (1.48)	4.08 (1.28)	11.65***
(2) Conductors, Insulators, Resistance, Series Circuit and Parallel Circuit	5	9	5.36 (1.85)	5.66 (1.72)	2.78*
(3) Design of an Electronic Alarm with the Application of Knowledge about Electric Circuits	1	4	0.36 (0.70)	0.59 (0.84)	5.21***
Total	**8**	***18***	***8.86 (3.15)***	***10.33 (2.92)***	***8.68******

$*p < 0.05$ $***p < 0.001$

Figure 3 gives an example that a student made great improvements in completing a test question before and after the pedagogy: (i) from no recognition (in pre-test) to a correct matching (in post-test) of electronic symbols with three target electronic components; (ii) from the correct identification of only one target conductor (in pre-test) to all three conductors and three insulators (in post-test); and (iii) keeping the correct understanding (from pre-test to post-test) of how the change in resistance influences the flow of electric current and the brightness of bulb-lighting in an electric circuit

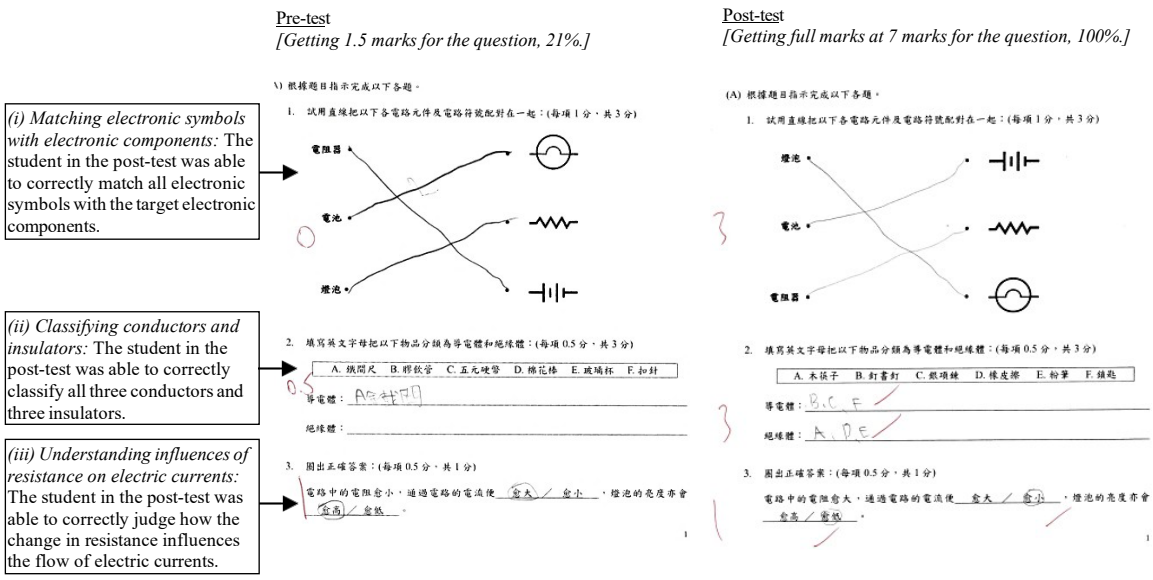


Figure 3. An example of the improvements made by a student in completing a test question before and after the pedagogy on "Electric Circuits".

3.2 Students' Perception of the Pedagogy for Developing Knowledge of Electric Circuits and Problem-Solving Skills

From the questionnaire survey results as shown in Table 3, the students positively perceived the pedagogy for developing knowledge of electric circuits and problem-solving skills. The students demonstrated a positive perception persistently that the pedagogy promoted them to repeat attempts for solving problems successfully. It is noteworthy that the students had a statistically significant increase in the level of confidence in their capability of all four elements of problem-solving skills. This shows that the pedagogy, from students' perspectives, can help with the development of problem-solving skills. Such finding serves a good indicator that the pedagogy is a pleasing and promising way for students to develop problem-solving skills.

Table 3. *Students' Questionnaire Survey Results on the Pedagogy on "Electric Circuits" (N =369)*

In the process of learning STEM …	***Pre-survey***		***Post-survey***		**t-test**
	Mean (1-5) [#]	***(SD)***	***Mean (1-5)*** [#]	***(SD)***	
I can design and create things in the correct order.	3.52	(1.11)	3.78	(1.15)	3.79***
I can solve a problem with repeated attempts until success.	3.62	(1.13)	3.77	(1.11)	2.24*
I can form a system through the interaction between devices.	3.50	(1.15)	3.74	(1.16)	3.40***
I can set different conditions (If… then…) to get desirable results.	3.44	(1.13)	3.70	(1.11)	3.87***
My overall problem-solving skills are improved.	3.53	(1.19)	3.70	(1.23)	2.32*
I can apply the problem-solving skills learned from STEM to other academic subjects.	3.42	(1.14)	3.68	(1.18)	2.67**
I can explain what cause the change and I can expect the result.	3.40	(1.12)	3.66	(1.14)	3.78***

[#]*Note: 1 = "strongly disagree"; 2 = "disagree"; 3 = "neutral"; 4 = "agree"; 5 = "strongly agree".*
**p < 0.05 **p < 0.01 ***p < 0.001*

From the focus group interviews results as shown in Table 4, students' positive perception of the pedagogy in STEM classrooms were further confirmed.

Table 4. *Students' Focus Group Interview Feedback on the Pedagogy on "Electric Circuits" (N = 27)*

Major interview feedback
Help in the development of knowledge of electric circuits and problem-solving skills
• Many students were able to differentiate the various types of electronic components and electric circuits, to draw circuit diagrams, and to connect electric circuits for building the target STEM artifacts. • Many students realized the necessity to think carefully about the casual relationship between the use of different types of electronic components, to correctly connect the given electronic components for a workable electric circuit.
Enjoyment in and satisfaction with the pedagogy
• Many students liked and enjoyed the STEM activities with coding tasks very much because they can do the hands-on practice of constructing various electric circuits. • Many students felt the experience of building STEM artifacts interesting and useful; and were highly satisfied with the effectiveness of the approach of learning by doing in the STEM activities.
Challenges in and recommendations for the pedagogy in STEM classrooms
• Some students were not familiar with coding with Micro:bit at the beginning and found challenging to complete the coding tasks in the STEM activities within the limited class time. • Some students recommended to extend the teaching scope of the pedagogy to introduce more different types of sensors for building STEM artifacts to solve problems in daily life.

The students asserted the help in the development of knowledge of electric circuits and problem-solving skills. Almost all of the student respondents considered that the pedagogy helped them to develop knowledge of electric circuits and problem-solving skills. For the development of topic-specific knowledge, the student respondents reflected that after the pedagogy they were able to recognize electronic symbols for comprehending and drawing circuit diagrams, and to identify different types of electronic components and connection methods for constructing varying closed circuits. Four student respondents further indicated that apart from learning about the rationale behind the operation of traffic lights, they were also stimulated by the pedagogy to explore the internal structures of common electrical devices in daily life. For the development of problem-solving skills, the student respondents indicated that the pedagogy helped them to go through the problem-solving process. Nearly one-third of the student respondents noted that they understood more about each of the four elements of problem-solving skills, especially the ones of sequencing and causal reasoning. Five student respondents noted that after the pedagogy they had more confidence in tackling challenging problems successfully through iterative trials and errors. One of these student respondents further illustrated that the impact of the pedagogy on stimulating interest and effort of thinking about the internal structure and operation model of various electrical devices in daily life. Another student respondent pointed out that the logical thinking built through the pedagogy can be extended to the learning process in other subjects, such as mathematical calculation in the mathematics subject.

The students expressed their high level of enjoyment in and satisfaction with the designed pedagogy for learning the topic of "Electric Circuits" in STEM classrooms. Almost all of the student respondents appreciated that the pedagogy brought them a happy and interesting experience in learning the target topic. The student respondents indicated that they liked the hands-on activities and became more engaging and attentive in General Studies lessons, comparing with subject lessons taught through the traditional teaching mode. Five of these student respondents further expressed their great sense of achievement after the successful completion of coding tasks and the correct connection of operable electric circuits in hands-on activities. More than four-fifths of the student respondents valued that the hands-on activities in the pedagogy exposed them to the real-life experience in handling electric circuits. Nine of these student respondents further indicated that such learning experience not only enabled them to deepen their understanding of the topic-specific knowledge, but also stimulated them to extend learning and thinking after class about the topic-related knowledge outside traditional textbook coverage for solving problems in daily life.

The students indicated challenges in and made suggestions on the pedagogy for General Studies lessons in STEM classrooms. Around one-fifth of the student respondents expressed their challenges when coding with Micro:bit at the beginning; and therefore had difficulties in completing the coding

tasks in the STEM activities within the limited class time. More than half of the student respondents recommended the pedagogy to include more hands-on tasks on designing other STEM artifacts, such as automatic cars, electric switches, mini fans, mini air-conditioners, fire alarm, intruder alarms, etc. Three of these student respondents further suggested that more types of sensors for building STEM artifacts can be introduced in the hands-on tasks, to expand the learning scope among students in this subject topic. Three student respondents further expected for introducing hands-on tasks with different levels of difficulty to cater for the learning diversity among students. One student respondent suggested on extending the use of pedagogy in Grade 5 General Studies lessons themed in other subject topics such as reflections and refractions of light rays.

4. Conclusion and Future Work

This study innovated a pedagogy in STEM classrooms for developing knowledge of electric circuits and problem-solving skills among Grade 5 students at Hong Kong. The pedagogy "To Play and Learn, To Inquire, To STEM and Code", supplementary with seven hands-on activity worksheets and 19 coding product templates and samples, was implemented in an eight-lesson trial teaching for 23 Grade 5 classes from nine Hong Kong primary schools, involving a total of 369 students. From the pre-post-tests, the pedagogy can effectively support the students to develop knowledge of electric circuits. From the questionnaire survey and focus group interviews, the students positively perceived their learning experiences and outcomes gained from the pedagogy, and significantly enhanced their awareness and confidence in applying problem-solving skills.

The pedagogy in this study was recommended to be promising for expansion with the concerns on the diversity of hands-on activities and the introduction to other subject topics. Future work will explore the potential suggestions collected from this study to further enhance the design and widen the implementation of the pedagogy, to expose students to a more appealing and helpful process in learning senior primary General Studies curriculum. The research design of the future work will try to arrange a control group for the trial teaching, in view of the limitation in this study that there was no control group involved in the evaluation of the effectiveness of the pedagogy.

Acknowledgements

The authors would like to acknowledge the support of the project "Use Coding as a Pedagogy for Teaching Subject Knowledge in Mathematics, General Studies and English Language and Fostering Computational Thinking" funded by the HKSAR Quality Education Fund.

References

Blackley, S., & Howell, J. (2019). The next chapter in the STEM education narrative: Using robotics to support programming and coding. *Australian Journal of Teacher Education*, *44*(4), 51-64.

Cederqvist, A. M. (2022). Designing and coding with BBC micro:bit to solve a real-world task – A challenging movement between contexts. *Education and Information Technologies*, *27*(5), 5917-5951.

Curriculum Development Council. (2017). *General Studies Curriculum Guide for Primary Schools (Primary 1 – Primary 6)*. Hong Kong: Curriculum Development Council.

Fokides, E., & Papoutsi, A. (2019). Using Makey-Makey for teaching electricity to primary school students. A pilot study. *Education and Information Technologies*, *25*(2), 1193-1215.

García-Carrillo, C., Greca, I. M., & Fernández-Hawrylak, M. (2021). Teacher perspectives on teaching the STEM approach to educational coding and robotics in primary education. *Education Sciences*, *11*(2), 1-16.

Lee, I., Grover, S., Martin, F., Pillai, S., & Malyn-Smith, J. (2020). Computational thinking from a disciplinary perspective: Integrating computational thinking in K-12 science, technology, engineering, and mathematics education. *Journal of Science Education and Technology*, *29*, 1-8.

Li, Y., Schoenfeld, A. H., diSessa, A. A., Graesser, A. C., Benson, L. C., English, L. D., & Duschl, R. A. (2020). On computational thinking and STEM education. *Journal for STEM Education Research*, *3*(2), 147-166.

Morado, M. F., Melo, A. E., & Jarman, A. (2021). Learning by making: A framework to revisit practices in a constructionist learning environment. *British Journal of Educational Technology*, *52*(3), 1093-1115.

Preston, C., Hubber, P., Bondurant-Scott, M., & Gunesekere, I. (2020). A representation construction approach to learning about electrical energy in year 6. *Teaching Science*, *66*(2), 5-19.

Priemer, B., Eilerts, K., Filler, A., Pinkwart, N., Rösken-Winter, B., Tiemann, R., & Zu Belzen, A. U. (2019). A framework to foster problem-solving in STEM and computing education. *Research in Science and Technological Education*, *38*(1), 105-130.

Trilles, S., & Granell, C. (2020). Advancing preuniversity students' computational thinking skills through an educational project based on tangible elements and virtual block-based programming. *Computer Applications in Engineering Education*, *28*(6), 1490-1502.

Iyer et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Computational Thinking Development: Validating an Instrument for Self-regulated Learning Using Animation

Siu-Cheung KONG[a*], Ye HU[b] & Ming LAI[c]
[a]*Department of Mathematics and Information Technology, The Education University of Hong Kong, Hong Kong*
[b]*Centre for Learning, Teaching and Technology, The Education University of Hong Kong, Hong Kong*
[c]*Centre for Learning, Teaching and Technology, The Education University of Hong Kong, Hong Kong*
*sckong@eduhk.hk

Abstract: Animation is a useful tool for supporting students' self-regulated learning (SRL). We present the validation of an instrument for assessing students' SRL using animations during their computational thinking (CT) development. The participants were 442 junior secondary students who took part in an initiative using block-based programming for CT development. Animations were created for each curriculum unit, and the students could use these animations to monitor their own learning. Following the SRL literature, we identified four theoretical SRL sub-constructs: goal setting, performance monitoring, time management, and self-reflection. The items were adapted from existing instruments for use in assessing students' SRL using animations. Confirmatory factor analysis showed that the four-factor model fitted the data well. This result indicated a better fit than a three-factor model based on a three-phase theoretical model for SRL. The four-factor model also indicated invariance across genders. The results provide evidence for the validity of the instrument; therefore, it can be used to assess students' improvements based on their SRL using animations.

Keywords: Animation, computational thinking, self-regulated learning

1. Introduction

Computational thinking (CT) is an essential skill set for the next generation (Wing, 2006). While several block-based programming environments have been developed to support CT development activities, students' CT development should not be taken for granted (Zhang & Nouri, 2019). Self-regulated learning (SRL) emphasises students' role in monitoring their own learning (Zimmerman, 2000), which follows the importance of student empowerment in CT education (Kong, Chiu, & Lai, 2018). Animation is a useful tool for supporting students' SRL (Iyer, Sharma, Sahasrabudhe, Garg, & Lokhande, 2021) and addressing learner diversity (Barak & Dori, 2011). However, the relationships among animation, SRL, and CT are less explored. In this study, we develop and validate an instrument for assessing students' SRL using animations during their CT development. Following the development of this instrument, we can explore students' learning using animations during their CT development and evaluate whether animations can be a possible way to address the issue of learner diversity, which is identified as a problem in CT education (Angeli & Giannakos, 2020). We address the following research questions: What factors should be included in the instrument for students' SRL using animations? Can the factor models be confirmed using the students' data?

2. Literature review

2.1 Self-regulated Learning

Garcia, Falkner and Vivian (2018) described SRL as students' active control of their learning process through managing their thoughts, behaviours and emotions to navigate their learning experiences successfully and achieve their learning goals. In addition, SRL has been recognised as a critical predictor of students' academic motivation and achievement. Scholars have shown that young learners

are engaged by SRL activities (Alvi & Gillies, 2021). Zimmerman (2000) developed a three-phase cyclical model using forethought, performance, and self-reflection phases. The forethought phase concerns preparing for and increasing learning attempts, the performance phase concerns self-control and the self-monitoring of one's performance during the learning process, and the self-reflection phase concerns performance evaluation, which helps students generate their self-reactions on how to improve their performance of the task in the future.

2.2 Computational Thinking and Self-regulated Learning

Peters-Burton, Cleary and Kitsantas (2018) highlighted the relationship between SRL and CT as involving a goal-directed process whereby the learner is required to identify a problem, examine relevant data to inform a solution and develop and evaluate their solution. According to Wing (2006), CT is a universal attitude and ability that anyone can acquire, not just computer scientists. Students are exposed to CT during programming activities, which involves resolving problems through the application of computer science concepts and practices, such as abstraction, debugging, remixing and iteration (Wing, 2006). Programming activities can efficiently assist kindergarten to year 12 (K-12) students in their CT development (Fessakis, Gouli, & Mavroudi, 2013). However, students do not learn programming skills easily (Angeli & Giannakos, 2020). Compared with traditional programming language learning, block-based programming languages, such as Scratch and App Inventor, reduce unnecessary syntax; therefore, they are frequently used in K-12 contexts (Lye & Koh, 2014).

2.3 SRL using animations for CT Development

Animation-based learning interactively combines text, images, sound, and video to assist student learning (Sastradika, Iskandar, Syefrinando, & Shulman, 2021). Animations help students to understand their learning materials, for enhancing their learning outcomes (Dori, Barak, & Adir, 2003) and motivation to learn (Rosen, 2009). Educators extensively use animations to provide their students with immediate, engaging feedback (Morales Díaz & Gaytán-Lugo, 2016). Iyer et al. (2021) found that interactive animations can play a role in students' learning by helping them to observe their learning process and reflect on their learning outcomes via feedback, which fosters their SRL (Zimmerman, 2000). Moreover, Animations provide students with enjoyable and interactive learning experiences, making it an effective tool for teaching CT (Morales Díaz & Gaytán-Lugo, 2016). Therefore, SRL using animations is a potential scaffold in assisting students' CT development.

2.4 Measurement of SRL

Several survey instruments have been developed to measure SRL (e.g., Barnard, Lan, To, Paton, & Lai, 2009; Schraw & Dennison, 1994) and can be mapped onto the three-phase theoretical model for SRL by Zimmerman (2000). Important sub-constructs in each of these three phases were also identified (Peters-Burton et al., 2018; Schunk, 1990; Wolters & Brady, 2021). In the forethought phase, 'goal setting' is one of the first steps for students to engage in learning activities since it provides standards for assessing learning performance. (Schunk, 1990). In the performance phase, 'time management' is crucial to academic success (Wolters & Brady, 2021) because it helps students use time effectively to achieve learning goals (Claessens, van Eerde, Rutte, & Roe, 2007). In addition, it is important for students to monitor their performance using strategies and resources (Barnard et al., 2009). Therefore, 'performance monitoring' is also important during this phase. In the self-reflection phase, students evaluate and react to their performance to judge how well they performed and why. The students can adjust their learning strategies afterward (Peters-Burton et al., 2018). Despite all of these being important SRL sub-constructs in the literature, they have seldom been included in a single instrument for assessing students' SRL (Jansen, van Leeuwen, Janssen, Kester, & Kalz, 2017). To the best of our knowledge, no measurement assesses the role of animations in students' SRL. In studying the importance of SRL and how the use of animations facilitates students' SRL for CT development, an instrument is needed to measure students' SRL using animations.

3. Method

3.1 Research Context

The background of this study is a pilot scheme for promoting students' SRL about programming using animations in their CT development in junior secondary high school. Four local secondary schools were recruited. The students learned a curriculum across eight teaching units using App Inventor to create mobile apps in a block-based programming language. After completing the eight units, the students were required to create an open-ended final project. To ensure that the teachers could guide their students to learn programming using animations in CT development, a 1.5-hour workshop was organised for the teachers to help them understand the basics of CT, how to use App Inventor for CT development, the curriculum structure, and how to guide their students' engagement in SRL using animations and worksheets. All of the learning and teaching materials, including the unit outlines, worksheets, and animations for each teaching unit, were presented to the teachers at the workshop.

3.2 Design of the Animation

The design of the animation was according to the steps outlined by the seven-step model of CT teaching (Kong & Lai, 2021) based on the technological pedagogical content knowledge (TPACK) framework (Mishra & Koehler, 2006). It began with the introduction of the new technological components such as list in App Inventor, followed by the CT concepts and practices involved, such as data structures and testing and debugging respectively. Then the animation guided the students to think how to approach the programming task, such as decomposing it into sub-tasks. Then it guided the students to code in the programming environment of App Inventor for finishing the app. Afterwards, the students were guided to think about the possible use of the technological components in other occasions and reflect on the concepts and practices that they had learned.

3.3 Participants and Procedures

The participants were 442 S2 (year two of junior high school) and S3 (year three of junior high school) students (aged 13 and 14 years) from four local schools (S2: N = 338, 76.6%; S3: N = 103, 23.4%), with 39.8% male (N = 176) and 60.2% female (N = 266) students. Before they started learning the programming curriculum for CT development using animations, the students were asked to complete an online SRL survey for programming using animations, which took them around 10 minutes.

3.4 Development of the SRL Instrument

In developing our SRL using animations instrument based on Zimmerman's four sub-constructs, namely goal setting, performance monitoring, time management and self-reflection, we searched for existing instruments and identified two instruments: the Online Self-regulated Learning Questionnaire (OSLQ) (Barnard et al., 2009) and the Metacognitive Awareness Inventory (Schraw & Dennison, 1994). We adapted the items from these two instruments to make them suitable for assessing SRL using animations. For example, for goal setting, we adapted the item '*I set short-term (daily or weekly) goals as well as long-term goals (monthly or for the semester)*' (Barnard et al., 2009) by modifying it into '*I know how to set up sub-goals to achieve the goal of watching the animation*'. Sub-goals could include problem formulation and problem-solving in finishing the final project. The resulting instrument has 12 items (i.e., four goal-setting items, three performance-monitoring items, three time-management items and two self-reflection items), which are each answered using a 5-point Likert scale (see Table 1). To ensure face validity of the instrument, all items were reviewed and revised by an expert with computer science and educational research backgrounds, followed by another expert with computer science background and teaching experience in secondary schools to further ensure that the items could be understood by secondary students.

3.5 Data Analysis

We first checked whether the assumption of normal data distribution could be verified. The skewness and kurtosis values were all within an absolute value of 1, suggesting that there was no violation of the assumption (see Table 1). We then validated our instrument by checking whether the collected data were compatible with the theoretically constructed factor model through a confirmatory factor analysis (CFA). To check the model's goodness of fit, we used the root mean square error of approximation (RMSEA) with a value smaller than .08 indicating a reasonable error of approximation, in addition to the Tucker–Lewis index (TLI) and comparative fit index (CFI), both with a value larger than .90 indicating an acceptable fit and .95 indicating a good fit. We compared the four-factor model described above with the three-factor model based on the three-phase model for SRL (Zimmerman, 2000).

4. Results

4.1 Confirmatory Factor Analysis

The goodness-of-fit indices showed that the four-factor model provided an acceptable fit to the data, (Chi square / degree of freedom (χ^2/df) = 3.121, $p < .001$, RMSEA = .069, CFI = .975, TLI = .965). The standardised factor loading estimates from the factors to the observed variables were all higher than the benchmark value of .50 (Raykov & Marcoulides, 2008). The 12 items and their corresponding factors are presented in Table 1. To check whether the two 'performance phase' factors of time management and performance monitoring could be combined as suggested by the three-phase model for SRL (Zimmerman, 2000), we fitted the three-factor model to the data. We kept goal setting as Factor 1 and self-reflection as Factor 3, whereas the items belonging to time management and performance monitoring formed Factor 2. The CFA results indicated the inadequate fit of the three-factor model to the data (χ^2/df = 5.584, $p < .001$, RMSEA = .102, CFI = .942, TLI = .966). In the comparison of the two models, a significant χ^2 reduction ($\Delta\chi^2 = 134.967$, Δdf = 3, $p < .01$) was found from the three-factor to the four-factor model, indicating that the latter had a better fit to the data.

Table 1. *Loadings of the Items based on the Four-factor Model and their Descriptive Statistics*

Sub-constructs	*Factor loading*	*Mean*	*SD*	*Skewness*	*Kurtosis*
Goal setting					
1. I know how to set up sub-goals to achieve the goal of watching the animations.	0.81	3.42	1.01	−.258	−.196
2. I receive guidance from my teachers on how to achieve the goals in stages.	0.81	3.45	.99	−.442	.156
3. I have autonomy to achieve the goals in my own way.	0.83	3.49	1.05	−.412	−.094
4. I am clear about the main goal for watching the animations.	0.72	3.63	.96	−.295	−.188
Performance monitoring					
5. I shall use the animations and student guide when I have difficulties during the coding process.	0.83	3.43	1.11	−.330	−.371
6. I watch the animations because they help me to understand coding and computational thinking.	0.86	3.41	1.07	−.287	−.247
7. If I cannot understand the concepts and coding practices in the animation fully, I will watch it again.	0.81	3.49	1.10	−.356	−.392
Time management					
8. I can arrange an appropriate time and location to watch the animation.	0.77	3.44	1.05	−.315	−.214

9. I can schedule enough time to watch the animations.	0.78	3.60	1.03	−.416	−.203
10. I make sure I keep up with the overall progress in the course by watching the animations.	0.85	3.59	1.06	−.446	−.214
Self-reflection					
11. I like the part of the animations that lead my reflection on the goal of coding and connection to daily life.	0.84	3.43	1.03	−.269	−.136
12. In my learning process for coding, I continuously reflect on what I have learned in completing my final programming project.	0.85	3.44	1.04	−.274	−.200

Note: SD stands for Standard Deviation

To evaluate the measurement invariance across genders, we conducted a multi-group analysis (male and female students) based on the four-factor model. The results are presented in Table 2, with the fitting of a configural invariance model (baseline, Model 1), a metric invariance model (Model 2) and a scalar invariance model (Model 3), which all indicated an acceptable fit. The comparison between Models 1 and 2 indicated metric invariance because there was no significant change in χ^2 ($\Delta\chi^2 = 3.520$, $\Delta df = 8$, $p = .898$). The comparison between Models 1 and 3 also indicated scalar invariance ($\Delta \chi^2 = 9.259$, $\Delta df = 20$, $p = .980$). Therefore, the four-factor model indicated invariance across genders.

Table 2. Measurement Invariance across Genders

	χ^2	df	*p*-value	RMSEA	CFI	TLI
Configural invariance (Model 1)	283.527	96	<.001	.067	.954	.937
Metric invariance (Model 2)	287.047	104	<.001	.063	.955	.943
Scalar invariance (Model 3)	292.786	116	<.001	.059	.957	.951

5. Conclusion and Implication

Empowerment is an important aspect in students' development of CT (Kong et al., 2018). The emphasis on SRL in education facilitates students' control of their learning process, which subsequently empowers them in their learning. Animations have been found to be useful in supporting students' SRL (Iyer et al., 2021). As such, they can be used as a teaching resource in students' CT development through programming activities (Kong & Lai, 2021). In this study, we developed and validated an instrument for assessing students' SRL using animations for CT development. Using this instrument, we can assess students' learning gains when they use animations for CT development. Based on the SRL literature, we identified goal setting, performance-monitoring, time-management and self-reflection sub-constructs, which can be mapped onto the three-phase theoretical SRL model proposed by Zimmerman (2000), with goal setting belonging to the forethought phase, performance monitoring and time management belonging to the performance phase and self-reflection belonging to the self-reflection phase. This four-factor model was confirmed by the CFA results, which indicated a better fit compared to the three-factor model with the middle two sub-constructs combined as one sub-construct. The invariance analysis also suggested that the four-factor model can be verified among both male and female students. The establishment of this instrument allows researchers to further investigate changes in students' SRL using animations following interventions. Animation-based learning enhances students' engagement and knowledge transfer (Rosen, 2009). Teachers can use our instrument to explore whether the involvement of animations in teaching can help students to achieve a higher level of SRL. To this end, it is important for schools and policymakers to provide financial support for the creation of learning-oriented animations, as well as teacher development programs to facilitate animation-based learning in schools. A limitation of this study is that since the project is ongoing, we have not yet collected sufficient post-test data for CFA. The data analysed were based on pre-test findings, before the students started learning CT using animations. We will continue to examine students' improvement in SRL after using the animations to support their CT development.

References

Alvi, E., & Gillies, R. M. (2021). Promoting self-regulated learning through experiential learning in the early years of school: A qualitative case study. *European Journal of Teacher Education*, *44*(2), 135–157.

Angeli, C., & Giannakos, M. (2020). Computational thinking education: Issues and challenges. *Computers in Human Behavior*, *105*, 106185.

Barak, M., & Dori, Y. J. (2011). Science education in primary schools: Is an animation worth a thousand pictures? *Journal of Science Education and Technology*, *20*(5), 608–620.

Barnard, L., Lan, W. Y., To, Y. M., Paton, V. O., & Lai, S.-L. (2009). Measuring self-regulation in online and blended learning environments. *The Internet and Higher Education*, *12*(1), 1–6.

Claessens, B. J. C., van Eerde, W., Rutte, C. G., & Roe, R. A. (2007). A review of the time management literature. *Personnel Review*, *36*(2), 255–276.

Dori, Y. J., Barak, M., & Adir, N. (2003). A web-based Chemistry course as a means to foster freshmen learning. *Journal of Chemical Education*, *80*(9), 1084.

Fessakis, G., Gouli, E., & Mavroudi, E. (2013). Problem solving by 5–6 years old kindergarten children in a computer programming environment: A case study. *Computers & Education*, *63*, 87–97.

Garcia, R., Falkner, K., & Vivian, R. (2018). Systematic literature review: Self-regulated learning strategies using e-learning tools for Computer Science. *Computers & Education*, *123*, 150–163.

Iyer, M., Sharma, R., Sahasrabudhe, S., Garg, A., & Lokhande, G. (2021). Project OSCAR: Open source animations repository to foster self-regulated learning. *Bulletin of the Technical Committee on Learning Technology*, *21*(2), 4-10.

Jansen, R. S., van Leeuwen, A., Janssen, J., Kester, L., & Kalz, M. (2017). Validation of the self-regulated online learning questionnaire. *Journal of Computing in Higher Education*, *29*(1), 6–27.

Kong, S.-C., Chiu, M. M., & Lai, M. (2018). A study of primary school students' interest, collaboration attitude, and programming empowerment in computational thinking education. *Computers & Education*, *127*, 178–189.

Kong, S.-C., & Lai, M. (2021). A proposed computational thinking teacher development framework for K-12 guided by the TPACK model. *Journal of Computers in Education*, 9, 379-402.

Lye, S. Y., & Koh, J. H. L. (2014). Review on teaching and learning of computational thinking through programming: What is next for K-12? *Computers in Human Behavior*, *41*, 51–61.

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, *108*(6), 1017–1054.

Morales Díaz, L., & Gaytán-Lugo, L. S. (2016). Computer animation as a vehicle for teaching computational thinking. In F. J. Mata & A. Pont (Eds.), *ICT for promoting human development and protecting the environment* (pp. 53–59). Cham: Springer International Publishing.

Peters-Burton, E. E., Cleary, T. J., & Kitsantas, A. (2018). Computational thinking in the context of science and engineering practices: A self-regulated learning approach. In D. Sampson, D. Ifenthaler, J. M. Spector, & P. Isaías (Eds.), *Digital technologies: Sustainable innovations for improving teaching and learning* (pp. 223–240). Cham: Springer International Publishing.

Raykov, T., & Marcoulides, G. A. (2008). *An introduction to applied multivariate analysis* (1st ed.). New York: Routledge.

Rosen, Y. (2009). The effects of an animation-based on-line learning environment on transfer of knowledge and on motivation for science and technology learning. *Journal of Educational Computing Research*, *40*(4), 451–467.

Sastradika, D., Iskandar, I., Syefrinando, B., & Shulman, F. (2021). Development of animation-based learning media to increase student's motivation in learning physics. *Journal of Physics: Conference Series*, *1869*(1), 012180.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460–475.

Schunk, D. H. (1990). Goal setting and self-efficacy during self-regulated learning. *Educational Psychologist*, *25*(1), 71–86.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33–35.

Wolters, C. A., & Brady, A. C. (2021). College Students' Time Management: A self-regulated learning perspective. *Educational Psychology Review*, *33*(4), 1319–1351.

Zhang, L., & Nouri, J. (2019). A systematic review of learning computational thinking through Scratch in K-9. *Computers & Education*, *141*, 103607.

Zimmerman, B. J. (2000). Chapter 2—Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). San Diego: Academic Press.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Rural Teacher ICT Literacy Professional Learning and Teaching Practice in China: A Multiple Case Study Investigation

Yanfen HUANG [a,b], Sharifah Intan Sharina SYED-ABDULLAH[a*], Nurul Nadwa ZULKIFLI[c] & Norliza GHAZALI[a]

[a]*Faculty of Educational Studies, Universiti Putra Malaysia, Serdang, Malaysia*
[b]*Faculty of Teacher Education, Zhaoqing University, Guangdong, China*
[c]*Faculty of Humanities, Management and Science, Universiti Putra Malaysia, Bintulu, Malaysia*
sharifahintansharina@upm.edu.my

Abstract: China is currently vigorously developing educational information, and implementing nationwide projects to improve teachers' ICT literacy. However, the progress of such projects is significantly unbalanced between the rural and urban areas. This study uses a multi-case study approach to investigate the approaches and effectiveness of ICT professional learning programs received by seven rural teachers in the past decade. The study suggests that regardless of those programs, rural teachers' teaching practice and professional development, as well as the student's academic results remain limited due to the flawed local educational environment. The study examines the professional learning approaches applied in rural ICT professional learning programs and identifies both the effective ones and the barriers.

Keywords: Rural teacher, ICT literacy, teaching practices, teacher professional development

1. Introduction

The sustained development and popularization of information and communication technology (ICT) have great potential to accelerate social development, bridge the digital divide, and support the development of an inclusive knowledge society (OECD, 2015). In recent years, China has attached great importance of ICT to its educational system. The Ministry of Education of China proposed the comprehensive implementation of the National Primary and Secondary School Teachers' Information Technology Application Ability Promotion Project in 2013 and 2018 (respectively, Promotion Project 1.0 and 2.0). These projects aim to comprehensively improve the ICT-based teaching ability of primary and secondary school teachers and promote innovation and increase the use of ICT in teaching before 2022. In addition, Promotion Project 2.0 proposed the Development Framework of Information Education and Teaching Ability of Primary and Secondary School Teachers (DFIETAT) (MOEC, 2021).

ICT in education is a vital measure to bridge the gap between urban and rural education development (Wu & Lu, 2020). It is found that compared with urban teachers in China, the ability and creativity of rural teachers in applying ICT in their teaching are still at an unsatisfactory level (BNU, 2020). Therefore, understanding and developing the ICT literacy of rural teachers is an urgent need to be achieved. High-quality professional development plays a crucial role in improving the quality of teachers' classroom practice, which will influence the overall quality of students' learning (Eliahoo, 2016; Vangrieken et al., 2017). However, few studies investigate how professional learning affects teachers' subsequent teaching practice in rural areas in China.

The purpose of this study is to explore rural teachers' views on the potential benefits to themselves and their students from their participation in ICT literacy professional learning activities. This study aims to provide an in-depth overview of how rural teachers engage in ICT learning, and how their learning is disseminated to their students from professional development projects. It was hoped that this study will provide evidence for understanding which methods hold promise, identify barriers to success, and provide information for future implementation of rural teacher professional development.

2. Methodology

This study adopted the multi-case study method to investigate the phenomena of ICT learning and the practice of rural teachers. It aimed to explore mechanisms related to rural teachers' ICT literacy professional learning and teaching practice development through the viewpoints of the participants. The interview protocol in this study was designed according to the effective teacher professional development framework proposed by Desimone (2009). The framework consists of four dimensions: core features of professional development, increased teacher knowledge and skills, change in attitudes and beliefs, change in instruction, and improved student learning.

This study used semi-structured interviews as the main data collection technique. Purposive sampling technique is adopted in this study to identify respondents to participate in the cases. This study ended up with seven participants who volunteered to participate in the interviews. They all come from different cities and teach different subjects.

For data analysis, in the first step, this study adopted a relying on theoretical proposition analytical strategy (Yin, 2018) to conduct a deductive analysis of the data. This was then followed by the "Grounded up" analysis as proposed by Yin (2018). The inductive analysis method was used to re-analyze the data line by line. Label the data source that has not been encoded, and discover new categories. Third, analyze all themes and categories, looking for core types and associating categories to determine the relationship between each category and theme and draw the internal logical relationship between them.

3. Findings

Four themes and 13 categories were obtained by a cross-case analysis in this study. The conceptual model is shown in Figure 1. It also demonstrates the four synthesized findings of this study.

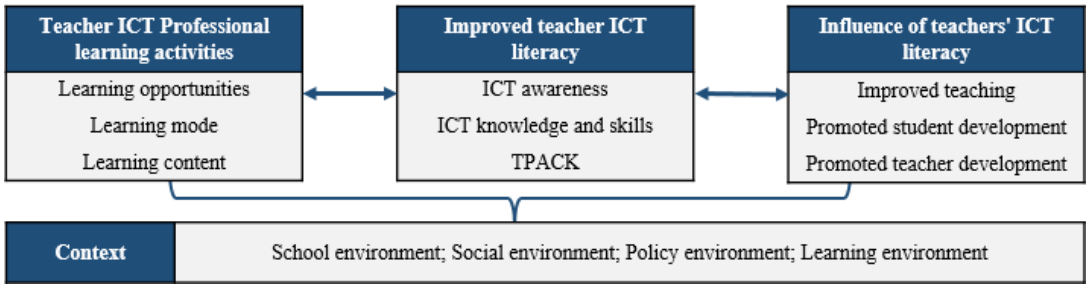


Figure 1. A conceptual Model of ICT Literacy Professional Learning and Teaching Practice for Rural Teachers

Finding 1: Rural teachers expect access to high-quality and tailored ICT professional learning programs.

Although Chinese education authorities emphasize that all primary and secondary school teachers must participate in ICT professional learning, rural teachers have fewer opportunities than their urban counterparts, especially for face-to-face learning. Online learning is less efficient than online ones in improving rural teachers' ICT competencies. They need high-quality blended training. In addition, not all rural teachers are interested in advanced technologies such as artificial intelligence, they prefer those that can be applied to daily teaching practice in their schools.

Finding 2: The improvement of rural teachers' TPACK requires continuous follow-up instruction. Constant follow-up instructions are needed to …

ICT training programs have been proved to have improved teachers' ICT awareness, knowledge and skills, but they cannot access the needed resources to successfully incorporate ICT into their classrooms. They need preparation time and continuous practical guidance from experts in the field of technology and primary and secondary school teachers with extensive practical experience.

Finding 3: Rural teachers are less confident and motivated than urban teachers in using ICT for instructional innovation.

The improvement of rural teachers' ICT literacy has contributed to changing classroom teaching, enriching students' knowledge and promoting students' development. Better ICT literacy

helps to transform/facilitate classroom teaching and achieve better teaching results. At the same time, it has contributed to their professional development, especially in multidisciplinary teaching skills. However, they mostly showed low confidence in using ICT to continuously assess and reflect on professional practice for innovation and improvement.

Finding 4: Multiple factors influence rural teachers' teaching practice and professional development

Rural teachers' ICT literacy professional learning and teaching practices are primarily influenced by their school environment, social environment, policy environment, and learning environment. The impact of school resources and leadership support is particularly evident. In addition, they especially need a close-knit community of teachers to interact and collaborate with each other.

4. Conclusion and Limitation

With the increasing influence of technology on social development, the Chinese government attaches more and more importance to the application of ICT in education. Even in economically underdeveloped provinces, the construction of school information facilities and ICT training of teachers have been guaranteed to a certain extent. This study believes that the blended professional learning model may provide more learning opportunities for rural teachers and improve their sense of participation and confidence in practice. However, how to better integrate online and offline learning needs further discussion. In addition, rural teachers face a more complicated teaching environment than urban teachers. Therefore, building a blended learning model and professional development community for rural teachers has become an issue that researchers need to study further. The learning model and community should combine online and offline and on-campus and off-campus. It is also an important means to promote the professional development of rural teachers and the equity of urban and rural education.

The main limitation of this study was the sample size. However, these participants teach almost all subjects in Chinese primary schools, and multidisciplinary teaching is common among teachers in rural schools in China (An & Yin, 2017), so the results of this study have general implications for the study of rural teachers in China.

References

An, X. & Yin, L. (2017). An investigation on teacher professional development in small-scale rural schools. *Journal of Shanghai Educational Research*, (07),5-9.

BNU (Beijing Normal University). (2020, December 12). "2020 White Paper on Teacher Training in China" released. http://www.edu-gov.cn/edu/15392.html,2020-12-22.

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. *Educational Researcher,* 38(3), 181–199.

Eliahoo, R. (2016). Teacher educators: Proposing new professional development models within an English further education context. *Professional Development in Education*, 43(2), 1–15.

MOEC (Ministry of Education of the People's Republic of China). (2021, March 21). Opinions on implementing the national primary and secondary school teachers' information technology application ability promotion Project 2.0. http://www.moe.gov.cn/srcsite/A10/s7034/201904/t20190402_376493.html

OECD. (2015). *Transforming our world: The 2030 agenda for sustainable development.* United Nations: New York, NY, USA. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N15/291/88/PDF/N1529188.pdf

Vangrieken, K., Meredith, C., Packer, T., & Kyndt, E. (2017). Teacher communities as a context for professional development: A systematic review. *Teaching and teacher education*, 61, 47-59.

Wu, F., Liu, S. (2020). Obstacles and Breakthrough Paths of Information Eliminating the Digital Divide between Urban and Rural Education. *China Educational Technology*, (2): 30-36.

Yin, R. K. (2018). *Case study research: Design and method (6th ed.)*. Thousand Oaks, CA: Sage.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

How Teachers Used ICT in Education in 2015 and 2022 in Croatia: A Cross-Sectional Study

Ivica BOTICKI*, Marcela MATAS, Lana JURAK & Bruno SEGVIC
Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia
*ivica.boticki@fer.hr

Abstract: In this paper, the results of a cross-sectional study comparing two surveys on educational technology use are presented. The first survey included 1103 teachers and was completed in 2015 to explore the areas of in-class technology use, digital contents use, professional development, and other relevant issues in primary and secondary education in Croatia at the time. The second survey was administered seven years later and included 838 teachers in total. The two survey results were compared to identify changes in teachers' educational technology use and related issues. The results show large increase in technology and digital contents use in everyday primary and secondary school classrooms confirming these tools become more widespread and more accepted nowadays. On the other hand, the analysis shows pedagogical, and teacher professional development issues remain challenges even after 7 years, indicating those areas should be the focus on attention in terms of educational technology sustained use.

Keywords: ICT, teacher professional development, technology use, cross-sectional survey

1. Introduction

ICT (Information and Communication Technology) has become an integral part of education and is applied in a wide range of areas including lectures, communication with students and generally enables access to learning resources and is becoming a standard part of curriculum implementation (Vuk et al., 2020). ICT is often used to save time on highly redundant tasks so that teachers get to focus more on high quality work with the students. In the time of Covid-19 pandemic, classes without ICT would have been for the most part impossible.

A prior survey conducted by the authors in 2015 titled "The application of ICT in primary and secondary schools in Republic of Croatia" (in Croatian) analyzed the use of ICT with a focus on innovative use of technology in teaching and learning (Pović et al., 2015). The 2015 survey revealed that although 25% of teachers used technology in class, they were still keen on using even more technology in their classes but were not adequately educated for its use. In 2015, the lack of resources was often managed by the teachers themselves by buying equipment to be used in class, finding donors etc.

In order to map out the changes in the primary and secondary education between 2015 and 2022, the authors conducted a new survey which is presented in detail in this paper, in which teachers were asked about the current state of the use of ICT in education similarly as in the 2015 survey. The research was based on questions from the 2015 survey. Both surveys will be carefully compared side by side based on the teachers' answers. With the 2022 survey it was possible to draw the comparison between the change of technology use and the perception of technology use in education between 2015 and 2022 in Croatia. In addition to that, the teachers reported how Covid-19 affected their use of technology in class and how it has affected the quality of teaching. The outcome of this research contributes to better understanding the current situation of ICT use by the teachers and the possibilities for advanced implementation of ICT in Croatian schools (Ministry of Public Administration (Croatia), 2017; Ministry of Regional Development and EU Funds, 2018).

2. Teacher Technology Use and Adoption

Digital technologies present an important tool that is used for education and skill development. To take advantage of the full capacity of digital learning, a lot of prerequisites need to be met for both teachers

and students to gain access to digital equipment and fast internet connection. Furthermore, teachers need to be educated about the technology they will be using in their teaching in order to be competent to pass on the knowledge on adequate technology use to their students who will soon become teachers (Wei, 2020). The study carried out for the European Commission called "ICT in Education" discusses access and use of technology in education by surveying teachers and students from EU, Norway, Iceland, and Turkey to provide more data and evidence regarding digitalization in education (European Commission, 2019). The study explored teachers' and students' confidence in their own digital competence. One of their key findings is that teachers state there is insufficient number of digital tools such as laptops, computers, and tablets available. Moreover, although teachers are most confident in the areas of communication, safety, and basic ICT activities, they lack confidence in more complex tasks, such as coding.

Given the variety and the sheer number of resources, digital technologies can be transformative to the teaching process and improve decision-making in education. Covid-19 pandemic has further accelerated the digitalization process. There are different ways in which the digital technologies can be used for education systems such as: 1) objectives of learning for the students, 2) to support students in their learning, 3) to support teachers and 4) to support the management of schools and education systems (OECD, 2020). Given the constant technological progress individuals need a mix of skills rather than more specialized digital skills. For the teachers to teach they not only need to be skillful with the technology, but they also need to coordinate the use of technology, pedagogy, and subject matter. This seems to be *au pair* with the analysis done by PISA which warrants for a more in depth rethinking of the governance, funding, distribution, and management of ICT resources (OECD, 2019).

The survey conducted by Gudmundsdottir and Hatlevik (2020) has concluded that the teacher-students are poorly educated on the ways to find reliable content, on how to adhere to copyright rules and how to teach pupils safe and responsible ways of using ICT. Most of their knowledge regarding ICT and digital responsibility came from things they learned on their own, and not from the teacher development sessions (CARNet, 2017). Also, their experience regarding ICT during practice placement in schools is varied. Some were told to keep off technology as it is not used in the school, while the others were encouraged to use it and were even taught how to use it by teacher educators (Gudmundsdottir et al., 2020). The survey emphasizes that teacher-educators who taught teacher-students how to use technology focused more on the technical aspects and less on pedagogical or responsible ways of using ICT. In the first part of the survey, which was conducted before teacher-students' in-school sessions, teacher-students were able to check the reliability of online content by methodologically examining it. However, the methods for checking reliability of content are difficult to pass onto students, especially when it comes to knowledge on advanced search procedures (Avalos, 2011). Limited awareness of privacy issues was reported, while copyright rules were acknowledged but ignored due to convenience.

3. Methodology

The goal of the 2022 survey was to observe the evolution of teacher educational technology use in Croatia in the seven preceding years. To make the comparison of the two survey results possible, a decision was made to employ the original survey questions again in 2022 and to supplement them with new developments in the ICT use and COVID-19 specifics. As in the original survey, the teachers across elementary and high schools across Croatia were randomly chosen as the target survey population

54 questions general demographic data, electronic class register system called e-Diary, the use of computers, tablets and smartphones in lessons, use of digital educational content and smart boards, education on safe ways of using ICT, problems teachers encounter while using ICT in class and the effect of Covid 19 on ICT usage were devised. Survey was then sent to the ethical committee of the Faculty of Electrical Engineering and Computing University of Zagreb for the ethics approval of the study. The survey was then implemented using Google Forms and distributed to teachers using e-mail. Data collection occurred over one month - from 17th of January until 18th of February 2022. This specific time frame was chosen as this is typically the time of the year when teachers have less obligations and there are no school holidays, so they are more available.

838 teachers took part in the 2022 survey. Data analysis of the collected data was done in three parts: quantitative analysis, qualitative analysis, and the comparison with the 2015 survey results. Quantitative analysis was done in RStudio using the R programming language and in Excel. Qualitative

analysis on open-ended questions was done manually by grouping answers into categories. That way, spelling and grammatical errors were eliminated, and similar answers got grouped. Then, the frequency of answers in each group was calculated. The two survey results were compared in relative terms due to slight difference in the number of participants.

4. Results

a. General Information and Technology Use Frequency

In the 2015 survey there were 1103 participants, while the 2022 survey got completed by 836 teachers. As Table 1 shows, the gender structure of the participants remained comparable, while slight changes occurred in the demographic composition of the participants, where there is increase in more senior age group (>50) in 2022 in comparison to 2015. This can most likely be attributed to the increasingly aging population in Croatia and general emigration trends in recent years. At the same time the youngest age group (<30) shrank due to decreasing demand for teachers (since less students get enrolled).

Table 1. Gender and demographic structure of survey participants (2015 and 2022)

	2015	**2022**	**2015 (%)**	**2022 (%)**
Female	832	667	75	80
Male	271	169	25	20
< 30	150	88	14	10
30 - 40	411	262	36	32
40 - 50	327	249	30	30
> 50	215	237	20	28

The percentage of teachers who use computers, tablets and smartphones in education remained very high in both surveys (97% and 99% in 2015 and 2022, respectively), but the frequency of use increased. Table 2 shows that the frequency use of desktop computers and laptops increased across almost all-time options. The only decrease appeared for the "once a week" option which is compensated with a very notable increase in the "almost every lesson" option, indicating the teachers are adopting regular computer/laptop use in their classrooms.

Table 2. Frequency of technology in education use: "How often do you use computers/laptops in lessons?"

	2015	**2022**	**2015 (%)**	**2022 (%)**
Never	11	10	1	1
A couple of times per year	87	16	9	2
Once a month	79	42	8	5
Once a week	219	66	22	8
Almost every lesson	604	684	60	84

Table 3 shows that the general frequency of use of tablets increased across all options, with the first three options (a couple of times per year, once a month, once a week) showing the greatest increase.

Table 3. Frequency of technology in education use: "How often do you use tablets in lessons?"

	2015	**2022**	**2015 (%)**	**2022 (%)**
Never	766	434	77	53
A couple of times per year	65	119	6	15
Once a month	50	94	5	11
Once a week	47	101	5	12
Almost every lesson	72	70	7	9

The most substantial increase was recorded in smartphone frequency use (Table 4) with a drastic decrease in the number of teachers who never used smartphones in their lessons (from 66% to

23%) and very notable increase across all time options, with all the options surpassing the tablet use and some options even surpassing the desktop or laptop use.

Table 4. Frequency of technology in education use: "How often do you use smartphones in lessons?"

	2015	2022	2015 (%)	2022 (%)
Never	655	187	66	23
A couple of times per year	115	136	12	17
Once a month	74	164	7	20
Once a week	75	154	7	19
Almost every lesson	81	177	8	21

b. Digital Learning Contents Use

Digital learning contents examined in this paper come in a variety of forms, ranging from video tutorials to YouTube videos, web tools, smart games, and other digital educational contents. Table 5 shows that video educational contents were more frequently used across all time options in 2022 compared to 2015.

Table 5. Frequency of digital contents use – "How often do you use video tutorials from YouTube or similar sites?"

	2015	2022	2015 (%)	2022 (%)
Never	107	33	11	4
A couple of times per year	291	130	29	17
Once a month	255	244	26	29
Once a week	265	295	27	36
Almost every lesson	82	116	7	14

There is a very notable increase in Web 2.0 use across all time options in 2022 (Table 6). The proportion of teachers who never used such tools almost halved, while the number of teachers who use such tools at almost every lesson increased dramatically.

Table 6. Frequency of digital contents use – "How often do you use Web 2.0 tools?"

	2015	2022	2015 (%)	2022 (%)
Never	637	279	63.7	34.1
A couple of times per year	150	146	15.0	17.9
Once a month	87	151	8.7	18.5
Once a week	93	137	9.3	16.7
Almost every lesson	33	105	3.3	12.8

The use of smart games follows similar patters as the use of Web 2.0 tools, where teachers show more smart games adoption in general and more frequent use across most time options (Table 7).

Table 7. Frequency of digital contents use – "How often do you use smart games?"

	2015	2022	2015 (%)	2022 (%)
Never	514	173	51.4	21.2
A couple of times per year	249	186	24.9	22.7
Once a month	138	186	13.8	22.7
Once a week	79	190	7.9	23.3
Almost every lesson	20	83	2.0	10.1

Due to increased availability of native (Croatian) digital contents there is a notable increase of native contents use (Table 8).

Table 8. Digital contents use – "Do you use digital content made by Croatian publishers?"

	2015	2022	2015 (%)	2022 (%)
Yes	510	629	51.0	76.9
No	490	189	49.0	23.1

c. Access Control and Communication

Table 9 indicates only some teachers, and their schools, were able to provide access to computers/tablets to their students but this number rose significantly (from 17% to 56% of the surveyed teachers from 2015 to 2022, respectively). Nevertheless, access control mechanisms remain the same in 2015 and 2022, indicating there was no substantial development in this area over the years.

Table 9. Access control – "If students have access to computers/tablets, how do you control what they do?"

	2015	2022	2015 (%)	2022 (%)
Personally	122	267	64	57
Software	45	119	24	26
Combination of personally and software	20	48	11	10
Time limit	1	16	1	4
Do not control them	0	15	0	3

Communication between teachers and students reportedly improved with less teachers reporting they do not communicate with their students at all (Table 10).

Table 10. Communication – "How do you communicate with students outside of class?"

	2015	2022	2015 (%)	2022 (%)
I do not	286	6	31	2
E-mail	258	76	29	30
Social networks	124	43	14	16
Calls and/or text messages	91	63	10	24
Teams	84	69	9	26
Google Classroom	60	5	7	2

5. Discussion

The analysis presented in the preceding chapters reveals there exist very notable trends in ICT in education use that happened over the last 7 years in Croatia, since the initially administered 2015 survey. These trends happen in demographically and globally challenging times, where on one side the overall Croatian population is decreasing due to low birth rate and emigration, while on the other side the teacher population generally ages and is not supplemented with new junior teachers as fast as expected (Keržic et al., 2021).

Technology use remains high in classrooms in 2022, as expected. However, the frequency of technology use across all types of devices increased a lot. Standard classroom equipment such as desktop/laptop computers becomes widely available and is being predominantly used by the teachers during almost every lesson. Although frequency of use increased for tablet computers, the use of smartphones dramatically rose indicating such tools become more widely accepted in today's Croatian classrooms. Although such changes are most definitely fueled by high availability and lower prices of high-tech equipment and the Covid-19 pandemic drive for online learning, it must be noted that in 2022 compared to 2015 more devices used by the teachers come from the schools' and less are own teachers' devices. In terms of digital contents use, there is some increase in the use of video content at almost every lesson, and surprising changes are noted with a massive increase in the use and the frequency of use of Web 2.0 tools and smart games by the teachers. This change can be attributed to the structured and EU-supported approach to the development of a digital educational contents repository in Croatia

which was made available to all Croatian students and teachers. This is well confirmed by a large increase in the number of teachers using localized educational contents in 2022.

Although there is more than apparent use of educational tools and educational digital contents, the changes become less dramatic when teacher professional development or student resource access is concerned. Teacher professional development remains a challenge in 2022 with only a handful of initiatives available, most triggered due to the Covid-19 situation. Students' device access control is still being managed mainly on the personal basis, while more structured approaches to controlling students' device use in class remains on the 2015 levels. In terms of communication with the students out of class, teachers continue to use e-mail and social networks on the 2015 levels, and in 2022 used more calls/text messages and specialized software. This can be explained by the Covid-19 pandemic and the need to establish communication channels with the teachers in a time of need.

6. Conclusions

In this paper a survey on primary and secondary teacher use of ICT in education is presented. The survey was conducted in 2022 and compared to the survey conducted in 2015. This cross-sectional approach was employed to investigate the change in teacher educational technology use and compare the changes in the 7-year period. There is an increase in educational technology use and digital learning contents use in 2022 in comparison to 2015. What is more, the increase is even more emphasized in the technology and digital contents frequency of use, where teachers got empowered with the availability of educational tools and use them on almost daily basis in their classrooms. The challenges related to teacher professional development remain, where teachers need to manage their growing technological resources on their own and find ways of devising adequate pedagogies and manage different facets of daily technology and digital content use on their own. In this respect it is to be noted that mere technology and digital contents have their limits in truly transforming the educational arena and need to be truly embedded everyday in-class pedagogies to unleash their true potential.

References

Avalos, B. (2011). Teacher professional development in Teaching and Teacher Education over ten years. In *Teaching and Teacher Education* (Vol. 27, Issue 1). https://doi.org/10.1016/j.tate.2010.08.007

CARNet. (2017). *e-Škole: Mnoge hrvatske škole su digitalne početnice [e-Schools: May Croatian schools are digital beginners]*. https://www.e-skole.hr/hr/e-skole-mnoge-hrvatske-skole-su-digitalne-pocetnice

European Commission. (2019). 2nd Survey of Schools: ICT in Education. Technical Report. In *European Commission*.

Gudmundsdottir, G. B., Gassó, H. H., Rubio, J. C. C., & Hatlevik, O. E. (2020). Student teachers' responsible use of ICT: Examining two samples in Spain and Norway. *Computers and Education*, *152*. https://doi.org/10.1016/j.compedu.2020.103877

Keržic, D., Danko, M., Zorko, V., & Decman, M. (2021). The effect of age on higher education teachers' ICT use. *Knowledge Management and E-Learning*, *13*(2). https://doi.org/10.34105/j.kmel.2021.13.010

Ministry of Public Administration (Croatia). (2017). *e-Croatia Strategy 2020*.

Ministry of Regional Development and EU Funds. (2018). *National development strategy Croatia 2030*. http://www.hrvatska2030.hr

OECD. (2019). OECD. (2019). PISA 2018 insights and interpretations. OECD Publishing, 64. https://www.oecd.org/pisa/PISA 2018 Insights and Interpretations FINAL PDF.pdfPISA 2018 insights and interpretations. *OECD Publishing*.

OECD. (2020). *ICT resources in school education: What do we know from OECD work?*

Pović, T., Veleglavac, K., Čarapina, M., Jagušt, T., & Botički, I. (2015). Primjena informacijsko-komunikacijske tehnologije u osnovnim i srednjim školama u Republici Hrvatskoj [Use of information-communication technology in primary and secondary schools in Croatia]. *Proceedings of the Carnet User Conference 2015*.

Vuk, B., Mornar, V., & Boticki, I. (2020). ICT in School Education in Croatia. In *Lecture Notes in Educational Technology*. https://doi.org/10.1007/978-981-15-6157-3_3

Wei, B. (2020). An Exploratory Study of Teacher Development in the Implementation of Integrated Science Curriculum. *Research in Science Education*, *50*(6). https://doi.org/10.1007/s11165-018-9768-x

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

The Influence of an Online Community of Practice on Professional Development for Online Teaching: Case Study on Kindergartens

Yuk Mui HEUNG[a*], Alex Wing Cheung TSE[b*]
Faculty of Education, The University of Hong Kong, Hong Kong
[a*]u3573782@connect.hku.hk

Abstract: The unexpected closure of schools, including kindergartens, due to the coronavirus pandemic in the past two years transformed face-to-face teaching into online teaching, which was not a common practice for kindergarten teachers. As research reveal, kindergarten teachers are generally inexperienced in teaching online, so they need ongoing support and interactive professional training like building a community of practice (CoP) for switching to online teaching. Yet, limited research has investigated the impacts of building an online CoP as a teacher professional development approach in early childhood education. Thus, this study examined the effects and sustainable factors of the CoP groups provided by a university for online teaching in early childhood education through intensive training and quadripartite collaboration among kindergarten teachers, teacher trainers, preservice teachers, and eLearning specialists. To have an in-depth investigation, this study employed a case study of three kindergartens, semi-structured interviews with eight kindergarten teachers and three university teacher trainers were conducted via Zoom. Then, multi-sourced data, including artifacts and interview data, were analyzed by thematic analysis based on the CoP Theory to elucidate participants' experiences of an online CoP as an approach for teacher professional development. Based on the findings, this study revealed that the CoP empowered kindergarten teachers to teach online by (1) developing their technical and leadership skills through (2) building up a supportive network for idea exchanges; and (3) accumulating resources, skills, and tools they needed for online teaching from the domain, community, and practice dimension accordingly. In addition, the findings indicated that content relevance to the school curriculum, availability of time for collaboration, continuous technical support, and self-learning resources are important factors for sustaining the online teaching practices learned from the CoP groups. This study recommended an establishment of an ongoing online CoP for facilitating more sustainable teacher professional development in early childhood settings.

Keywords: Community of Practice, Teacher Professional Development, Online Teaching, Early Childhood Education

1. Introduction

The outbreak of coronavirus disease in 2019 resulted in the closure of kindergartens and the forced transition from in-person education to online teaching in Hong Kong, which aroused the attention of teacher training on this aspect, especially for kindergarten teachers. Inadequate professional training and support were found to have negatively impacted the effectiveness of kindergarten teachers' online instruction (Hu et al., 2021). Seeing the pressing needs, the Education University of Hong Kong (EdUHK) launched the community of practice (CoP) project for online teaching in early childhood education in 2021. The project was meant to strengthen the capability of kindergarten teachers to teach online through intensive training and quadripartite collaboration, specifically among kindergarten teachers, teacher trainers, preservice teachers, and eLearning specialists. Although some existing research has outlined the benefits of a CoP for upskilling teachers' workforce and enhancing the quality of online teaching (Printy, 2008), limited research has investigated its impacts on kindergarten teachers' professional development in early childhood education (To et al., 2021). This paper aimed to explore whether the online CoP groups affected kindergarten teachers' professional development for online teaching by examining the factors that might affect the sustainability of the online teaching practices supported by the CoP groups.

2. Literature Review

Kindergarten Teachers' Unpreparedness for Online Teaching

A study conducted by Hu et al. (2021) showed that kindergarten teachers had little preparation time and school training for online teaching. Deficiencies in teachers' capacity for technology adoption, technical support, technological resources, and up-to-date training hindered technology integration in the early childhood education settings in Hong Kong (Han, 2003; Hu & Yelland, 2017). Kindergarten teachers encountered significant difficulties teaching online as there was a gap between the teacher education courses and the ICT level they incorporated (Kalogiannakis, 2010). The research found that the suitability of technological tools, teachers' capability, and professional development opportunities have affected teachers' online teaching competence (Saxena & Prasad, 2021). In summary, previous studies underline the needs to enhance teachers' efficacy in online teaching.

CoP as a Practical Approach to Enhancing Teachers' Professionalism

Although the Hong Kong Education Bureau encouraged kindergarten teachers to participate in training organized by the learning communities of schools (Curriculum Development Council, 2017), little evidence of implementing learning communities such as CoP in kindergartens has been found. Hence, it was worth conducting a case study to investigate the possible impacts of CoP on kindergarten teachers' professional development in Hong Kong.

Reasons for Studying CoP for Kindergarten Teacher Professional Development

Teacher professional development for kindergarten teachers generally consists of a series of one-off workshops and meetings and an ongoing learning process through engagement in practice, comprised of informal activities such as peer sharing (Schaler & Fusco, 2003). After teachers' reflection, the approach embodies accumulating new knowledge and applying it to existing practices (Moore, 2008). Effective teacher professional development focusing on job-embedded responsibilities can yield positive outcomes when it is sustained over time in the CoP (Bokiev et al., 2017). Still, limited evidence of ongoing professional development activities recurred in a CoP was found in the kindergartens of Hong Kong. Therefore, this study focused on the impacts of communities of practice on kindergarten teachers' professional development through co-developing online learning materials.

Communities of Practice (CoP)

From the sociocultural perspective, CoP considers teachers' professional development in a social context requiring participation and reification among groups of people that share mutual engagement, a joint enterprise, and a shared repertoire of ways of doing things regularly (Lave & Wenger, 1991; Wenger, 1998). Members exchange ideas and knowledge to find solutions with common interests in a subject or problem. The theoretical framework of this paper was based on the three crucial dimensions of a CoP theory (Wenger, 1998; Wenger-Trayner & Wenger-Trayner, 2015), namely the domain, the community, and the practice. As shown in Figure 1, the domain refers to mutual engagement as an identity is defined by a shared domain of interest. Members are committed to the domain and therefore share a sense of competence; the community represents the joint enterprise as the mutual relationships sustained through interactions and collaborations among the members; the practice is related to the shared repertoire of resources: experiences, stories, tools, and solutions to recurring problems (Wenger, 1998). Situated learning emerges through a process of participation and reification. Teachers' involvement in the CoP is conducive to value creation by acknowledging that practice is an intrinsic condition of the existence of knowledge, forming a mechanism for the development, dissemination, and reification of knowledge (Hernaez, 2011). The characteristics and the key dimensions of the CoP have been clarified from the above literature reviews to guide the data analysis of this study.

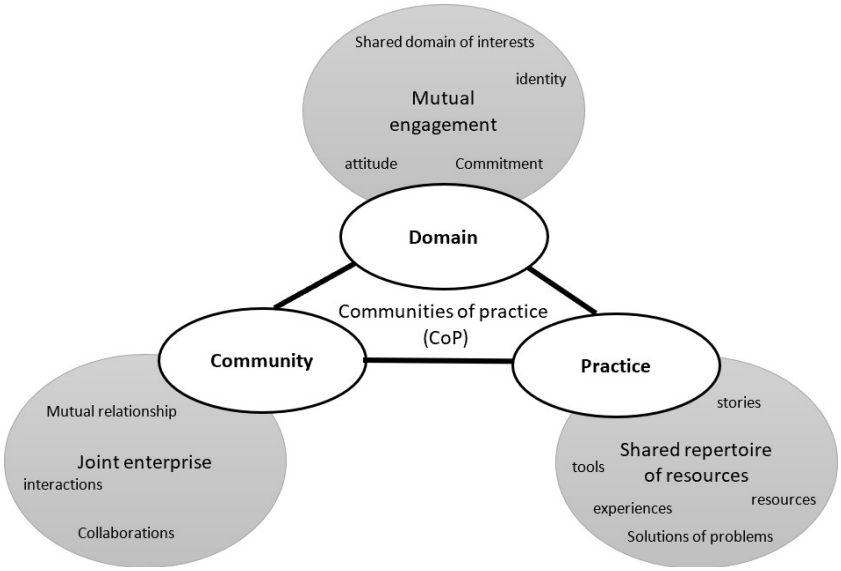


Figure 1. The Community of Practice Model (adapted from Wenger, 1998, p.72; Wenger-Trayner & Wenger-Trayner, 2015)

Prior research on CoP generally attributed its application to higher education (Cesareni et al., 2011; Correia & Davis, 2008). A CoP can function as vocational training for teachers to improve their use of technologies in schools, operating as an approach for teacher professional development and sharing reifications of teaching practices (Kirschner & Lai, 2007). Conversely, little research has focused on its implementation in early childhood education settings (Kirkby et al., 2019).

Regarding the sustainability factors of CoP, a sustainability model emerged from a study conducted by Kezar and Gehrke (2017, p.11), which includes leadership development; distribution and succession; a viable financial model; a professional staff; feedback and advice mechanisms; research and assessment; and an articulated strategy. This model suggested relevant factors that might affect the sustainability of applying online teaching practices by kindergarten teachers.

Communities of Practice Project for Online Teaching in Early Childhood Education

Communities of Practice (CoP) Project for Online Teaching in Early Childhood Education was a part of a university-level project. Given the pressing needs for training kindergarten teachers in designing online teaching materials as a substitution for face-to-face classes, the project intended to build the online teaching capability of kindergarten teachers through quadripartite collaboration and intensive training. Six CoP groups were formed, one for each of the six kindergartens in Hong Kong, when the project started in January 2021. Each group worked together from January to July 2022 to co-develop six sets of online teaching materials in the six learning areas in the Curriculum Framework of Kindergarten Education recommended by the Hong Kong Curriculum Development Council (2017).

3. Methodology

As mentioned in the introduction, few studies have focused on the influences of CoP on teacher professional development for online teaching in early childhood education. The present case study addressed this issue with the following research questions:

1. How does the online community of practice groups (CoP groups) affect the professional development of kindergarten teachers for online teaching?
2. What factors affect the sustainability of the online teaching practices of kindergarten teachers learned from the online community of practice groups (CoP groups)?

Research Design

This study employed a case study (Yin, 2018) to delve into the impacts and sustainability factors of CoP groups on teacher professional development for online teaching in an early childhood education setting.

Data Collection

This case study is a follow-up study to the CoP project that ended in July 2021. To have an overview of the project and the contextual background of each CoP group, this study began with reviewing the project documents and artifacts, including the project report and published online learning materials co-developed in the CoP groups. These documents later served as evidence to support participants' claims during the analysis of the interviews. The semi-structured individual interviews for about thirty minutes took place in mid-April 2022, nine months after the end of the project, with eight kindergarten teachers from three different kindergartens and three teacher trainers who supervised the corresponding groups. These participants were purposively targeted as they were the co-supervisors and members from three of the CoP groups. Participants' views on their current teaching practices and experience in the CoP groups were collected. The interview design involved reviewing relevant literature related to building a CoP (Cesareni et al., 2011; Printy, 2008) based on the theoretical framework of the CoP Theory in three dimensions (Wenger, 1998; Wenger-Trayner & Wenger-Trayner, 2015). All these participants clearly understood the primary purposes of this study. Before the main study, a project-in-charge EdUHK academic participated in the pilot study to provide suggestions on the instruments.

Triangulation of This Study

This study collected multi-sourced data for triangulation from participants, including artifacts and interview data with kindergarten teachers and teacher trainers, to validate the responses from various sources (Yin, 2018). Data collected from the interviews were checked and verified with the project documents and artifacts to yield the study's validity.

Ethical Measures

After the research plan was constructed, an ethical review application form was approved by the panels of the Office of Research. Consent forms that clarified this study's purposes and significance were sent to the participants. After receiving the signed consent forms, individual interviews were arranged in April 2022 for two weeks. Throughout the study process, participants were free to withdraw at any point without any negative effect. Information obtained was anonymized and treated as confidential.

4. Data Analysis

To conduct rigorous qualitative data analysis, the two investigators undertook a multistage process of examining, coding, categorizing, and interpreting data as outlined by Mills and Gay (2015) using a qualitative research coding software called Delve Tool. Deductive codes were derived from the theoretical framework of CoP Theory and the relevant literature related to the characteristics and applications of CoP (Wenger, 1998; Wenger-Trayner & Wenger-Trayner, 2015). Codes extracted from the data were reviewed to identify and match patterns for generating themes accordingly (Clarke & Braun, 2017). After generalizing the themes, the codes were rearranged based on the themes using Delve Tool. Then, the themes were regrouped into three dimensions: domain, community, and practice for analyzing the three kindergarten cases in the following section.

5. Results and Findings

Code of the Participants

The participants were coded in terms of their roles, namely teacher (T) and co-supervisor (CS) supporting that kindergarten (KG) to protect their privacy, as shown in Table 1.

Table 1. *Coding of participants*

Code	Representation	Example	Range of code
CS	Teacher trainer who co-supervised a kindergarten in the CoP group involved in the research	"CS1" stands for the co-supervisor who co-supervised the first kindergarten involved in the interview	From CS1 to CS3
T	Kindergarten teachers who participated in the CoP group involved in the research	"T1" stands for the first kindergarten teacher involved in the interview	From T1 to T3 (from KG1) From T4 to T6 (from KG2) From T7 to T8 (from KG3)

Contextual Background of the Kindergarten Teachers' Online Teaching Experience

In this study, most kindergarten teachers reported only half a year to a year of online teaching experience before joining the CoP project. They started online teaching in the first year when face-to-face classes were suspended in 2020 because of the pandemic (T1, T8, CS1, CS2). Without much training on online teaching (T2, T5), kindergarten teachers, thus, encountered difficulty from the very beginning when attempting to teach online, especially in creating online teaching materials for their students (T1, T7, CS3). For example, as T1 stated, "*We were not too skilled at editing the film or making it appealing to children. In the beginning, we did not know how to use green screens and create a virtual background.*" They were unsure about how to conduct online teaching properly and perceived it as less interactive than regular face-to-face classes. Some said the online teaching mode was less inspirational to students' thinking (T5, T8). It was challenging to sustain children's concentration online (T7). Overall, before joining this CoP project, kindergarten teachers were generally unconfident about online teaching suitability and underestimated the strengths of teaching online.

Research Question 1: Impacts of CoP on Supporting Kindergarten Teachers Switching to Online Teaching

The Domain Dimension

Aligning with the CoP project, the CoP theory was used to analyze the data involved; the results and findings are separated into three major parts, as shown in Table 2. Those about the domain dimension are discussed first.

Table 2. *Theme categorization and codes of the impacts of CoP groups in the domain dimension*

Dimension	Themes	Codes	Examples
Domain	Sharing domain of interests	Focusing on the same task; committed to completing the tasks	*"We commented to one another, intending to make high-quality teaching videos. We wanted to enhance the learning effectiveness of our students with online teaching."*

	Growing confidence	More confident; grew confidence	*"After finishing the project, we were more confident in our video shooting and post-production skills."*
	Job satisfaction	Felt satisfied	*"We felt satisfied. Our students recognized us when we taught with these videos online, they were delighted."*

As stated in the literature review, the domain dimension refers to mutual engagement as an identity is defined by a shared domain of interest. CoP group members are generally committed to the domain and share a sense of competence. In this study, each CoP group had a shared domain of interests as group members concentrated on developing six sets of online teaching materials in the six learning areas. As demonstrated in Table 2, T5 described: "*We commented to one another, intending to make high-quality videos. We wanted to enhance the learning effectiveness of our students with online teaching.*" CS1 added that she was glad to see all members commit to completing the tasks, resulting in a positive outcome. Half of the kindergarten teachers (T1, T2, T6, and T8) showed confidence in improving students' attention and teacher-student interaction through videos. As T2 put it: "*After finishing the project, we were more confident in our video shooting and post-production skills. […] The video retained students' concentration. The videos allowed our students to learn at home, combined with hand-held materials. The learning outcomes were achieved effectively."*

It was also found that the CoP groups had enhanced kindergarten teachers' job fulfillment. T1 felt satisfied with herself as she accomplished the tasks. She was thrilled to see her students answer her questions and identify her in the videos. What impressed her most was that some students who initially had weak concentration could stay focused for six to eight minutes when she used the co-developed materials in her class. T4 was touched by the parents who applauded their efforts in marking the videos as the learning experience became vivid: "*After using online teaching materials in our class, students' parents appreciated that. Our students were excited to dance with the animals in the videos.*" Above all, kindergarten teachers generally grew confidence and satisfaction with online teaching after committing to the quadripartite collaboration in the CoP groups.

The Community Dimension

As mentioned previously in the literature review, the community refers to the joint enterprise as the mutual relationships sustained through social interactions and collaborations among the members. As shown in Table 3, both teacher trainers and kindergarten teachers engaged in joint activities and discussions in the CoP groups.

Table 3. *Theme categorization and codes of the impacts of CoP groups in the community dimension*

Dimension	Themes	Codes	Examples
Community	Building a strong relationship	Close/harmonious/ comfortable/strong relationship	*"We had a good relationship. As our CoP group members had different strengths, we learned from one another during our discussions and collaborations."*
	Interacting socially	Provided advice; asked for support; consultations; discussions	*"When designing the teaching plans and the teaching events, the supervisor provided some advice on improving our lesson design, while we asked for support from eLearning specialists."*
	Collaborating as a team	Collaborated well; worked together; solved problems together	*"We pre-downloaded some software, and then our CoP group members taught us how to create animations. During the process, we solved problems together."*

By forming a networked partnership, teacher trainers and kindergarten teachers in the CoP groups generally had positive views on the relationship and collaboration among the members as they

learned from one another and enhanced the quality of the videos after negotiating and seeking consensus (T1, T5, T7, CS1, CS2). Various types of members contributed to the CoP groups based on their strengths. For example, as T7 explained, "*The supervisor provided some advice on improving our lesson design, while we asked for support from eLearning specialists*." CS2 added, "*The two preservice teachers learned the technical skills from the eLearning specialists for supporting the kindergarten teachers*." Teachers from the same kindergarten had extensive collaboration. For example, T1 recalled, "*I work closely with T2. We prepared lesson plans and revised contents based on our school curriculum.*" The above findings showed that the CoP groups generally supported the kindergarten teachers through social participation with different group members harmoniously in a supportive learning environment.

The Practice Dimension

The practice dimension is about a shared repertoire of resources, including experiences, stories, tools, and solutions to recurring problems (Wenger, 1998). As illustrated in Table 4, more than half of the kindergarten teachers noted that the shared resources provided in the CoP groups were useful for their online teaching. T1, T4, T6, and T8 shared that the resources posted on the online platform could inspire them in curriculum design as they gained insights by referring to others' teaching styles, contents, and the use of other eLearning tools with interactive game templates. T1 explained: "*Other preservice teachers' sharing of online teaching inspired us. […] We adopted game-based learning, making the teaching materials more exciting and diverse.*"

Table 4. *Theme categorization and codes of the impacts of CoP groups in the practice dimension*

Dimension	Themes	Codes	Examples
Practice	Utilizing tools	Tools; software; digital device	*"We created an interactive game with Word Wall for the topic of counting down numbers."*
	Strengthening techniques	Techniques; green screen; animations; virtual objects; video editing; video shooting; lesson plan design	*"We learned different techniques, such as creating stop-motion animations, removing green screens, and adding some virtual objects to the online teaching materials. My presenting skills have been improved."*
	Accumulating resources	Resources; sharing from other teachers; online teaching materials	*"Other preservice teachers' sharing of online teaching inspired us. We did not know stop-motion animation could be used in an instructional video."*

As mentioned in the introduction, most kindergarten teachers expressed concerns about the interactivity of online teaching before joining the community. In the CoP groups, most of them learned how to apply different techniques to make the videos interactive with students. As demonstrated by T6, "*We created an interactive game with Word Wall for the topic of counting down numbers. Our students followed the video and counted down the numbers together.*"

More importantly, most kindergarten teachers involved (T1, T2, T5, T7, T8) are regarded to have developed a sense of competence in technical skills, knowledge, and tools for lesson design and video production after participating in the project. For example, T1 said, "*We learned different techniques, such as creating stop-motion animations, removing green screens, and adding some virtual objects to the online teaching materials. My presenting skills have been improved.*"

The above examples demonstrated that through the professional development in the CoP, kindergarten teachers utilized online teaching tools to raise students' engagement. Kindergarten teachers indicated that they played a leadership role in leading teachers to incorporate online teaching approaches after being community members. Peer-to-peer teacher professional development activities were organized in the kindergartens. For example, T1 recalled, "*We organized a sharing session to share what we learned in CoP with our colleagues*." Meanwhile, T5 praised, "*Most teachers acquired animation creation skills through a mini-workshop, promoting what we learned in CoP with our colleagues*." This implied that good practice extended outside the CoP group by developing kindergarten teachers' leadership identity. Overall, a conclusion can be drawn that most kindergarten

teachers strengthened their technical skills for online teaching by utilizing the shared resources accumulated in the CoP among the kindergartens. Some also embraced their leadership role by promoting good practices with their colleagues.

Results and Findings Addressing the Impacts of CoP on Supporting Kindergarten Teachers Switching to Online Teaching

Overall, most kindergarten teachers in this study reflected that their participation in the CoP groups contributed to their professional growth in switching to online teaching. In the domain dimension, it was found that they became more confident and satisfied with their online teaching after sharing the commitment to creating online teaching materials with the CoP group members; in the community dimension, kindergarten teachers learned through socially interacting and collaborating with the group members harmoniously, building a supportive learning environment; in the practice dimension, most kindergarten teachers acquired technical skills for creating online learning materials with the help of the shared resources in CoP, which facilitated idea exchanges and accumulated resources by sharing the materials they co-created with other CoP groups among the kindergartens. Some of them also shared the good practice with their colleagues, leading to developing their leadership roles.

Research Question 2: Potential Factors Affecting the Sustainability of the Online Teaching Practices Learned from the CoP

This section discusses teacher trainers' and kindergarten teachers' views on the sustainability factors that might affect the online teaching practices they learned in the CoP groups from the domain, community, and practice dimensions. In the domain dimension, kindergarten teachers reported reusing the online learning materials they created during the CoP experiences in a different academic year as the contents were related to their school curriculum (e.g., T3, T5, T7). For example, T7 disclosed: "*Our principal appreciated our participation in this online CoP group and encouraged us to reuse our learning resources."* Similarly, T3 mentioned that *"The formation of the CoP was right at the time to support teachers in developing online teaching skills and video creation skills. It was complimentary to our school-based plans. We could see how the online learning materials were combined with the hands-on material packages for facilitating students' learning.*" A group of kindergarten teachers (e.g., T3, T7) perceived online teaching as a new trend in kindergartens to supplement students' learning.

Although the CoP groups in this project linked people with shared interests in creating online learning materials for kindergartens, there are two inhibiting factors affecting the sustainability issues: limitation of project timeframe and members' identity change effects. Members of the CoP groups had limited communication after the project ended as CS3 revealed that "*The community built in the project was terminable. The groups stopped growing after the end of the project. For promoting sustainability, gathering a group of people can drive the motives to the kindergartens as pioneers*". It shows that maintaining the CoP groups led by passionate educators on online teaching in early childhood education could facilitate participation and strengthen the sustainability of the CoP groups. A change in members' identity may also affect the bonding among group members. For example, T4 shared, "*The two preservice teachers have graduated and become kindergarten teachers. We did not keep in touch*." It shows that some kindergarten teachers raised concerns over connecting members from non-partnered kindergartens after changing their jobs or roles.

Regarding the community dimension, most kindergarten teachers from the second kindergarten considered continuous technical support from the community members effective for idea exchanges and discussion on new software and tools. Both the teacher trainers and kindergarten teachers responded that sufficient time for collaboration was notable for sustaining their practices in the CoP groups. For example, T1 emphasized, "*Sufficient time for collaboration within the CoP group is essential. We took one to two days to create a 10-minute video. We needed to re-watch it to check if there was any error.*" A sufficiency of kindergartens' workforce and resources are contributing factors as CS2 explained that "*Participation in the CoP is the extra workload added to the teachers who may have difficulty balancing their normal work and tasks in the CoP group.*"

In the practice dimension, most participants, both the teacher trainers and the kindergarten teachers, regarded access to self-learning resources as favorable criteria for sustaining kindergarten

teachers' practice for online teaching (e.g., T5, T6, T8, CS1, CS2). A typical response from CS1 showed: "*Self-paced learning materials upskill them [kindergarten teachers]. After posting tutorial videos on the platform, they [kindergarten teachers] can learn in their free time.*" T5 appreciated that the resource platform contained many resources for self-learning and free elements. Kindergarten teachers could access the learning materials even after the end of the project. Hence, they received the opportunities to learn and apply online learning practices sustainably.

Results and Findings Addressing the Potential Factors Affecting the Sustainability of the Online Teaching Practices learned from the CoP

In short, the above findings revealed that most kindergarten teachers sustainably applied what they learned in the CoP in their teaching practices because of the following favorable factors: (1) aligning the contents with the kindergarten curriculum, (2) having sufficient time and technical support for collaboration, and (3) given access to self-learning resources. On the other hand, the findings unveiled three unfavorable factors that may cause inhibitory effects: (1) limitation of project timeframe, (2) effects of identity change, and (3) limited workforce and resources.

6. Discussion

Based on the findings, it was generally found that kindergarten teachers positively developed a more robust collaborative partnership with the members of CoP. Kindergarten teachers have enriched the teaching and learning experience with online interactive tools and materials. The results reinforce the theory of Wenger (1998) that teachers can achieve learning goals through active involvement in implementing knowledge and skills with a group of people who share similar goals. Although this study focused on the implementation of CoP in early childhood education settings, it confirmed the findings of earlier studies in primary and secondary schools that teachers developed their technical and leadership skills, confidence, flexibility, and creative and innovative thinking abilities in the CoP (Hammersley-Fletcher & Kirkham, 2007; Mercieca & McDonald, 2021).

Regarding the factors for sustaining the practice, this study has emerged a list of sustainability factors of the CoP, partly coinciding with the sustainability model proposed by Kezar and Gehrke (2017); for instance, continuous technical support and time in the community and resources. Moreover, this study raised the alignment with the kindergarten's curriculum in context as an additional factor to sustain the online teaching practices kindergarten teachers learned in the CoP.

7. Conclusion

Limitations and Implications for Further Study

This case study was a follow-up study to a CoP project, given the time and resources, resulting in the professional growth of kindergarten teachers. Kindergarten teachers joined the project during the class suspension, which gave them sufficient time to co-create online learning materials. This study, however, might not reflect the cases of every kindergarten that participated in other CoPs or other periods. More in-depth research could be done regarding the ongoing implementation of the CoP project.

In conclusion, this study has unveiled that kindergarten teachers' participation in the CoP groups enhanced their readiness for online teaching by building a supportive professional network for idea exchanges, strengthening their online teaching competencies, and accumulating online teaching resources. It facilitated the evolution of online teaching practices in kindergartens. To sustain the legacy of the CoP groups, it is essential to consider sustainable factors such as aligning the contents with the kindergarten curriculum, providing self-learning resources, continuous technical support, and sufficient time for collaboration within the CoP groups. This study has contributed to research on the values of forming a long-term community of practice for kindergarten teachers' sustainable professional development in early childhood education.

Acknowledgments

We sincerely thank the individuals and kindergartens involved in this study.

References

Bokiev, U., Bokiev, D., Othman, M., Ismail, L., & Aralas, D. (2018). Online Teaching Portfolio as a Catalyst for Professional Development. *International Journal of Academic Research in Business and Social Sciences, 7*(14). https://doi.org/10.6007/IJARBSS/v7-i14/3680

Curriculum Development Council (CDC) (2017). *Kindergarten education curriculum guide: Joyful learning through play, balanced development all the way*. Hong Kong Special Administration Region (SAR): Retrieved from https://www.edb.gov.hk/attachment/en/curriculum-development/major-level-of-edu/preprimary/ENG_KGECG_2017.pdf

Cesareni, Martini, F., & Mancini, I. (2011). Building a community among teachers, researchers, and university students. A blended approach to training. *International Journal of Computer-Supported Collaborative Learning, 6*(4), 625–646. https://doi.org/10.1007/s11412-011-9126-8

Clarke, V., & Braun, V. (2017). Thematic analysis. *The Journal of Positive Psychology, 12*(3), 297–298. https://doi.org/10.1080/17439760.2016.1262613

Correia, A.-P., & Davis, N. (2008). Intersecting communities of practice in distance education: the program team and the online course community. *Distance Education, 29*(3), 289–306. https://doi.org/10.1080/01587910802395813

Hammersley-Fletcher, L., & Kirkham, G. (2007). Middle leadership in primary school communities of practice: Distribution or deception. *School Leadership & Management, 27*(5), 423-435.

Han, C. W. (2003). Challenges of using ICT in Hong Kong early childhood settings. *IFIP Working Group 3.5 Conference: Young Children and Learning Technologies*. Parramatta, Australia

Hernaez, O. R. (2011). *Handbook of research on communities of practice for organizational management and networking : Methodologies for competitive advantage*. IGI Global.

Hu, X., & Yelland, N. (2017). An investigation of preservice early childhood teachers' adoption of ICT in a teaching practicum context in Hong Kong. *Journal of Early Childhood Teacher Education, 38*(3), 259–274. https://doi-org.ezproxy.eduhk.hk/10.1080/10901027.2017.1335664

Hu, X., Chiu, M. M., Leung, W. M. V., & Yelland, N. (2021). Technology integration for young children during COVID-19: Towards future online teaching. *British Journal of Educational Technology, 52*(4), 1513– 1537. https://doi.org/10.1111/bjet.13106

Ignatow, G. & Mihalcea, R. (2017). *Thematic analysis, qualitative data analysis software, and visualization. In Text mining (pp. 74-87).* SAGE Publications, Inc, https://dx.doi.org/10.4135/9781483399782

Kalogiannakis, M. (2010). Training with ICT for ICT from the trainer's perspective. A local ICT teacher training experience. *Education and Information Technologies, 15*(1), 3–17.

Kirkby, J., Walsh, L., & Keary, A. (2019). A case study of the generation and benefits of a community of practice and its impact on the professional identity of early childhood teachers. *Professional Development in Education, 45*(2), 264–275. https://doi.org/10.1080/19415257.2018.1449003

Kirschner, P. A., & Lai, K.W. (2007). Online communities of practice in education. *Technology, Pedagogy and Education, 16*(2), 127–131. https://doi.org/10.1080/14759390701406737

Kezar, A., & Gehrke, S. (2017). Sustaining Communities of Practice Focused on STEM Reform. *The Journal of Higher Education (Columbus), 88*(3), 323-349.

Mercieca, B., & McDonald, J. (2021). *Sustaining Communities of Practice with Early Career Teachers*. Singapore: Springer Singapore Pte. Limited.

Mills, G., & Gay, L. (2015). *Educational research: Competencies for analysis and applications* (Eleventh edition, Global ed.). Boston, Massachusetts: Pearson.

Moore, B. (2008). Using technology to promote communities of practice (CoP) in social work education. Social Work Education, 27(6), 592–600.

Printy, S. (2008). Leadership for Teacher Learning: A Community of Practice Perspective. *Educational Administration Quarterly, 44*(2), 187-226.

Saxena, A. & Prasad, K. (2021, August 18-20). Tools for Distance Education in Hong Kong Kindergartens During Unexpected School Closures, *Proceedings of the 16th International Conference on Computer Science & Education (ICCSE 2021)*. Lancaster, UK.

Schaler, M. S., & Fusco, J. (2003). Teacher professional development, technology, and communities of practice: Are we putting the cart before the horse? *Information Society, 19*, 203-220.

To, K., Yin, H., Tam, W., & Keung, C. (2021). Principal leadership practices, professional learning communities, and teacher commitment in Hong Kong kindergartens: A multilevel SEM analysis. *Educational Management, Administration & Leadership*, 174114322110152.

Wenger, E. (1998). *Communities of practice: learning, meaning, and identity*. Cambridge; New York: Cambridge University Press.

Wenger-Trayner, E., & Wenger-Trayner, B. (2015). *Introduction to communities of practice. A brief overview of the concept and its uses.* Accessed May 5, 2022.

Yin, R. K. (2018). *Case study research and applications: design and methods (Sixth edition.).* Thousand Oaks, California: SAGE Publications, Inc.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Development and Validation of a Brief Digital Pedagogy Competency Scale (DiPeCoS)

Shraddha RAWAT, Shreya TIWARI, Mayank SHARMA & Nandini CHATTERJEE*
UNESCO Mahatma Gandhi Institute of Education for Peace and Sustainable Development, 35, Ferozshah Road, New Delhi – 110 001, India
*n.chatterjee@unesco.org

Abstract: Digital pedagogy or the purposeful application of digital technology in teaching and learning, has the potential to significantly enhance learning experiences. Considering this potential and the rapid digitization in education, it has become critical to define, measure and build teacher's digital pedagogy competence. Although there are a number of self-report tools to evaluate digital pedagogy competence, there is a paucity of scenario-based tools for the same. Scenario-based assessments allow demonstration of knowledge and skills as well as its real-world application. We present here, DiPeCoS, a short tool that assesses a teacher's digital pedagogy competence through their decision making under teaching learning scenarios. Using an item response analysis, a pool of 10 items was reduced to an 8-item scale and validated on a sample of 1315 teachers in India. The DiPeCoSe demonstrated unidimensionality, and its constituent items showed acceptable levels of discrimination, difficulty and guessing parameters. Additionally, it also demonstrated acceptable values of reliability. We hope that this scale can be used to conduct training needs assessment amongst teachers and to assess the effectiveness of digital pedagogy training in future research.

Keywords: Digital Pedagogy, India, Teacher Competence Scale, Universal Design for Learning, Teacher Digital Competence

1. Introduction

The rapid advancement in technology in the 21st century presents a landscape shaped by technology that places new demands on the teaching learning experience and has broad implications both in terms of what it means to teach, and how one teaches and creates a learning experience. Present day teachers do not just need technological knowledge and ICT skills but also the ability to leverage technology affordances for modernizing teaching practice. Numerous efforts have been made to promote technology-enabled learning through ICT training for teachers. This approach has not considered the fact that the intentional use of technology affordances by teachers to achieve learning goals requires a refreshed set of digital pedagogy competence (Fernández-Batanero et al., 2021, Sailer et al., 2021).

Merely integrating technology in educational programs with the aim of 'building digital skills' does not fully utilize the power of meaningful application of technology to transform the learning experience. Research suggests that 'learning occurs when access to technology is combined with relevant and engaging content, a well-articulated instructional model, effective teaching presence, learner support, and an enabling learning environment' (UNICEF, 2020). However, such use of technology requires teachers to be equipped with a whole new skill set and perspective connected to application of digital competencies in teaching and learning. In this paper, we refer to these skills and perspectives as a teacher's digital pedagogy competence and develop the Digital Pedagogy Competence Scale (DiPeCoS) to assess it.

2. What is Digital Pedagogy?

An analysis of literature (Kivunja, 2013; Montebello, 2017; Sailin & Mahmore, 2018) highlights that mere integration of technology in teaching and learning does not qualify as digital pedagogy. Rather, the purpose of technology integration must be to enrich or enhance learning. Such an intentional use of technology warrants technological skills, pedagogical skills as well as the ability to integrate both.

Given that the purpose of enhancement or enrichment of learning is at the center of digital pedagogy, it is essential to define what 'enhanced' or 'enriched' learning entails. Although defining and measuring the quality of teaching and learning is complicated (Fink, 2003; Berman, 2003; McCabe and Layne, 2012), it is widely agreed that learners have variable characteristics, preferences, needs, and abilities which need to be considered in the design of a learning experience to improve the experience for all (Al-Azawei et al., 2016).

Universal Design for Learning (UDL) is a framework to improve and optimize teaching and learning for all learners based on insights from cognitive neuroscience (CAST, 2015). Literature suggests a positive relationship between application of UDL and student interest and engagement (Smith, 2012) as well as its potential to improve students' academic and social outcomes (Ok et al., 2017). There are 3 principles of UDL which guide the design of learning to suit a wide variety of learners. These are:

- Providing multiple means of engagement: This emphasizes the 'why' of learning and offers checkpoints on recruiting interest, sustaining effort & persistence and self-regulation.
- Providing multiple means of representation: This focuses on the 'what' of learning and captures checkpoints under - perception, language & symbols, and comprehension.
- Providing multiple means of action & expression: This focuses on the 'how' of learning and captures checkpoints under physical action, expression & communication, and executive function

Integrating digital tools in alignment with the UDL principles enables teachers to intentionally use technology to enhance learning. Such intentional use of digital tools in alignment with the three principles of UDL can reduce learning barriers and support students to meet learning and affective goals (Rao, 2021).

Using Universal Design for Learning as the reference framework, we propose that digital pedagogy entails leveraging digital technology with the purpose of: one, presenting information such that it can be perceived and comprehended by learners effectively; two, offering multiple strategies to engage learners such that they are motivated to learn; and three, enabling learners to navigate the learning environment and express what they know.

3. Existing Scales to Assess Digital Pedagogy Competence

Many tools have been developed to evaluate teacher's digital competence. Beyond pedagogical skills, these tools include application of technology in a range of work teachers carry out in the classroom, the educational institution, the community and in the context of their own personal and professional development (Lázaro-Cantabrana et al., 2019). Literature review of existing tools like the Self-reflection on Effective Learning by Fostering the use of Innovative Educational technologies (SELFIE) tool based on DigCompEdu (Redecker & Punie, 2017), Teachers' Digital Competencies Questionnaire based on the Common Digital Competence Framework for Teachers by INTEF (Tourón et al., 2018), Wayfind Teacher Assessment based on International Society for Technology in Education (ISTE) standards for teachers (Banister & Reinhart, 2013), COMDID based on other frameworks of teacher's digital competence (Lázaro & Gisbert, 2015) points that most tools measure teacher's digital competence, a concept that goes beyond digital pedagogy competence, and only a few tools have a targeted focus on evaluating a teacher's digital pedagogy competence.

Two tools that are closely related to the concept of teacher's digital pedagogy competence are the Survey of Preservice Teachers' Knowledge of Teaching and Technology (SPTKTT) by Schmidt et al. (2009) which is based on the TPACK model (Mishra & Koehler, 2006) and the UDL self-assessment tool by University of Waikato (2018) based on the UDL framework (CAST, 2015). Both the tools use Likert scale-based items to get teachers to reflect on their pedagogical practices (for e.g., "*I know how to select effective teaching approaches to guide student thinking and learning in mathematics*" on the SPTKTT and "*I encourage students to express their learning in multiple ways (e.g., essay, or video blog, poster or presentation)*" on the UDL tool. The problem with these tools is that they rely on self-report of respondents' behaviors, beliefs, perceptions, attitudes, or intentions, which are shown to be virtually uncorrelated with their on-the-job behavior (Thalheimer, 2018). Responding on self-report scales is also often affected by biases of social desirability (Van de Mortel, 2008), which can corrupt collected data. Thus, although self-report tools can be used for reflection and self-assessment that may

promote learning and improvements in performance (Andrade, 2019), they cannot adequately assess a teacher's digital pedagogy competence.

Any tool designed to measure competence must focus on the ability to apply knowledge and skills in real-life context. Thalheimer (2018) argues that assessing decision making is better than gauging self-perception of skills or behaviors. He also argues that one of the ways to evaluate realistic decision making is by presenting learners with realistic scenarios and prodding them to make decisions that are similar to the types of decisions they will have to make on the job. It is this form of scenario-based assessment that efficiently counters the shortcomings of self-report tools. Scenario-based assessments are based on the situated learning theory (Lave & Wenger, 1991) which states that learning and assessment best take place in the context in which they are going to be used (Kindley, 2002) and thus can be more efficient measures of assessment.

4. Purpose

There is a need for a tool to assess teacher's digital pedagogy competence which does not rely on self-report but rather makes use of scenario-based assessments in which respondents apply their subject knowledge, critical thinking and problem-solving skills in a real-world context to respond reliably. The Digital Pedagogy Competency Scale (DiPeCoS) for teachers has been developed to address this need. The following sections present the results of the validation of the DiPeCoS.

5. Materials and Methods

5.1 Participants

A total of 1315 English-speaking Indian teachers completed the virtually delivered scale. Participants reported a mean age of 42.1 years and had all received formal education in English language. 39% (N = 513) of the participants reported their gender as female and 61% (N = 802) as male. 3.6% of the teachers reported the setting of their work as "primary school", 28.5% as "middle school", 52.6% as "high school" and 15.3% as "others." Participants who reported their work setting as "others" included education counselors, consultants and freelance teachers. Teachers reported a mean teaching experience of 14.77 years (SD = 10 years, range 0-50 years).

5.2 Item Development

As mentioned above, DiPeCoS is based on the Universal Design for Learning (UDL) framework. It assesses the competence of teachers to purposefully use digital technologies to: 1) Present information such that it can be perceived and comprehended by learners effectively; 2) Offer multiple strategies to engage learners such that they are motivated to learn; and 3) Enable learners to navigate the learning environment and express what they know.

Ten multiple-choice, scenario-based items were developed to assess the ability of the respondents to leverage digital technology for the three pedagogical purposes mentioned above. For example, item 8 evaluated a teacher's ability to design a learning experience by using digital books and pointer tools to teach language was developed to assess the ability to use technology to represent information such that it can be perceived and comprehended by learners effectively. This purpose of technology integration is aligned with the UDL principle of multiple means of representation. Similarly, item 6 was developed to assess a teacher's ability to use technology to design a peer learning experience for 40 students to learn science at home. This item evaluated the use of technology to engage all the learners such that they are motivated to learn, this is aligned with the second principle of UDL, namely multiple means of engagement.

The pedagogical decision of the respondent in each scenario serves a dominant purpose and at the same time, enables other purposes too. For instance, while item 1 was developed to assess the ability of a teacher to leverage technology to facilitate peer sharing, which is aligned with the purpose of enabling learners to express what they know, peer sharing is also a strategy to engage and motivate learners and therefore serves a secondary function of engaging learners. Of the 10 items on the original scale, 3

questions were designed to assess the use of digital technology to represent information, 3 to engage learners and 4 to enable learners to navigate the learning environment and express what they know.

Popular guidelines on item development were followed, and it was ensured that the items were unbiased in terms of responding by government and private school teachers, as well as other forms of diversity. Items of the Digital Pedagogy Competency Scale (DiPeCoS) and their alignment to UDL principles can be found here: https://bit.ly/3BQOxws.

5.3 Procedure

Participants recruited for the validation of the tool were a part of a wider intervention that included taking an online course on digital pedagogies, and the scale developed in this study was utilized as a pre-post questionnaire to assess the impact of the course-based intervention. The results reported in this study were obtained from the pre-assessment responses. Recruited participants involved English-speaking teachers from English-medium schools across various parts of India, who filled out the questionnaire over a period of 4 months, from January to April 2022. Teachers working at private schools in urban areas as well as government-run residential schools operating in various parts across rural India were both a part of the study. Since many teachers did not provide details about their schools, it was not possible to determine the exact percentage of the different types of schools.

All participants were briefed about the aim of the study and given the opportunity to resolve queries related to the study and the questionnaire during a 1.5-hour long online workshop. After the workshop, participants created an account on Framerspace, an interactive learning platform (*www.framerspace.com*), where the questionnaire and the course were hosted. The questionnaire consisted of the scale described above, along with a short demographic form that collected demographic information like gender, age, teaching experience and teaching profile. Information on time required to complete the assessment was provided. Informed consent was sought from the participants. No personally identifiable information (such as name or email address) was collected from the participants. Instead, entries were recorded, stored and identified using anonymized IDs provided to all participants, which could not be linked to their identities.

5.4 Data Analysis

A total of 1824 entries were received on the questionnaire. After removal of duplicates, 1315 entries remained, which were used to perform a validation of the scale. All statistical analyses were performed on R version 4.0.2 (*https://www.r-project.org*). An item response analysis was conducted to validate the developed scale because Item response theory (IRT) is useful to investigate if the items in a scale do not have enough reliable information about the construct being measured. It can also differentiate item properties (e.g., discrimination and difficulty) among individuals across a much wider range of the construct at hand. If the analyses show that there is such a problem with some items, the researcher can remove/modify those items or add new items that help to measure these parts of the construct, thus, providing information that can differentiate people across a much greater range of the latent trait and increase the validity of the whole scale (Oishi, 2007).

To use IRT, basic assumptions pertaining to unidimensionality, local independence, monotonicity and differential item functioning (DIF) were first tested. Unidimensionality (items in the scale load on only one latent factor) was tested using factor analysis. Exploratory factor analysis (EFA) procedures such as eigenvalue extraction (Kaiser, 1960), scree test (Cattell, 1966) and parallel analysis (Horn, 1965) were used to test the presence of a unidimensional factor, which was later affirmed using confirmatory factor analysis (CFA). Goodness of fit for the CFA model was evaluated based on commonly used indices: $\chi 2_{S\text{-}B}/df < 5$ (Marsh & Hocevar, 1985), CFI, TLI > 0.90 (Bentler, 1990), RMSEA < 0.08 (Browne & Cudeck, 1992) and SRMR < 0.08 (Browne & Cudeck, 1992). All factor analytic procedures used polychoric correlations with Satorra-Bentler correction (Satorra & Bentler, 1994) given the dichotomous nature and non-normality of items (Morata-Ramírez & Holgado-Tello, 2013). Local independence was tested by examining if the chance of one item being answered was related to any other item(s) being answered or if responses to items were independent decisions taken by the test-takers. Monotonicity (meaning that the levels of a person's latent trait increase, as a monotonic function, as the probability to choose the answer in each item that represents the participant's actual

level of the trait increases) was tested using Mokken analysis and DIF was applied to investigate whether the item responses varied across gender.

After assumption testing, an appropriate IRT model was selected. Since the items were scored on a dichotomous scale (0 for incorrect response and 1 for correct response), a logistic IRT model was chosen for the study. From the 1-PL, 2-PL and 3-PL models, a 3-PL model was chosen since 3-PL models, even though more complex, make it worthwhile sacrificing parsimony by providing the best fit. Moreover, the 3-PL model is considered appropriate for multiple-choice tests (like the one in this study) where the probability of success from a very low-ability person on an item may be significantly higher than zero because of random guessing (Diamond & Evans, 1973). Generally speaking, both the 2-PL and 3-PL models are considered more suitable for cognitive tests as compared to 1-PL models.

Using the 3-PL model, item parameters were calculated, and item information curves (IICs) and item characteristic curves (ICCs) were plotted for all the items. Results of these analyses, complemented with the theoretical judgment of the researchers, were used to guide removal of items that were problematic and increase the overall validity of the scale. After the scale structure was finalized, the test information function was plotted to observe how the overall scale responded to individuals with different abilities. Finally, a reliability analysis was performed. All results reported in this paper use a *p*-value of alpha = 0.05.

5.5 Assumption Testing

We used EFA and CFA procedures to examine if the correlation among items in the scale was explained by a single latent factor. An unrestricted factor solution indicated that the magnitude of the first eigenvalue (4.11) was much greater than the magnitude of other eigenvalues (1.21, 0.91, 0.83, 0.72, 0.61, 0.56, 0.46, 0.36 and 0.18). The ratio of the first to the second eigenvalue was also greater than 3, which provided some evidence of unidimensionality (Sattelmayer et al., 2017). Complementing this, results from the scree plot and parallel analysis also hinted at single-factor solutions. A single-factor CFA model was thus built to confirm unidimensionality in the scale. The model converged normally and demonstrated a good fit: χ2S-B/df = 0.667, CFI = 0.998, TLI = 0.998 and RMSEA (90% C.I.) = 0.000 (0.000 - 0.066). All item loadings were significant at $p < .05$, except items 7 and 9.

Local independence was confirmed as: a) the chance of one item being answered was not related to any other item(s) being answered, and (b) the response to an item was every test-taker's independent decision, i.e., there was no cheating or group work involved.

Results from Mokken analysis indicated that the response function of the probability of getting a correct response on each item increased when a person's latent trait increased (for all items except item 9). Therefore, evidence of monotonicity was seen in all items except item 9.

Results of a DIF analysis indicated that all items were responded similarly by both males and females. Next, we calculated the information, difficulty, discrimination and guessing parameters for all items in order to decide whether they needed to be excluded from the scale.

5.6 ITR Model

A 3-PL IRT model was created and discrimination, difficulty and guessing parameters were calculated for each of the 10 items. All items fitted well in the model ($p > .05$). Discrimination parameter values can range from $-\infty$ to $+\infty$, but values typically fall in the range of 0 to +2.50. Item discrimination values of 0.01–0.34 are considered very low; 0.34–0.64 low; 0.65–1.34 moderate; 1.35–1.69 high; and 1.70 and above very high (Baker, 2001). Item difficulty estimates vary from -4 to 4, where -4 represents most easy, 0 represents average and +4 represents most difficult. The guessing parameter, c has a theoretical range of [0,1], but in practice, values above 0.35 are not considered acceptable. As seen from Table 1, item 5 demonstrated "low" discrimination value, items 1, 2, 4 and 7 demonstrated "medium" discrimination values, item 8 and 10, "high" and items 3, 6 and 9, "very high." Broadly speaking, items 1, 2, 3, 4, 6, 8 and 10 were "easy" items and items 5, 7 and 9 were "difficult" (based on the polarity of the difficulty parameter). Items 7 and 9 demonstrated high guessing parameters.

Table 1. *3-PL IRT model for the scale items*

Item	Guessing	Difficulty	Discrimination	X^2	P (> X^2)
Item 1	0	-0.355	1.18	31.2758	0.8713
Item 2	0	-0.506	1.081	52.6484	0.1782
Item 3	0.001	-1.538	3.619	42.2773	0.5347
Item 4	0	-1.058	1.139	34.2256	0.5941
Item 5	0	0.967	0.565	25.2309	0.5842
Item 6	0	-1.063	2.266	45.2733	0.9406
Item 7	0.432	1.236	1.08	23.3789	0.5842
Item 8	0.002	-0.965	1.64	46.9475	0.703
Item 9	0.227	1.448	7.549	58.4472	0.4851
Item 10	0	-0.815	1.552	56.7459	0.4455

Item characteristic curves (ICC) and item information curves (IIC) were also plotted for the scale items (Figures 1 and 2, respectively). For both the graphs, theta (θ) represents a person's true latent trait (factor), which has been standardized to follow a normal distribution with a range from - 4 to 4, where 0 represents the average score (Baker, 2001). For the ICC, P(θ) represents the probability of a correct answer, while for the IIC, I(θ) represent the information function, or how well each item contributes to score estimation precision (higher levels of information leading to more accurate score estimates). In an ICC, discrimination is defined as how well an item can differentiate between examinees having abilities below the item location and those having abilities above the item location. Consequently, items with ICCs which are more "spread out" indicate lower discriminability, ICCs which are farthest on the plot indicate higher difficulty, and ICCs that have a finite value of y-intercept indicate that there is a higher probability of guessing. In a similar way, IICs peak at the difficulty value (point where the item has the highest discrimination), with less information at ability levels farther from the difficulty estimate. As seen from Figure 1, items 5, 7 and 9 demonstrated higher levels of difficulty since they were placed on the right-hand side of the graph, indicating that the probability of responding to these items correctly would be high only for individuals with high ability. On the contrary, item 3 showed the lowest level of difficulty. Items with curves that were the least spread, for example, items 3, 6, 8, 9 and 10, indicated highest levels of discrimination. Since items 7 and 9 also had significant positive y-intercepts, they had a higher probability of being guessed. These results coincide with the ones inferred from Table 1.

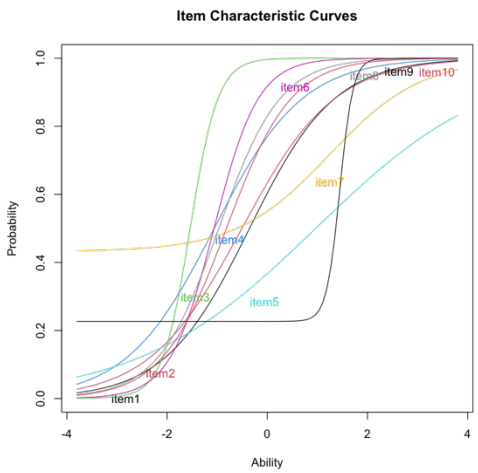


Figure 1. Item characteristic curves (ICCs) for the 10 items forming the scale

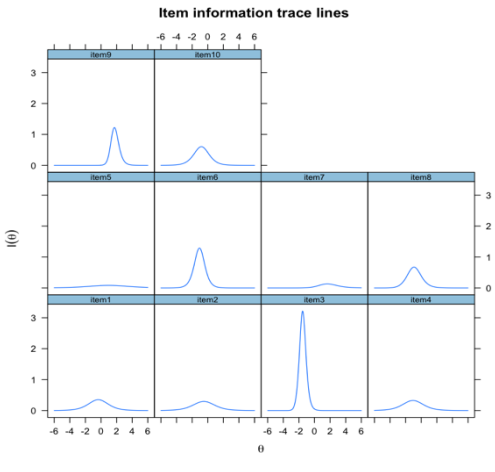


Figure 2. Item information curves (IICs) for the 10 items forming the scale

Some additional information was gathered from Figure 2. For instance, item 3 peaked very high at an ability level of θ = -2.1 and also demonstrated a narrow IIC, indicating that the item provided most information about low ability individuals. This was different as compared to, say item 4, which demonstrated a curve that was much wider spread and peaked at an ability level of θ = -0.1.

A Test Information Function (TIF) was also plotted for the overall scale (see Figure 3a), which was simply the sum of information functions of all items of the scale. As seen in the figure, the TIF was

a bimodal curve, with two peaks at ability levels θ = -1.8 and θ = 1.9. To the extent possible, the TIF should be a unimodal curve centred around θ = 0 so that the scale serves as an unbiased assessment of low and high ability individuals. In order to improve the quality of information provided by the scale, the researchers decided to remove items 3 and 9 based on the results obtained from the IRT analysis. Through a retrospective exercise, researchers concluded that it is possible that item 3 could be significantly affected by the bias of social desirability as response options (other than the correct one) provided on the item were "morally incorrect" for any teacher to answer (see details in Table 1). On the other hand, item 9 had ambiguity in the response options provided, due to which many of the responses could possibly be "correct." After removal of these items, the TIF was re-plotted. As seen from Figure 3(b), the distribution improved. It demonstrated a unimodal peak around θ = - 0.2 and was reasonably well spread. The 8-item scale was finalized.

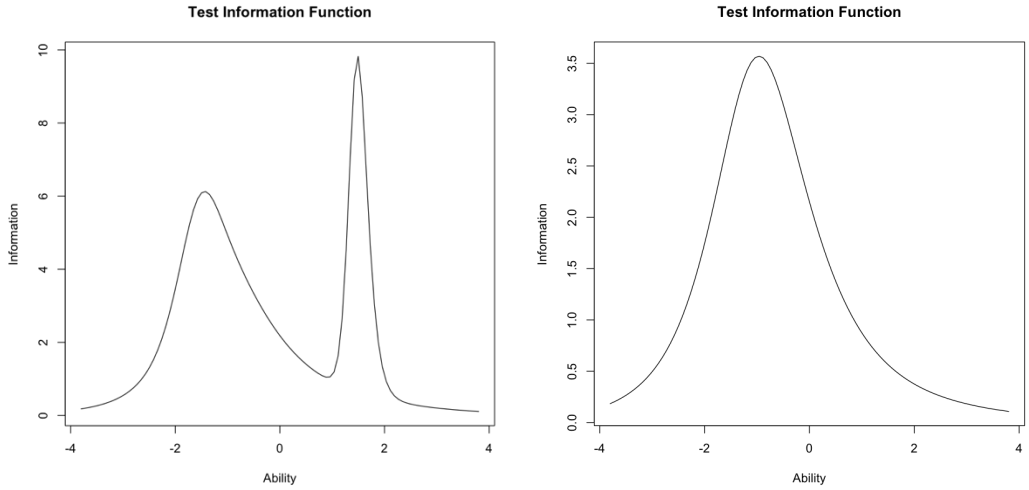


Figure 3. Test information curves (TIFs): a) before removing items 3 and 9, b) after removing items 3 and 9

5.6.1 Reliability analysis

As demonstrated in Figure 4, the scale demonstrated the highest reliability for individuals with ability levels around θ = -1. There was almost no reliable information about below −2.5 and about above 2.00, and the standard error increased quickly for both smaller and larger θ values. The marginal reliability for the scale was approximately 0.62.

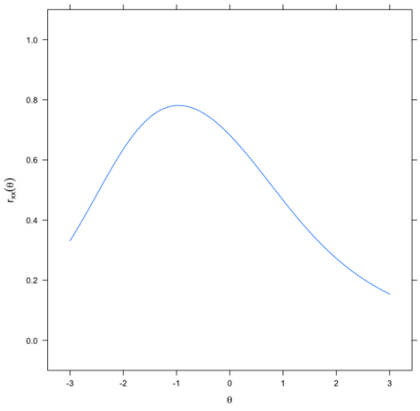


Figure 4. Reliability analysis of the scale

6. Discussion and Conclusion

Considering the demand for teacher's digital pedagogy competence, especially after the rapid digitization in education due to COVID-19 lockdowns, it has become critical to define, measure and build teacher's digital pedagogy competence. A literature review of the tools available to assess

teacher's digital pedagogy competence points at the need for a tool that does not rely on self-reporting to evaluate their digital pedagogy competence as data obtained from self-reports can be often unreliable. The Digital Pedagogy Competency Scale (DiPiCoS) has been developed to address this need and contains scenario-based items which are situated in real life learning and teaching contexts.

An item response analysis for the scale was conducted with responses from 1315 English-speaking teachers from various parts of India. The scale demonstrated unidimensionality and most of its constituent items showed acceptable levels of discrimination, difficulty and guessing parameters in an item response analysis constructed using a 3-PL model.

Results of the item response analysis showed that item 3, which was developed to assess whether or not a teacher practices inclusion in online classes, was answered correctly by most respondents. Through a retrospective exercise, researchers concluded that responses on this item were significantly affected by the bias of social desirability. On the other hand, item 9, which pertained to the integration of technology in pedagogical practice, had ambiguity in the response options provided, due to which many of them could possibly be "correct." Items 3 and 9 were removed from the 10-item scale and an 8-item scale was finalized. This scale demonstrated good reliability around ability levels $\theta = -1$ and a marginal reliability value of approximately 0.62.

The validated DiPeCoS can be used as pre and post assessments for teacher training programs on digital pedagogy. The tool can also be used for screening purposes to identify educators with high and low digital pedagogy scores. With multiple teaching-learning scenarios being captured in the tool, the tool can be used with educators across K-12 to higher education institutions, across domains in any setting.

As future lines of research, external measures, which were not included in this study to save time and avoid respondent fatigue (Morgado et al., 2017), could be used to establish external discriminant and convergent validity of the scale. It might also be useful to examine some significant psychometric properties of the scale, such as test-retest reliability. Also, although assessment of responder's decision making with regard to pedagogical practices is a better measure of their digital pedagogical competence when compared with self-report, it still does not predict translation of these choices into real life practice.

Finally, the methods, strategies and goal of digital pedagogy continues to evolve with the emergence of new digital technologies and their affordances. Like any other tool designed to measure digital competence, this tool needs to be updated regularly to reflect new evidence and insights in the field of learning sciences that informs pedagogy as well as new advances in education technologies and their opportunities and constraints.

References

Andrade, H. L. (2019). A Critical Review of Research on Student Self-Assessment. In Frontiers in Education (Vol. 4). Frontiers Media SA. https://doi.org/10.3389/feduc.2019.0008

Al-Azawei, A., Serenelli, F., & Lundqvist, K. (2016). Universal Design for Learning (UDL): A content analysis of peer reviewed journals from 2012 to 2015. Journal of the Scholarship of Teaching and Learning, 16(3), 39-56.

Baker, F. B. (2001). The Basics of Item Response Theory. College Park: ERIC Clearinghouse on Assessment and Evaluation.

Banister, S., & Reinhart, R.V. (2013). Assessing NETS•T Performance in Teacher Candidates: Exploring the Wayfind Teacher Assessment, 29(2), Journal of Digital Learning in Teacher Education.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Berman, Elena. 2003. "A Short Guide to Evaluating Teaching." Tucson, AZ: University of Arizona. Accessed October 4, 2016. http://www.tlataskforce.uconn.edu/docs/ resources/A_Short_Guide_to_Evaluating_Teaching.pdf.

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. Sociological Methods & Research, 21(2), 230–258. https://doi.org/10.1177/0049124192021002005

Cattell, R. B. (1966). The Scree Test For The Number Of Factors. Multivariate Behavioral Research, 1(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Center for Applied Special Technology, CAST (2015), https://www.cast.org/impact/universal-design-for-learning-udl

Diamond, J., & Evans, W. (1973). The Correction for Guessing. In Review of Educational Research (Vol. 43, Issue 2, pp. 181–191). American Educational Research Association (AERA). https://doi.org/10.3102/00346543043002181

Fernández-Batanero, J.M.;Román-Graván, P.; Montenegro-Rueda, M.; López-Meneses, E.;Fernández-Cerero, J. (2021). Digital Teaching Competence in Higher Education: A Systematic Review. Educ. Sci.

Fink, L. Dee. 2003. Creating Significant Learning Experiences: An Integrated Approach to Designing College Courses, Higher and Adult Education Series. San Francisco, CA: Jossey-Bass

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika 30, 179–185. https://doi.org/10.1007/BF02289447

Kaiser, H. F. (1960). The application of electronic computers to Factor Analysis. Educational and Psychological Measurement, 20(1), 141–151. https://doi.org/10.1177/001316446002000116
Kindley, R.W.(2002). Scenario-based e-learning: A step beyond traditional e- learning. ASDT
Kivunja, C. (2013). Embedding digital pedagogy in pre-service higher education to better prepare teachers for the digital generation. International Journal of Higher Education, 2(4), 131–142.
Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation.* Cambridge University Press. https://doi.org/10.1017/CBO9780511815355
Lázaro-Cantabrana, J. L., Usart-Rodríguez, M., & Gisbert-Cervera, M. (2019). Assessing Teacher Digital Competence: the Construction of an Instrument for Measuring the Knowledge of Pre-Service Teachers. In Journal of New Approaches in Educational Research (Vol. 8, Issue 1, pp. 73–78). University of Alicante. https://doi.org/10.7821/naer.2019.1.370
Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First- and higher order factor models and their invariance across groups. Psychological Bulletin, 97(3), 562–582. https://doi.org/10.1037/0033-2909.97.3.562
McCabe, Kimberly A., and Leslie S. Layne. 2012. "The Role of Student Evaluations in Tenure and Promotion: Is Effective Teaching Really Being Measured?" The Department Chair: A Resource for Academic Administrators 22 (3):17–20
Mishra, P., & Koehler, M. J. (2006). Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge. Teachers College Record, 108(6), 1017–1054. http://one2oneheights.pbworks.com/f/MISHRA_PUNYA.pdf
Montebello, M. (2017). Digital pedagogies for teachers' CPD [Conference presentation]. IADIS International Conference on Educational Technologies, Sydney, Australia.
Morata-Ramírez, M. de, & Holgado-Tello, F. P. (2013). Construct validity of Likert Scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. International Journal of Social Science Studies, 1(1). https://doi.org/10.11114/ijsss.v1i1.27
Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: ten main limitations and recommendations to improve future research practices. Psicologia: Reflexão e Crítica, 30(1). Springer Science and Business Media LLC. https://doi.org/10.1186/s41155-016-0057-1
Oishi, S. (2007). The application of structural equation modeling and item response theory to cross-cultural positive psychology research, in Series in Positive Psychology. Oxford Handbook of Methods In Positive Psychology, eds A. D. Ong and M. H. M. van Dulmen, (New York, NY: Oxford University Press), 126–138.
Ok, M. W., Rao, K., Bryant, B. R., & McDougall, D. (2017). Universal design for learning in pre-k to grade 12 classrooms: A systematic review of research. Exceptionality, 25(2), 116-138.
Rao, K., Torres, C., & Smith, S. J. (2021). Digital tools and UDL-based instructional strategies to support students with disabilities online. Journal of Special Education Technology, 36(2), 105-112.
Redecker, C., & Punie, Y. (2017). European Framework for the Digital Competence of Educators: DigCompEdu. Luxembourg: Publications Office of the European Union.
Sailer, M., Murböck, J., & Fischer, F. (2021). Digital learning in schools: What does it take beyond digital technology? In Teaching and Teacher Education (Vol. 103, p. 103346). Elsevier BV. https://doi.org/10.1016/j.tate.2021.103346
Sailin, S.; Mahmor, N. (2018), Improving Student Teachers' Digital Pedagogy through Meaningful Learning Activities. Malaysian Journal of Learning and Instruction, v15 n2 p143-173 Dec 2018
Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye, & C. C. Clogg (Eds.), Latent variables analysis: Applications for developmental research (pp. 399-419). Thousand Oaks, CA: Sage
Sattelmayer, M., Hilfiker, R., Luomajoki, H., and Elsig, S. (2017). Use of Rasch analysis to investigate structural validity of a set of movement control tests for the neck. Musculoskelet. Sci. Pract. 27, 131–136. doi: 10.1016/j.math.2016.07.006
Schmidt, D. A., Baran, E., Thompson A. D., Koehler, M. J., Mishra, P. & Shin, T. (2009).Technological pedagogical content knowledge (tpack): The development and validation of an assessment instrument for preservice teachers. *Journal of Research on Technology in Education, 42*(2), 123-149. doi: 10.1080/15391523.2009.10782544
Smith, F. G. (2012). Analyzing a college course that adheres to the Universal Design for Learning (UDL) framework. Journal of the Scholarship of Teaching and Learning, 31-61.
Thalheimer, W. (2018). The learning-transfer evaluation model: Sending messages to enable learning effectiveness. Available at https://WorkLearning.com/Catalog This report is aligned with LTEM Version 11.
Tourón, J., Martín, D., Navarro Asencio, E., Pradas, S., & Íñigo, V. (2018). Validación De Constructo De Un Instrumento Para Medir La Competencia Digital Docente De Los Profesores (Cdd). In Revista Española De Pedagogía (Vol. 76, Issue 269). Funciva Ediciones Sl. Https://Doi.Org/10.22550/Rep76-1-2018-02
UNICEF. (2020). Putting the 'learning' back in remote learning. https://www.unicef.org/globalinsight/reports/putting-learning-back-remote-learning
University of Waikato. (2018). Universal Design for Learning Self-Assessment. https://www.waikato.ac.nz/__data/assets/pdf_file/0013/501007/for-web-UOW-UDL-self-assessment-tool-draft-printable-v3.1.pdf
van de Mortel, TF 2008, 'Faking it: social desirability response bias in self-report research', Australian Journal of Advanced Nursing, vol. 25, no. 4, pp. 40-48.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

How to Cluster Students Based on Their Digital Competence for Learning?

Külli KALLAS[a]*, Margus PEDASTE[a]
[a]*Institute of Education, University of Tartu, Estonia*
*kulli.kallas@ut.ee

Abstract: Finding the most relevant and effective teaching and learning methods is one of the core challenges in education. As the educational community is valuing personalized learning, student clustering for better outcomes is relevant. Therefore, the aim of the current study is to identify student clusters according to their level of digital competence for learning. 268 students (grades 3–8) from four schools in Estonia participated in the study and completed a test for assessing their digital competence for learning in ten dimensions. Four student clusters were identified using hierarchical cluster analysis: programmers, creators, casual users and beginners. The clusters were determined mainly by operational and programming skills. As all the students had rather positive attitudes towards using digital devices for learning, the levels of skills and knowledge ranged from very low to high. The results might be useful in designing long-term studies for monitoring the development of digital competence for learning throughout primary and lower secondary school. Furthermore, it is evidently clear from the findings that students do need tailored support in order to improve their digital competence for learning.

Keywords: digital competence for learning, cluster analysis, primary education, secondary education

1. Introduction

The importance of using digital devices, environments and materials in learning cannot be overrated. As the society moves towards digitalization, being an active member of it means having skills and knowledge to use digital devices. Therefore, schools have an important responsibility of making sure that children are ready to face the challenges of our more-and-more digital world. Furthermore, due to the COVID-19 outbreak, which started in 2020, students' ability to use digital devices for learning is now more relevant than ever. As we demanded more from our students than mere information seeking and document writing, students' skills were really put to test. Many students found the lockdown challenging due to their low level of digital skills (Carretero Gómez, 2021). Moreover, extra attention had to be paid to following the rules and communicating in the digital classroom. At the same time, as students were alone with their devices, they needed more knowledge regarding how to protect themselves in the digital world.

The digital skills needed for being an active society member are described in the Digital Competence Framework for Citizens (Vuorikari, Kluzer & Punie, 2022), which is describing five competence areas: information and data literacy, communication and collaboration, digital content creation, safety and problem solving. Digital competence itself is a set of knowledge, skills and attitudes that are required when using information and communication technology to perform tasks (Ferrari, 2012). Since the DigComp framework is describing the competence for citizens, it is a reasonable guideline for schools, because the role of schools is to help children to become active citizens.

Based on the DigComp framework, Estonian data from the EU Kids Online study and the Estonian Digital Competence test results, Pedaste, Kalmus and Vainonen (2021) defined the dimensions of the digital competence for learning, which serves as the basis for the digital competence assessment test created and validated by Kallas and Pedaste (2022). However, while we now could obtain the knowledge of students' skill level on different skills, attitudes and knowledge related to the digital competence for learning, clear guidelines on how to teach children based on this knowledge still remain a necessity for practitioners. As in education we steer towards teaching students as individuals (person-oriented approach, where individuals are seen as an organized whole; Bergman & Lundh, 2015), it would be ideal to personalize learning paths for each learner. However, as education often suffers from

teacher-student ratios which are far from optimal and teaching each student completely separately is nearly impossible in practice, classifying students based on their abilities could be the solution for more personalized learning. Classifying students' profiles through clustering gives us information about how the student groups are formed in light of the digital competence for learning and therefore provides practitioners with the option of also creating interventions for further digital competence development. Therefore, the goal of the current study is to determine the students' profiles based on their digital competence for learning.

2. Previous research on student profiling

It has previously been observed that students do differ according to their abilities and clustering is helpful in determining various student groups based on their traits. As the use of technology in schools and among young people in general is widespread, there are attempts to cluster them according to the use of technology in general and also more specifically in learning. Van den Beemt, Akkerman and Simons (2011) were looking into young people's (aged 10–23) media activities with the goal of obtaining valuable information about using media in learning. As a result, they proposed distinguishing four student profiles according to their type and frequency of media usage: *traditionalists*, *gamers*, *networkers* and *producers*. Pedaste and colleagues (2017) were clustering students who were using mobile devices for the purposes of learning science and mathematics. The authors concluded that five general groups and eleven subgroups could be distinguished according to the nature of their use of mobile devices. The two main clusters were *users* (further categorized as *digital natives* (use technology for different activities in learning), *communicating information students* (use social media in learning more than others), *information students* (share actively links to materials relevant in learning), *creation students* (often develop and adapt learning materials)) and *non-users*. Scherer, Rohatgi and Hatlevik (2017) analyzed the Norwegian ICILS 2013 dataset and found evidence of two profiles that described students' use of information and communication technologies (ICT) at school and outside of it. The profiles differed on the extent to which the students used ICT outside of school and for leisure-related activities – less consistent users with low frequency and advanced users, whose frequency and variety of ICT use was high.

These studies illustrate that attempts have been made to find connections between students' technology usage and skills, and behind it is a need for proposing more effective learning and teaching strategies. However, previous studies also indicate the need for more research in the area, as there is still a lack of knowledge regarding what kind of support students need for improving their digital competence for learning. In addition, this need might have significantly changed during the COVID-19 pandemic, which initiated rapid developments in education for using digital technologies.

3. Methods

3.1 Data collection

The data were collected in September 2021 with the digital competence assessment test called Digitest, which can be used to measure primary (from grade three) and lower secondary school children's digital competence in ten dimensions. The test (see https://survey.ut.ee/index.php/697392?lang=en) has been validated in Estonia by Kallas and Pedaste (2022). The ten dimensions of digital competence are assessed with 37 test items: social aspects (value of peers' use of digital tools; motivation to do what others expect; perception of others' support when using digital tools; 2 items), perceived control (students could decide whether to use digital devices; believing in one's own ability to use digital devices; 5 items), behavioral attitudes (anxiety towards using digital tools; belief in the value and simplicity of digital tools; 5 items), behavioral intention (students prefer digital devices or environments to other ways of doing things; 3 items), creation of digital materials/content (knowledge and skills required to create digital texts or visual materials; 4 items), programming digital content (knowledge and skills required for programming; 3 items), communicating in the digital world (knowledge and skills required for secure online communication in the digital world; 3 items), performing operations with digital tools (technical knowledge and skills acquired by using digital devices to solve problems; 4 items), protecting themselves and others in the digital world (the respondent's ability to protect themselves and others from the threats in the digital world; 4 items) and following the rules in the digital

world (the law-abiding nature of the respondent in the digital world; 4 items). The test was assessing skills and knowledge with different types of multiple-choice and open-ended questions. In the case of some tasks students had to perform practical assignments. Students' attitudes were assessed with statements on a six-point Likert type agreement scale.

The test was completed in an electronic environment where there was no option of moving back to the previous question. However, it was possible to skip questions which respondents perceived to be too difficult. There was no time limit for completing the test, but as the test was taken during the students' regular classes, the length of the lesson (45 minutes) was usually taken as a time frame for completion. The average time of completing the test was 46 minutes (SD = 11.8, range from 13 to 77). Taking the test was voluntary, students did not get a grade or any other benefit. The schools got feedback about their students' digital competence at class level.

3.2 Participants

The test was completed by 426 students from four Estonian schools. The age of the students was 9–15 (16 students from the 3rd grade, 112 from the 4th grade, 56 from the 5th grade, 87 from the 6th grade and 155 from the 8th grade). Of the students, 224 were males and 202 were females. As the items about attitudes towards digital devices and environments for learning were at the end of the test, the data about attitudes was incomplete. In order to have a more trustworthy analysis of student clusters, the students with missing data about attitudes were removed from the analysis. The final sample consisted of 268 students (137 males and 131 females).

3.3 Data analysis

In order to determine which student clusters could be distinguished based on their level of digital competence, hierarchical cluster analysis was carried out using the Ward method. A dendrogram was created for selecting a meaningful number of clusters. Cluster membership was saved and ANOVA analysis was used to find the differences between the clusters. Eta squared was calculated for determining the most weighty dimensions of each cluster.

4. Results

From hierarchical cluster analysis, four meaningful student clusters concerning digital competence for learning could be established (Figure 1). All the students have positive or average attitudes towards using digital devices and environments for learning. The skills and knowledge give a more varied picture, as high-, medium- and low-performing students could be distinguished between. The differences between the clusters are mainly in the dimensions of skills and knowledge (Table 1). All the differences are significant, except for the attitude dimension of social acceptance. Students' membership in the cluster is determined mainly by the level of their operational and programming skills.

High-performing and highly motivated students are profiled as *programmers*. The programmers have higher skills and knowledge levels in all the dimensions. What distinguishes them from all the other students is mainly their knowledge and skills in the programming. The second cluster is called *creators*. Creators exhibit high knowledge of creating digital materials and, at the same time, above-average operational and communication skills. Both programmers and creators have more positive attitudes than the rest of the students. The third cluster is labelled *casual users*, who have a low level of operational skills but an average level of knowledge and skills of creating materials and communication in the digital world. Casual users exhibit an average level of attitudes towards using digital devices in learning. The fourth cluster is called *beginners* due to their low level of operational skills and weaker performance in all the other dimensions describing skills and knowledge. Beginners have an average attitude level.

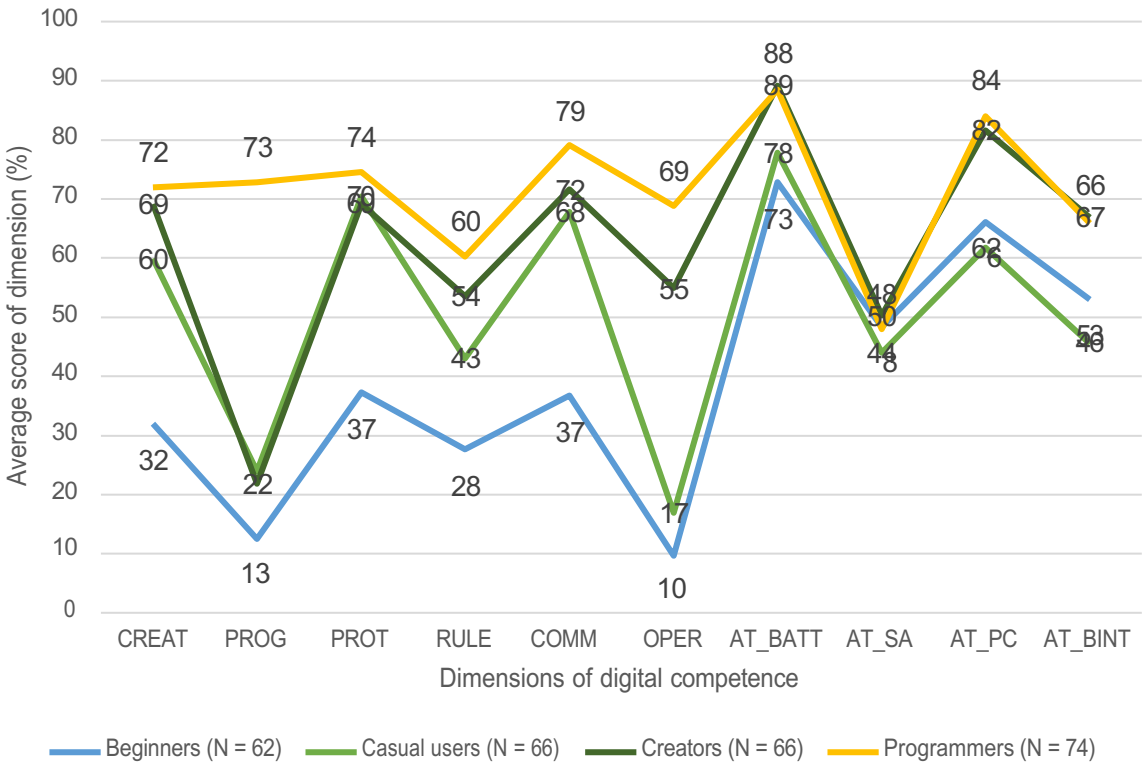


Figure 1. Student clusters based on their level of digital competence (CREAT – creating digital materials, PROG – programming digital content, PROT – protecting oneself and others in the digital world, RULE – following the rules in the digital world, COMM – communicating in the digital world, OPER – performing operations with digital tools, AT_BATT – behavioral attitudes, AT_PC – perceived control, AT_SA – social aspects, AT_BINT – behavioral intention)

Table 1. *ANOVA results comparing the dimensions of digital competence for learning across identified clusters and their importance in defining the student clusters*

Dimension of digital competence for learning	F	p	ε^2
CREAT	73.187	<.01	0.454
PROG	155.649	<.01	0.639
PROT	63.767	<.01	0.420
RULE	51.838	<.01	0.371
COMM	69.511	<.01	0.441
OPER	177.550	<.01	0.669
AT_BATT	14.870	<.01	0.145
AT_SA	0.973	>.05	0.011
AT_PC	26.288	<.01	0.230
AT_BINT	19.261	<.01	0.180

5. Discussion

This study aimed to identify user profiles, i.e. clusters of students based on their digital competence for learning. Four clusters were identified using hierarchical cluster analysis: programmers, creators, casual users and beginners. The programmers exhibit high scores in all the dimensions of digital competence, meaning that they have a sound set of skills for learning with digital devices in different learning environments. The biggest difference between programmers and creators lies in their programming skills. Casual users have mostly average scores in the dimensions of digital competence. What sets them apart from creators is their low level of operational skills and slightly lower level of attitudes. Casual users have the knowledge of creating materials, protecting their and other people's creative work and following the rules in the digital world, but shortcomings in operational skills may inhibit them from

learning as efficiently as they could. The lowest-performing students can be labelled beginners. They have moderate attitudes towards using digital devices and environments for learning; however, they seem to have less skills and knowledge to actually perform well.

Our findings bear some similarities with the previous studies. Scherer et al. (2017) observed two very distinguishable clusters of advanced ICT users and beginners, and these clusters are evident in the current study as well. Scherer et al. (2017) suggested the clustering according to frequency of technology use, which could also be a factor in this study. However, our study sheds a bit more light on the qualitative differences in the clusters. Similar to Pedaste et al. (2017), we also found more clusters which slightly differ from one another by distinctive factors. However, Pedaste's study was conducted in a narrower context of using only mobile devices for learning science and mathematics.

Approximately a quarter of the respondents (27%) in the current study belong to the programmers' cluster. These findings are somewhat surprising given the fact that other research shows that the number of students who have a well-developed set of digital skills is usually lower. In a study by Van den Beemt et al. (2010), the *creator* cluster representatives had more skills than others and only 6% of all the participants belonged to this group. Pedaste et al. (2017) also determined that only 5% of students in their study belonged to the *net generation* (also called *digital natives*) group. These somewhat different results compared to study by Pedaste and others (2017) could be explained by the latter study's higher number of clusters, which could mean that the *net generation* cluster only includes the highest-achieving students. Van den Beemt and colleagues (2010) were looking more into media production, which is a somewhat narrower skill than creating digital materials and different from programming, which could be the reason why the percentages of highest-achievers are not comparable. However, the changes might also be explained by the effect of the COVID-19 pandemic. For example, Rannastu-Avalos and Siiman (2020) indicated in their study conducted just five weeks into the pandemic that one of the teachers' challenges was establishing cognitive, social and teaching presence with young teenagers. Now, more than two years later the teachers might have found ways to support that and it might have had an effect on the learners. This explanation is also supported by Lepp, Aaviku, Leijen, Pedaste and Saks (2021), who found that during the COVID-19 pandemic teachers started to focus more on maintaining students' social interaction and supporting student motivation, valuing learners' well-being, socialization and motivation over subject matter competences. Digital competence might also have been viewed as one of the general competences, receiving more attention in the new situation. The study by Gerard, Wiley, Debarger, Bichler, Bradford and Linn (2021) even demonstrated how teachers had developed several strategies to support students' self-regulated learning in online settings in the context of the COVID-19 pandemic. Therefore, it is expected that students' digital competence has improved more than before the pandemic-related school closures when distance education was applied at a very limited level.

The current study brought forward differences between the student clusters, which could help us to create specifically tailored learning paths and interventions to enhance children's digital competence for learning. For example, programmers could benefit from more complex learning tasks, as their skills and knowledge related to the digital competence for learning are on a higher level. Creators could be further supported in completing more complex operational tasks, and they could benefit from tasks designed for learning programming as a way to create content. Casual users would profit from further support in improving their operational skills for learning, as they struggled with tasks which are quite basic for learning. Like creators, casual users would also benefit from knowing the basic programming concepts. Furthermore, as casual users have the lowest overall attitude scores, teacher initiatives for improving their attitudes towards digital devices for learning may prove beneficial. Beginners could be supported in becoming programmers, as the programmers have the best overall results in the test. Therefore, beginners would profit from learning more technical skills (operational and programming), since these are the areas they are evidently struggling in.

With regard to the research methods, some limitations of the study need to be acknowledged. As the study was based on a limited number of respondents only in Estonia, the number of identified clusters may either increase or decrease with a larger international sample. Furthermore, as there was an open call to participate in the study, which took place in a time where COVID-19 had an impact on the functioning of schools, only the most motivated and digitally active schools might have responded and therefore, the picture of the clusters might be somewhat distorted. In conclusion, this study could be repeated with a larger sample both in Estonia and internationally in order to get more accurate results.

6. Conclusions

This study showed that different student profiles, i.e. clusters determined by their level of digital competence exist, thus pointing out the need for personalizing learning with digital devices and environments. Moreover, although young people are considered to be more digitally competent, we still need to teach them the digital skills for learning purposefully, as the clusters show that there are significant differences in the levels of some skills. The findings of the current study point to the need for further research on how to support different student clusters in order for them to have higher-level digital competences for learning; also, the findings could provide a basis for understanding student learning throughout primary and lower secondary school. We also believe that having the knowledge of different student clusters based on their skill levels offers the teachers more information about designing their learning paths, and, for researchers, it is a great basis for intervention development.

7. Acknowledgments

We would like to thank all schools, teachers and students who have taken the time to participate in our study. This study has been supported by the DIGIVARA5 project "The Effect of Using Digital Learning Materials for Learning and Teaching in the Context of Estonian Basic School (1.05.2020 − 30.04.2023)" financed by the Ministry of Education and Research in Estonia.

References

Bergman, L. R., & Lundh, L.-G. (2015). Introduction: The Person-oriented approach: Roots and roads to the future. *Journal for Person-Oriented Research, 1(1–2),* 1–6. https://doi.org/10.17505/jpor.2015.01

Carretero Gómez, S. (2021). Skills for Life: Digital Skills. Inter-American Development Bank. https://doi.org/10.18235/0003126

Ferrari, A. (2012). Digital competence in practice: An analysis of frameworks. Publications Office. https://data.europa.eu/doi/10.2791/82116

Gerard, L., Wiley, K., Debarger, A. H., Bichler, S., Bradford, A., & Linn, M. C. (2021). Self-directed Science Learning During COVID-19 and Beyond. *Journal of Science Education and Technology*, 1–14.

Kallas, K., & Pedaste, M. (2022). How to Improve the Digital Competence for E-Learning? *Applied Sciences*, *12*(13), 6582.

Lepp, L., Aaviku, T., Leijen, Ä., Pedaste, M., & Saks, K. (2021). Teaching during COVID-19: The decisions made in teaching. *Education Sciences*, *11*(2), 47.

Pedaste, M., Kalmus, V., & Vainonen, K. (2021). Dimensions of digital competence and its assessment in basic school. *Eesti Haridusteaduste Ajakiri. [Estonian Journal of Education]*, 9(2), 212–243. https://doi.org/10.12697/eha.2021.9.2.09

Pedaste, M., Must, O., Leijen, Ä., Mäeots, M., Siiman, L., Kori, K., & Adov, L. (2017). Nutiseadmete kasutamise profiilid loodusainete ja matemaatika õppimise kontekstis. [Profiles of students who use mobile devices for the purposes of learning science and mathematics.] *Eesti Haridusteaduste Ajakiri. [Estonian Journal of Education], 5(1)*, 99–129. https://doi.org/10.12697/eha.2017.5.1.04

Rannastu-Avalos, M., & Siiman, L. A. (2020, September). Challenges for distance learning and online collaboration in the time of COVID-19: Interviews with science teachers. *In International Conference on Collaboration Technologies and Social Computing* (pp. 128–142). Springer, Cham.

Scherer, R., Rohatgi, A., & Hatlevik, O. E. (2017). Students' profiles of ICT use: Identification, determinants, and relations to achievement in a computer and information literacy test. *Computers in Human Behavior, 70*, 486–499. https://doi.org/10.1016/j.chb.2017.01.034

Van den Beemt, A., Akkerman, S., & Simons, P. R. J. (2011). Patterns of interactive media use among contemporary youth. *Journal of Computer Assisted Learning, 27(2),* 103–118. https://doi.org/10.1111/j.1365-2729.2010.00384.x

Vuorikari, R., Kluzer, S. and Punie, Y. (2022). DigComp 2.2: The Digital Competence Framework for Citizens. Publications Office of the European Union. doi:10.2760/115376.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Copyright in the Scope of Education

Branko KIRIN[a*] & Rwitajit MAJUMDAR[b]
[a]*KIRIN Attorney-at-Law Office, Croatia*
[b]*Academic Center for Computing and Media Studies, Kyoto University*
*branko.kirin@gmail.com

Abstract: Copyright is a legal term used to describe the rights that creators have over their works. Copyright grants the author a set of exclusive rights of protecting his personality in connection with his work (moral rights) and of exploiting the work (economic rights). In the context of education, copyright is important as copyrighted works used in course materials or lecturing, are often and common. When reproducing or distributing materials such as literature, publications, videos or music, lecturers need to understand the copyright of any material that is used. With this paper, we shall present the scope of copyright and its exceptions for the purpose of education. First, we shall explain the meaning and exceptions of copyright within the international legal framework agreed by the Bern Convention that set the standard of copyright and its exceptions. Further, we shall explain exceptions that affect teaching-related uses of copyrighted work and describe how respective issue is regulated within the legislature of the European Union.

Keywords: Copyright, e-education, Berne Convention

1. Introduction

Copyright is a form of intellectual property law that protects original works of authorship, as soon as an author fixes the work in a tangible form of expression, including literary, dramatic, musical, and artistic works, such as poetry, novels, movies, songs, computer software, and architecture. Copyright does not protect facts, ideas, systems, or methods of operation, although it may protect the way these are expressed. One may express ideas in writing or drawings and claim copyright in its description, but the copyright will not protect the idea itself as revealed in your written or artistic work. Copyright protection extends only to expressions, and not to ideas, procedures, methods of operation or mathematical concepts as such. By its nature, copyright belongs to the person who creates the copyrighted work, simply known as the author.

In general, copyright (or author's right) is a legal term used to describe the rights that creators have over their works. Copyright grants the author a set of exclusive rights of protecting his personality in connection with his work (moral rights) and of exploiting the work (economic rights). Moral rights would in principle be the right to claim authorship of the work and the right to object to any mutilation, deformation, or other modification of, or other derogatory action in relation to, the work that would be prejudicial to the author's honor or reputation. Moral rights consist of and include, inter alia the right of first publication, the right of recognition and respect of the authorship, and the right to protect the reputation of the authorship and the author itself. On the other hand, economic rights consist of the right to reproduce, the right to distribute, the right to inform the public of the work and the right to process and transform it. A virtue of the copyright is the ability that all mentioned rights come into exitance automatically with the expression of the author's work regardless of form. While moral rights could not be separated from the author, economic rights, on the other hand, are transferable under the consent of the author.

The use of copyrighted works in education providing course materials or in lecturing is often and common. When reproducing or distributing materials such as literature, publications, videos or music, lecturers need to understand the copyright of any material that is used.

Although copyright protection is strong and one should generally ask for a license or permission of the author, it is not unlimited, as legislators have constructed number of exceptions and limitations in using authors' work under specific conditions for specific purposes. One of such limitation is using copyrighted work for the purpose of education, both face to face and distance education. E-education revolves around the use of Information and Communication Technologies (ICT) to accelerate the achievement of education goals with the ability to apply ICT skills to access,

analyze, evaluate, integrate, present and communicate information; to create knowledge and new information; to enhance teaching and learning through communication and collaboration by using ICT; and function in knowledge society by using appropriate technology and mastering communication and collaboration skills (Natarajan, 2005).

The goal of this paper is to present the scope of copyright and its exceptions for the purpose of education. First, we shall explain the meaning and exceptions of copyright within the international legal framework agreed by the Bern Convention that set the standard of copyright and its exceptions. Further, we shall explain exceptions that affect teaching-related uses of copyrighted work and describe how respective issue is regulated within the legislature of the European Union.

2. Copyright and its Exceptions

Notably, the UN Universal Declaration of Human Rights in Article 27 (2) recognizes the right of the author: "Everyone has the right to protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."

There are consequently differences between the national laws governing the terms of protection of copyright and related rights as each country has its own legislature creating legal differences and uncertainty, for both authors (right holders) and users.

However, much of the internationally recognized standards in the field of copyright are agreed upon and accepted by the countries that accepted and are contracting parties of the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, July 24, 1971, amended on September 28, 1979, Berne Convention). However, basic principles and legal standards in the field of copyright set by the Berne Convention are not effective until applied within the national legislature of each sovereign state signed by the Berne Convention. On the other hand, the community legal framework of the European Union additionally contributed to the development of respective copyright standards accepted within the EU legislature, which will be further explained in this paper.

The Berne Convention lays down only minimum terms of protection of the rights, namely the life of the author and 50 years after his death, leaving the Contracting States freedom to grant longer terms. Notably, according to the Article 1 (1) of Directive 2006/116/EC: "The rights of an author of a literary or artistic work within the meaning of Article 2 of the Berne Convention, shall run for the life of the author and 70 years after his death, irrespective of the date when the work is lawfully made available to the public".

Some limitations are adopted based on major public interest considerations, some others are founded on the defence of fundamental rights and freedoms; the promotion of education is one of them. The public interest in using copyright protected works is particularly compelling in respect of education, which is not only a public good, but is a prerequisite for other public goods, such as the development of skills necessary for both the economy and the state, and an informed and empowered citizenry (Papadopoulou, 2010). E-Education for its purpose requires learning materials that are copyright protected, thus it is required to seek a license from the author or to rely on existing legislative exceptions and limitations.

Further in the paper relevant copyright exceptions or limitations in using copyrighted works for the purpose of educational activities through e-education are presented, by analyzing the international legal framework (Berne Convention) and relevant provisions of the community legal framework of the European Union.

3. Exceptions That Affect Teaching Related Uses of Copyrighted Work

The Berne Convention, concluded in 1886, was revised in Paris in 1896 and in Berlin in 1908, completed at Berne in 1914, revised in Rome in 1928, in Brussels in 1948, in Stockholm in 1967 and in Paris in 1971, and was amended in 1979 set the standard of copyright and its exceptions (Paris Act, July 24, 1971, amended on September 28, 1979, the Berne Convention).

The Berne Convention allows certain limitations and exceptions on economic rights, that is, cases in which protected works may be used without the authorization of the owner of the copyright, and without payment of compensation. These limitations are commonly referred to as "free uses" of protected works and are outlined in Articles 9(2) (reproduction in certain special cases), Article 10 (quotations and use of works by way of illustration for teaching purposes), Article 10bis (reproduction

of newspaper or similar articles and use of works to report current events) and Article 11bis(3) (ephemeral recordings for broadcasting purposes).

Therefore, Berne Convention provides two exceptions that affect teaching-related uses of copyrighted work.

3.1 Exception for Quotations (Article 10 (1))

Although it does not refer solely to teaching it represents one of two mentioned exceptions for using copyrighted work: "It shall be permissible to make quotations from a work which has already been lawfully made available to the public, provided that their making is compatible with fair practice, and their extent does not exceed that justified by the purpose, including quotations from newspaper articles and periodicals in the form of press summaries" (Article 10 (1) of the Paris Act, July 24, 1971, amended on September 28, 1979, the Berne Convention).

Meaning of "fair practice" is the condition of quotations system. The practice is "fair" when quotations are used in support of the ideas expressed by the person quoting, when quotations are used to illustrate someone's views to criticize someone else's work and generally speaking, when quotations do not merely substitute for the use of the work that has been quoted from (Drier, Drier/Hugenholtz, 2006). This exception is not restricted to certain, prescribed uses, thus can be used for any purpose, as long as they are used within the stipulated provision, as well as for the purpose of education.

3.2 Use of Works for Teaching Purposes (Article 10 (2))

"It shall be a matter for legislation in the countries of the Union, and for special agreements existing or to be concluded between them, to permit the utilization, to the extent justified by the purpose, of literary or artistic works by way of illustration in publications, broadcasts or sound or visual recordings for teaching, provided such utilization is compatible with fair practice" (Article 10 (2) of the Paris Act, July 24, 1971, amended on September 28, 1979, the Berne Convention)).

A relevant exception regulating the use of copyrighted works for teaching purposes is not mandatory for the members of the Berne Convention but could be a matter of national legislature. Therefore, it only sets the limits under which such regulation may be introduced within a specific legislature.

Additional requirements for the use of copyrighted works should be a) "to the extent justified by the purpose" – meaning that the amount of work used should not be more than necessary in order to complete its purpose and b) "compatible with the fair practice" – should not conflict with the normal exploitation of the work for its used purpose and the legitimate interest of the author.

Therefore, respective limitation allows teachers at all levels of education to incorporate selections of copyrighted works as an illustration using e-learning tools, under the condition that the use is compatible with fair practice.

3.3 Indication of Source and Author (Article 10 (3))

"Where use is made of works in accordance with the preceding paragraphs of this Article, mention shall be made of the source, and of the name of the author if it appears thereon" (Article 10 (3) of the Paris Act, July 24, 1971, amended on September 28, 1979, the Berne Convention).

In accordance with Article 10 (3) where the use of copyrighted works is made in accordance with the exception for teaching or quotation, the name of the author and the source should be mentioned.

4. Copyright and its Exceptions Within European Union

4.1 Directive 2001/29/EC on the Harmonization of Certain Aspects of Copyright and Related Rights in the Information Society

Directive 2001/29/EC concerns the legal protection of copyright and related rights in the framework of EU internal market. The provisions of the Article 2 of the Directive 2001/29/EC stipulate that the

Member States shall provide the exclusive right to authorize or prohibit direct or indirect, temporary, or permanent reproduction by any means and in any form, in whole or in part copyrighted work. Regardless of a high level of protection, Article 5 of the Directive 2001/29/EC introduces broad list of optional limitations and exceptions for use of copyrighted works, meaning that Member States are not allowed to introduce any other limitation into their national legislation apart from the listed in the Article 5. Nonetheless, education is one of the stipulated exceptions as in accordance with the Preamble (Recital 14) of the Directive 2001/29/EC: "This Directive should seek to promote learning and culture by protecting works and other subject-matter while permitting exceptions or limitations in the public interest for the purpose of education and teaching" (Recital 14 of the Directive 2001/29/EC).

4.1.1 Exception for the Purposes of Education (Article 5(3)(a)))

"Member States may provide for exceptions or limitations to the rights provided for in Articles 2 and 3 in the following cases: (a) use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved" (Article 5(3)(a))) of the Directive 2001/29/EC)))).

Relevant exception regulating the use of copyrighted works for the sole purpose of illustration for teaching or scientific research is not mandatory, as Member States should implement it within the sole legislature.

On the other hand, respective Article 5(3)(a) does not define category (universities, schools etc.) or nature of educational establishment (private or public, profit or non-profit etc.) that can use the respective exception. Although, Recital 42 of the Directive 2001/29/EC could clarify this by elaborating: "When applying the exception or limitation for non-commercial educational and scientific research purposes, including distance learning, the non-commercial nature of the activity in question should be determined by that activity as such. The organizational structure and the means of funding of the establishment concerned are not the decisive factors in this respect." Therefore, the non-commercial nature of education activity is the mere purpose of respective exception regardless of the organization structure or funding policy of the educational institution itself.

Related to the above, another condition to legally use the respective exception is to properly indicate the source and author's name.

In the end, an exception for the purpose of education is an optional one, so some Member States implemented it, e.g. Croatia (Copyright Act, Official Gazette 111/2021), while other legislatures define it differently.

4.2 Directive 2019/790 of the European Parliament and of the Council on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC society

Directive 2019/790/EU lays down the rules which aim to harmonize further EU law applicable to copyright and related rights in the framework of the internal market, considering digital and cross-border uses of protected content. It also lays down rules on exceptions and limitations to copyright and related rights. Furthermore, it introduces a new term "text and data mining" which refers to any automated analytical technique aimed at analyzing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations. Furthermore, the use of works and other subject matter in digital and cross-border teaching activities is stipulated as well.

4.2.1 Exception or Limitation for Text and Data Mining (Article 4)

Member States shall provide for an exception or limitation to the copyrighted works for reproductions and extractions of lawfully accessible works and other subject matter for the purposes of text and data mining. Such made reproductions and extractions may be retained for as long as is necessary for the purposes of text and data mining.

4.2.2 Use of Works and Other Subject Matter in Digital and Cross-border Teaching Activities (Article 5)

Member States shall provide for an exception or limitation to the copyrighted works to allow the digital use of works and other subject matter for the sole purpose of illustration for teaching, to the extent justified by the non-commercial purpose to be achieved, on condition that such use: a) takes place under the responsibility of an educational establishment, on its premises or at other venues, or through a secure electronic environment accessible only by the educational establishment's pupils or students and teaching staff, b) is accompanied by the indication of the source, including the author's name, unless this turns out to be impossible.

As the respective provisions are mandatory, all Member States should have implemented them and harmonize their legislature with respective provisions by June 7, 2021 (Article 29 of the Directive 2019/790/EU).

5. Conclusions

Copyright is not only legal but also a civilization standard affecting both the author and the user. It protects the author's rights and sets the standards for using copyright-protected work by the user. Giving such protection encourages the author's creativity but should not be unlimited and thus an obstacle for using copyrighted works by others. Copyright protection should remain strong, but legislators have constructed number of exceptions and limitations in using copyrighted works under specific conditions for specific purposes, such as education.

Exceptions for the purposes of education whether regulated within the international legal framework or as a part of the community legal framework of the European Union (only for EU Members applicable) could be applied in the national legislature of each sovereign state. As with any other international legal framework subject, its optional character opens the possibility for differences and diversity in national legislature regulating such matters. Therefore, differences in national legislature result in legal uncertainty. It would be of interest to seek uniform solutions to achieve harmonized legislature in providing a high level of legal certainty. Even more, having in mind constant change and implementation of modern methods for the purpose of improvement of education, legal science should also take advantage of current momentum and adapt to new trends and requirements.

References

Berne Convention for the Protection of Literary and Artistic Works (Paris Act, July 24, 1971, amended on September 28, 1979, Berne Convention)

Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights

Natarajan, M (2005). Implications of Intellectual property Issues in E-Learning. *World Conference on E-Learning in Corporate, Government, Health care and Higher Education, October 2005*

Papadopoulou, M.D. (2010) Copyright Limitations and Exceptions in an E-Education Environment, *European Journal of Law and Technology, Vol.1, Issue 2, 2010*

Drier, T and Hugenholtz, P (2006), Concise European Copyright Law (the Hauge: Kluwer Law International)

Directive 2001/29/EC on the harmonization of certain aspects of copyright and related rights in the information society

Directive 2019/790 of the European Parliament and of the Council on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC society

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

A First-order Action Research Study to Uncover Students' Conceptual Gaps in an Online Statistics Course using Extended Matching Questions

Ishaan TANEJA[a*], Prajish PRASAD[b], Jayakrishnan WARRIEM[a]
[a]*Indian Institute of Technology Madras, India*
[b]*Department of Computer Science, FLAME University, India*
*ishaan.taneja.stats@gmail.com

Abstract: The COVID-19 pandemic has forced institutions worldwide to accelerate their practices on digital transformation, especially in the use of Online Learning. The rapid pace of this transition has forced instructors also to adapt quickly to the online instructional design and in turn reduced opportunities of systematic action research within such a setting. In this paper, we present the action research study by an online instructor to tackle the problem of understanding conceptual gaps among learners. The instructor adapted an existing solution widely used in medical learning, called extended matching items, to design 'Extended Matching Question (EMQ)' for large scale online students (N=1525) of an introductory Statistics course. The instructor modified the structure of this assessment by allowing learners to first identify the concepts involved to solve a particular question and then use those concepts to reach the final numeric answer in the topic 'Permutations and Combinations'. Comparisons of this assessment were made with regular assessments (multiple choice or numerical answers) on the efficacy to make students apply their conceptual understanding for solving the problems. The analysis also enabled the instructor to identify the exact conceptual points where the students lag and showed strong association between identification of concepts and getting the final numeric answer correct. In addition, there was a significant increase in the average score of the test takers in the EMQ assessment compared to that of regular assessment. The study also presents pointers to universities involved in hybrid learning regarding the need for setting up research-practice partnerships within their institutions to tackle their own problems of practice.

Keywords: Action research, Research Practice Partnership, Extended Matching Questions, Statistics, Online Assessment

1. Introduction

The COVID-19 pandemic has forced institutions worldwide to accelerate their practices on digital transformation, especially in the use of Online Learning. Researchers have already pointed out the significant differences between carefully planned online instructional design and a rapid transition to online teaching (Hodges et al., 2020) that results in widespread rejection of online teaching methodologies adopted during the COVID-19 pandemic. Two practical difficulties exist in the systematic design of online instructional processes in such a setting - (i) inability for teacher professional development programmes to match the rapid pace at which institutions are going online (Byrne et.al., 2021), (ii) Limitations of the technology infrastructure in accommodating pedagogical designs. However, this also presents a unique opportunity for researchers to trace the evolution of online instructional practices, especially that of teacher action research in an online setting.

The primary goals of action research studies are to improve the practice with a subordinate goal to produce knowledge that may then be utilized for improving practice (Elliot, 1998). Posch (2019) explained these two goals as two different interests - developmental interest (What would I like t

develop or change?) and research or cognitive interest (What would I like to find out and understand?). The action research studies catering to the former interest are called first-order action research (Hanfstingl & Pflaum, 2022) and will typically involve identification of a problem, suggest possible solutions, implement the solutions and collect information to see effectiveness of their solutions (Altrichter et. al., 2002). The context of large-scale hybrid online learning (similar to Massive Open Online Courses) brings in another complexity to the process as there is a limitation in conducting qualitative studies on specific learning instances across the duration of the course. In this paper, we present a first-order action research study that explains adaptation of an assessment strategy to uncover students' conceptual gaps in an online Statistics course offered at scale.

2. Identifying a problem of practice in the operating context

2.1 Description of the Statistics course in the hybrid online programme

The context of this action research is a large-scale introductory Statistics course which is part of a hybrid Undergraduate degree program in Programming and Data Science at a University in India. The primary mode of teaching-learning within the program is Online (using LMS and Video conferencing services), and there is a mix of online and in-person assessments (conducted simultaneously across various assessment centers across the country). The course is of 12-week duration with weekly online assessments and the in-person assessments occur after the fourth week. The scores in the weekly assessment determine the eligibility of the student to write the in-person assessment. The scores in the in-person assessment contribute to the overall grade of the learners and determine their eligibility to pursue the longer degree programme.

The 12 weeks of the introductory Statistics course is split into three parts: (i) first part (i.e. Week-1 to Week-4) covers the concepts relevant to summarization and visualization of the different types of the data, (ii) second part (i.e. Week-5 to Week-8) of the coursework covers the skill of estimating the chance by applying laws of probability, (iii) the last part (i.e. Week-9 to Week-12) covers the understanding and properties of Random variables along with some probability distributions. The concepts and skills are aimed at helping students proceed to higher courses in Statistics and Machine learning that are required for data scientists.

The primary form of instruction were video lectures which were recorded beforehand by course faculty, and were released every week. The course also had instructors who were responsible for day-to-day orchestration of the course (The first author was an instructor of this course). The primary responsibilities of instructors were to create assessments and interact with students. Interactions with students happened synchronously through live sessions, and asynchronously through discussion forums. These synchronous and asynchronous interactions were useful in addressing students' concerns and difficulties they had with regard to contents of the course.

2.2 Description of the Course Learning Management System

In the course portal, students get access to all the content of coursework in which they are enrolled. The content in the portal is updated as per the release date of weekly contents. The front view of the portal is shown in Figure 1 below.

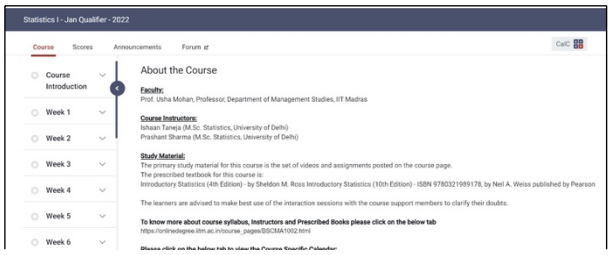


Figure 1. Student view of the portal

By clicking on the Week 'X', the student can access the content of X^{th} week (an example shown in Figure 2), which consists of lecture videos each followed by assessments to test the students' understanding of that specific lecture. At the end of all the video lectures, there are some tutorial videos recorded by the instructors followed by the practice assignments and the graded assignment (formative assessments) which test students' knowledge on the week's content.

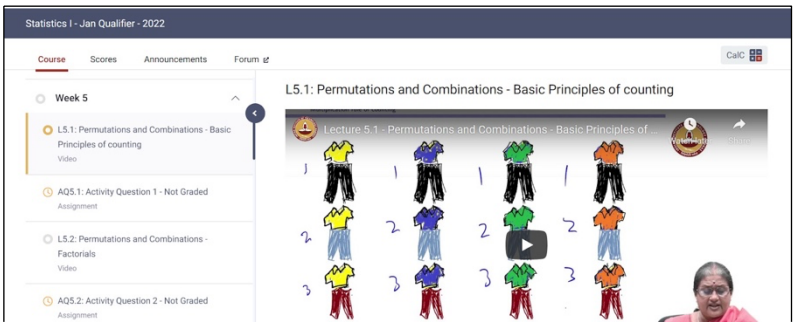


Figure 2. Weekly content view on the portal

2.3 Assessments used in the statistics course

In the Statistics-1 coursework, we test students using a variety of assessments, i.e. Activity questions, Practice assignments, Graded assignments and Quizzes. The 'Activity Questions' are to be attempted after going through each lecture video, in which the students are tested only on the example and conceptual statements used in the lecture video, to make sure that they understood what's been explained in the respective lecture video.

The Practice assignments and Graded assignments contain numerical answer tests (NAT), multiple choice questions (MCQs) and multiple select questions (MSQs). Both Practice and Graded assignments are similar in terms of pattern (approximately 50 % NAT, and remaining 50% MCQ and MSQ) and difficulty. Both of these assessments test the ability of students to apply the concepts taught in the respective week's lecture videos. There's a deadline to submit the Graded assignment, for each week students have a week of time to go through weekly content and solve this assessment. A student can also make multiple submissions to it, but only the last one will be considered for the scoring.

2.4 Limitations in assessments for measuring student learning of statistics concepts

As mentioned in Section 2.3, the type of assessment questions used to test students' learning from the content provided consists of MCQ/MSQ and NAT questions. On examining the assessment data from the LMS, the instructors observed that many students fail to get correct answers for the NAT questions. Further analysis of individual answers revealed that most often students lose out on the marks due to calculation errors. Hence, students might know the concepts and steps required to solve the answer, but NAT type of questions fail to measure their conceptual understanding.

Two of the most effective ways to overcome the drawbacks of such assessments are: a) use questions from Statistics Concept Inventory (SCI), or b) to replace them with subjective assessments. SCI consist of MCQs having options that correspond with common misconceptions that students may have in a particular topic (Stone et. al., 2003). An understanding of the misconceptions present among the students can further help the instructor plan for redress strategies. However, the instructor will not be able to track the exact position where the student develops this misconception by just using SCI as there could be intermediate steps in solving the problem. Similarly, it is not feasible to conduct and evaluate subjective assessments weekly at scale. Hence, we needed a method which can address this gap of conceptual testing of students at scale.\

In order to address this problem, we analyzed existing literature for assessment strategies which test conceptual understanding of students. We found that Extended Matching Questions (EMQ) assessment

type turns out to be one of the potential solutions that not only overcome the limitations of subjective assessments but is also a good measure of conceptual knowledge of the test takers. EMQs have been adopted in multiple contexts like medical and healthcare education, numeracy assessments (Wood, 2003; Cann, 2005). However, we have not seen any reported instances of use of EMQs in teaching-learning of Statistics, especially in a fully online mode of delivery with a diverse learner population (in age and ability). In the next section, we explain the details of this solution, and how we incorporated it into a fully online course setting.

3. Design of Solution - Extended Matching Questions

3.1 Description of Extended Matching Questions (EMQ)

Extended Matching Questions (EMQ), have long been prevalent in medical education, and are similar to multiple choice questions but with one key difference, that they test knowledge in a far more applied, in-depth, sense (Wood, 2003). EMQ consists of a list of more than 8 options along with a few number of questions that will have some of these options as a valid intermediate solution (as a diagnosis or root cause) or final solution. The design of questions and options are roughly similar in structure and content, to test the in-depth knowledge.

A diagrammatic structure for an EMQ is given in Figure 3. In the designing of a list of options for EMQ assessment, one needs to ensure that each question must have at least one best/correct answer from among the series of options given. One option can be an answer to more than one question, or may not even be an answer to any of the questions. This will ensure that the probability of getting the correct answer by chance remains reasonably low. For example, from Figure 3 we see that for Question-1, option-1 is the only correct option in the list of 7 options, which implies that the chances of getting it completely correct by guesswork is approximately 14 percent. Similarly, for Question-2, since three options out of the seven are correct, the total number of ways to answer the question is 35 ($^7C_3$). Hence, chances of getting it completely correct by luck is approximately 3 percent. If this was an MCQ with a list of 4 options, the chances of getting it correct would be approximately 25 percent.

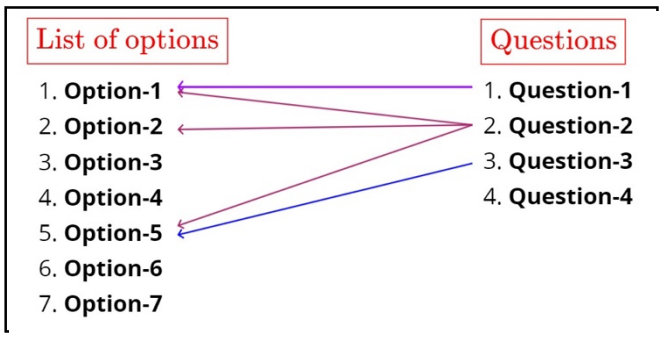


Figure 3. Diagrammatic Structure for EMQ

EMQ has often been used in medical education and other healthcare subject areas for diagnostic reasoning (Case & Swanson, 1993; Wood, 2003; Samuels, 2006; Vuma S, Sa B, 2017). It is evident from research that students attempting this test item format have a greater chance of answering incorrectly if they cannot synthesize and apply their knowledge (Case & Swanson, 1993). Furthermore, it also ensures that the probability of getting the correct answer by chance remains reasonably low. One of the major aims while designing an Extended Matching Question is to condense the list of options, as we don't want a list of 40-50 options for a 10 questions test. Hence, options were designed in such a way that if a student makes a conceptual mistake and gets some answer, say, 'X', then that answer will also be one of the options. Also, that option might be the solution for another question in the list.

We believe that EMQ have several advantages over an MCQ and NAT assessment. In an MCQ assessment with four options, the chances of getting any question correct is 25% by a guess (in an ideal

case). However, in case of an EMQ assessment, this probability reduces significantly. Also, in case of an NAT, there is a good chance that the person taking the exam has good conceptual knowledge in the topic tested but he/she may have got the answer wrong due to some calculation error. There is no way to assess whether the error is due to a lack of conceptual knowledge or due to a calculation error. EMQ assessments can help address this and can indicate to the course instructor which topics a student is lacking in and needs more attention.

3.2 Implementation of EMQ in the online statistics course

The basic understanding of 'Permutations and Combinations' plays an important role in all the upcoming weeks. Therefore, it is necessary to identify as an instructor where the students are lagging and more specifically the exact point of conceptual gap. We implemented EMQs in the Week-5 content of Introductory Statistics course that covered the content - 'Permutations and Combinations'. The idea was to split the graded assignment of the week into two parts. The first part will be the regular assignment (NAT and MCQ) while the second part will be EMQ assessment. Each question in the first part will have a corresponding implementation in the EMQ part. Example questions of regular assessment types are shown below in Figure 4 and Figure 5.

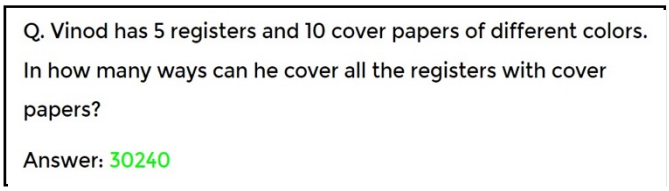


Figure 4. NAT variation of a question

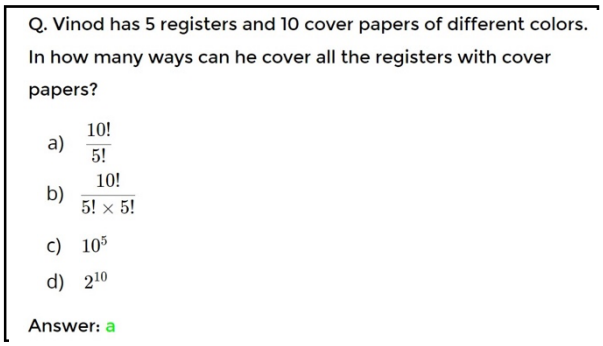


Figure 5. MCQ variation of a question

To ensure that drawbacks of MCQ and NAT are handled, we designed the corresponding EMQ variation as shown below in Figures 6 and 7.

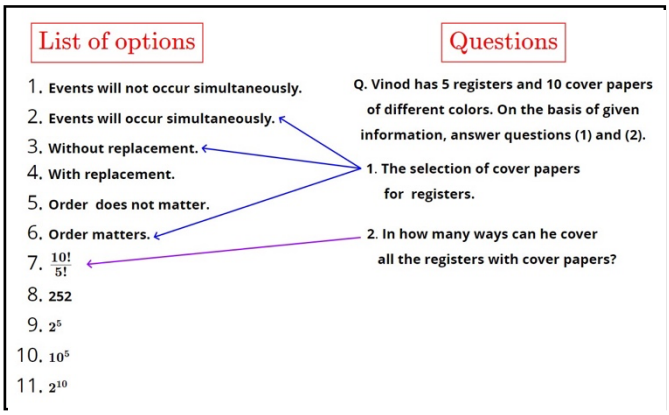


Figure 6. Diagrammatic representation of EMQ variation of a question

Each EMQ question is divided into two subparts, where the first part asks about the conceptual steps required to solve the question, while the second part requires solving and reaching the final solution by applying the conceptual steps selected in the first part of the question. For example, as shown in Figure 7, the first part of questions requires the test takers not to solve the question but just to identify the steps that are necessary to compute the final answer, while the second part requires the computational part to be done. In this way, we can get an idea if students need special attention to this topic, then what to focus more on. Since, the set of questions are from a specific topic, there is a high chance that the conceptual steps used to solve any numeric question relevant to that topic may be the same. As shown in Figure 6, there are six options which are relevant to step identification part of the question from the section 'Permutations and Combinations'. All questions in this assignment will be required to identify steps from the given list of six options only.

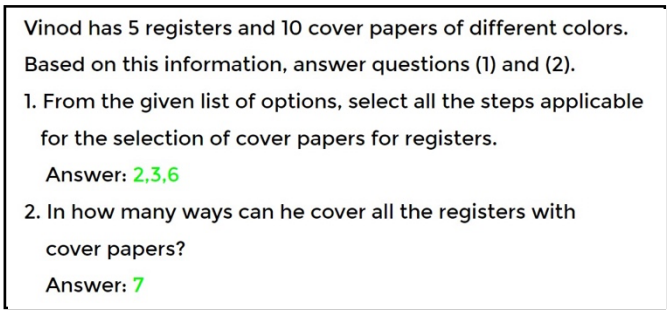


Figure 7. EMQ variation of a question

Before giving such assessments as graded assignments, we also needed to make sure that the test takers are aware of such types of questions and know how to answer them. Hence, we gave them practice EMQ assignments along with the Week-2 (Describing Categorical Data) and Week-4 (Association between two variables) content of the Statistics-1 course. There were also live video conferencing sessions explaining the motive of such assessments and details of the steps to be followed to submit the answers.

3.3 Extending the functionalities of course portal to incorporate EMQ assessments

The course portal was designed to conduct the regular assessments such as MCQ/MSQ, NAT and programming assessments (in input-output testing mode and functional testing mode) only. To enable EMQs, we made use of the functional testing mode of the programming assessments that allowed evaluations by custom scripts written by the course instructors. The functionality allows students to upload their answers to the EMQ via a '.txt' file which is then evaluated using the script provided by the instructors (see Figures 8 and 9). The use of script also allowed the instructors to provide formative feedback to the students.

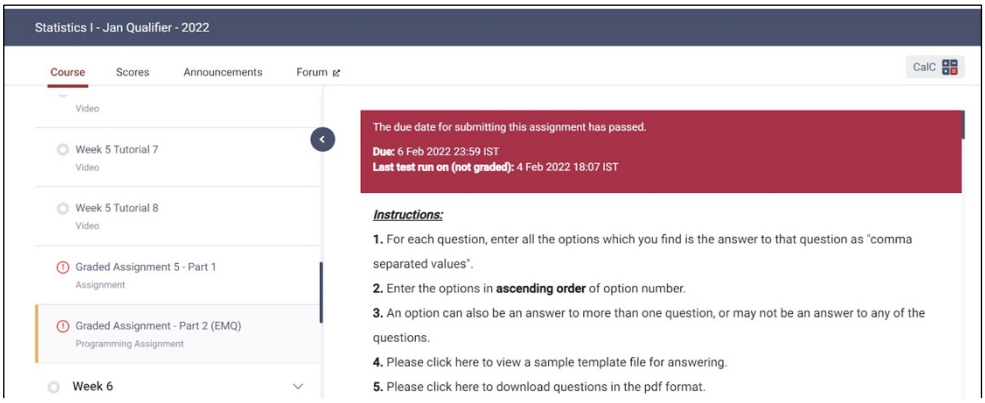


Figure 8. Student view of EMQ assessment on the portal

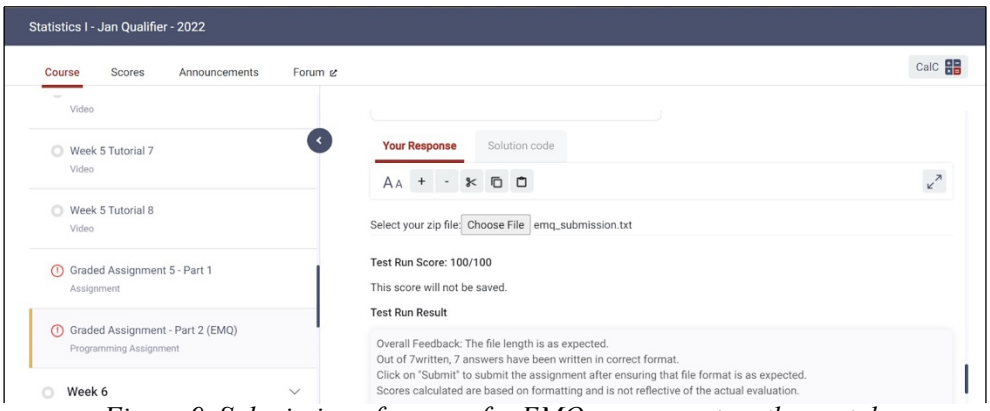


Figure 9. Submission of answer for EMQ assessment on the portal

4. Effectiveness of the EMQ Solution in addressing limitations of previous assessments

4.1 Students' Performance in EMQ Assessments

In the EMQ assessment, we received responses from 1109 students of which 1004 agreed for the use of their inputs and aggregate scores to be used for research purposes. While in regular assessment (MCQ and NAT), 1316 of 1525 students agreed for the same. Firstly, we analyzed the inputs of test takers to step-wise questions in EMQ assessment. Secondly, we considered those students who were able to identify at least one of the steps correctly and analyzed if it affects the percent of correct final answer to the problem, i.e. second sub-part of the question in which the test takers have to identify the final numeric answer to the problem. We also computed the result for regular assessment, of all those students who attempted the question, and tabulated it in Table 1.

Table 1. *Comparison of EMQ and Regular assessment*

Question Number (Type of Question)	EMQ (At least one step correct)	EMQ (All step correct)	% of students correct in NAT/MCQ
Question - 1 (MCQ)	93.2 %	98.4 %	66 %
Question - 2 (MCQ)	48.4 %	78.1 %	53.8%
Question - 3 (NAT)	48.8 %	36.4 %	41.4 %
Question - 4 (NAT)	93.4 %	95.1 %	77.9 %
Question - 5 (NAT)	91.2 %	94.7 %	57.4 %

The first two columns of the table correspond to the EMQ assessment result, and the last column is relevant to the result of regular assessment questions. The first column of the Table represents the percent of students who got the final numeric answer correct given that they identified at least one of the steps correctly, and the second column represents the percent of students who got the final numeric answer correct given that they identified all the steps correctly. While the third column represents the percentage of students who attempted the regular assessment question and got it correct.

From the tabulated data, it can be concluded that if the test takers identify all the steps involved in solving the problem, then the chances of getting the correct final numeric answer increases significantly and even if at least one step is identified correctly the chances of applying conceptual knowledge to get the final answer correct is more than that of regular assessment question. The pattern noticed for Question - 2, is a bit different from others and on further analysis we identified that the test takers struggled in identifying the step if order matters or not, which helps the faculty and instructors to focus more on such similar questions and give students a better understanding of it, i.e. where the test takers are facing problem. Another strange pattern is noticed in Question - 3, where the correct percent to final answer remains consistent irrespective of assessment type and steps identification in EMQ assessment. It is because this question involves the concept from Week-1 content as well, i.e. we connected the topic 'Scale of measurements' with the 'Permutations and Combinations', which students may not be able to recall while solving and hence we saw a big dip in the percentage of students who got it right. Therefore, from this, it can be concluded that without any intermediate steps related to

Week 1 and 2 in the EMQ option, the correct percentage falls.

The aggregate scores of test takers in both the assessments are tabulated in Table 2. In the regular assessment, only the five questions which were similar in concept to EMQ assessment questions were considered for this.

Table 2. *Average scores of EMQ and Regular assessment*

	EMQ (%)	NAT/MCQ (%)
Overall	62.09 %	57.42 %
Students who attempted at least one question	75.11 % (830 students)	57.50 % (1318 students)

We performed t-test for independent means to test if the average score in EMQ assessment is significantly greater than that of regular assessment score of those students who attempted at least one of the questions in the assessment. The t-test gives a p-value of 0.000 (< 0.05). Hence, it can be concluded that the scores of EMQ assessment is significantly greater than that of regular assessment at 5 percent level of significance.

Thus, all this implies that EMQ assessment is not only a better way to test the students but is equally helpful for faculty and instructors. As it makes students apply their conceptual knowledge to solve the question and not by the rote memorization of when to do multiplication/addition, which results in improved scores of the test takers. On the other hand, the faculty and instructors are able to identify the weak spots of students in 'Permutations and Combinations' in certain questions and can accordingly work on the contents and interaction sessions with students.

4.2 Students' Perspectives on EMQ

We sent a feedback form to students to get their feedback on EMQ after the EMQ type practice assignments in Week-2 and Week-4 content of the Statistics-1 coursework. The main aim of the feedback form was to get an idea of students' view towards EMQ and any trouble they were facing in answering the EMQ assessments. In the Week-2 feedback form, we had 97 responses, in which approximately 60 percent of test takers found the regular assessment difficult compared to the EMQ assessment, and more than 70 percent of test takers wanted more such assessments in the upcoming weeks. But the majority of the audience wanted the way of answering to be more convenient, as they have to repeatedly scroll up and down to see the list of options and read the question.

To resolve this issue rapidly before the Week-4 content releases, we provided the PDF of questions along with the assessment too in order to avoid the struggle of test takers to repeatedly scroll up and down. But the way to submit answers, i.e. uploading ".txt" file is still a tedious task from students perspective and could be taken care of in future updates of the portal where they will be provided a textbox within the portal to input the answers.

After the Week-4 EMQ assessment, we received 27 responses to the feedback form, in which the majority of students (more than 55 percent) believed that even though EMQ assessment takes more time, it requires more understanding of the underlying concepts as compared to other types of questions. In addition, after these assessments the test takers believe that *"EMQ not only tests their understanding of concepts but also avoids the tedious calculations which helps in better decision-making."*

5. Conclusion

In this paper, we described how instructors adapted an assessment strategy to uncover students' conceptual gaps in an online Statistics course offered at scale. Based on student difficulties encountered in the course, instructors adapted 'Extended Matching Question (EMQ)' which has been primarily used in medical education, to identify the gap in students' conceptual understanding in the topic 'Permutations and Combinations'. This assessment type overcomes the limitations of the regular assessments, i.e. NAT and MCQ, as it helps students think and apply their knowledge conceptually to reach the final numeric answer rather than by the memorization of steps or guesswork. In addition, it also helps instructors capture the exact conceptual point where students need more attention, which helps in the improvement of contents in the coursework.

We evaluated the effectiveness of EMQ assessments by measuring student performance and taking perspectives of students and course instructors. Students attempted both assessments, i.e. regular type and EMQ type, each based on the similar concepts. Comparison of EMQ and regular NAT/MCQ assessments show that there is a significant increase in the scores of students in the EMQ type assessment as compared to that of regular assessment. Based on the analysis of students' answers to the EMQ type assessment, it is found that as the ability to apply the conceptual understanding of concepts increases, there are very high chances of getting the final numeric answer correct. Instructors also realized that students were struggling with certain concepts and question types, and this helped them plan their tutorial sessions with the students in the upcoming weeks.

This study is also a first order action research where the focus has been on addressing the development interest of the action researcher (Byrne et. al., 2021). We observe that the action researcher has systematically studied the problem at hand, designed and implemented a solution, collected data related to the implementation and analyzed the outcome of the implementation (Altrichter et. al., 2002). We observe that the online setting makes it easier for instructors to collect data related to their interventions and thus making quantitative methods of research more amenable. Further the roles of research mentors (in this case, second and third authors) in the research-practice partnership is critical in providing a meaningful structure for the action research. In the current study, the first author had consistent meetings with the research mentors to identify systematic ways of validating the findings and further structuring the write-up for the entire study. Universities involved in hybrid learning can invest in creating such research-practice partnerships as they provide these institutions with solutions to immediate problems of practice as well as larger research problems related to second-order that they can invest in for a longer term.

There are two limitations to our study. The first limitation is on the difference in the count of students who attempted both the assignments. We observe that there's a small portion of students who have not attempted the EMQ assignment or whose inputs were not stored due to formatting error while submission, but the size of the audience is large enough to overcome this effect while analyzing by aggregating the results of each of the two assignments. The second limitation is that, it could also be possible that the few may have attempted it just to get an idea of the assessment for future reference, without going through the content as they also have an option to opt out from the course end assessments. We intend to address these limitations by conducting longitudinal studies by providing EMQ assessments across weeks. This can give us a better understanding of the benefits of EMQ and how students' conceptual understanding is improving as the course progresses.

We believe that Extended Matching Questions can be adapted to other topics in Statistics and Maths as well. As part of our future research, we intend to use EMQs in other relevant Statistics and Math courses in the programme in a more longitudinal format. We believe that the results of this study will generate sufficient interest among instructors to move towards the higher levels of scholarships in teaching and learning.

References

Altrichter, H., Kemmis, S., McTaggart, R., & Zuber-Skerritt, O. (2002). The concept of action research. *The Learning Organization*, 9(3), 125-131

Byrne, V. L., Hogan, E., Dhingra, N., Anthony, M., & Gannon, C. (2021). An exploratory study of how novice instructors pivot to online assessments strategies. *Distance Education*, *42*(2), 184-199.

Cann, A. J. (2005). Extended matching sets questions for online numeracy assessments: a case study. Assessment & Evaluation in Higher Education, 30(6), 633-640.

Case, S. M., & Swanson, D. B. (1993). Extended-matching items: A practical alternative to free-response questions. *Teaching and Learning in Medicine: An International Journal*, *5*(2), 107-115.

Elliot, J. (1991). Action research for educational change. McGraw-Hill Education (UK).

Hanfstingl, B., & Pflaum, M. (2022). Continuing professional development designed as second-order action research: outcomes and lessons learned. *Educational Action Research*, *30*(2), 223-242.

Hodges, C. B., Moore, S., Lockee, B. B., Trust, T., & Bond, M. A. (2020). The difference between emergency remote teaching and online learning. *EDUCAUSE Review*, Accessible at: https://er.educause.edu/articles/2020/3/the-difference-between-emergency-remote-teaching-and-online-learning

Kearns, L. R. (2012). Student assessment in online learning: Challenges and effective practices. *Journal of Online Learning and Teaching*, 8(3), 198-208.

Posch, P. (2019). Action research–conceptual distinctions and confronting the theory–practice divide in Lesson and Learning Studies. Educational Action Research, 27(4), 496-510.

Samuels, A. (2006). Extended matching questions and the Royal Australian and New Zealand College of Psychiatrists written examination: an overview. *Australasian Psychiatry*, *14*(1), 63-66.

Stone, A., Allen, K., Rhoads, T. R., Murphy, T. J., Shehab, R. L., & Saha, C. (2003, November). The statistics concept inventory: A pilot study. In 33rd Annual Frontiers in Education, 2003. FIE 2003. (Vol. 1, pp. T3D-1). IEEE.

Vuma, S., & Sa, B. (2017). A descriptive analysis of extended matching questions among third year medical students. *Int J Res Med Sci*, *5*(5), 1913-1920.

Wood, E. J. (2003). What are Extended Matching Sets Questions?, *Bioscience Education*, 1:1, 1-8.

Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Teachers' Classroom Digital Technology Integration Practices

Doris Kristina RAAVE[a*], Eric ROLDAN ROA[a], Margus PEDASTE[a] & Katrin SAKS[a]
[a]*Institute of Education, University of Tartu, Estonia*
*doris.raave@ut.ee

Abstract: Classroom digital technology integration (CDTI) is increasingly common to facilitate technology-enhanced learning. However, in the Estonian context, there is a lack of understanding of the CDTI's practices that is yet necessary to evaluate the effect of purposeful CDTI to provide research-based advocation for CDTI. Our study aims to contribute to the field by identifying Estonian basic education teachers' practices regarding CDTI, considering the use of digital technology and its purposes. To that end, we conducted in-class observations of basic education teachers' lessons to note the use of digital technology. We analyzed the observed uses based on the Substitution, Augmentation, Modification and Redefinition (SAMR) model. Furthermore, we interviewed the teachers to understand their purposes for the DTI and categorized the given explanations by categories of technology-enhanced learning. We conducted cluster analyses to identify the types of practices that we then described and compared. We identified four types of teachers' practices: introducing, facilitating, motivating, and deepening, which differed in the purpose and use of digital technology in instructional processes. These results make way for evaluating the effect of different practices on learning outcomes.

Keywords: classroom digital technology integration, basic education, SAMR, technology-enhanced learning, technology-enhanced teaching

1. Introduction

Classroom digital technology integration (CDTI) is prominent in modern society and education. Previous research suggests the pandemic created an opportunity for educational innovations previously proposed but not fully implemented, advocating for a need to reimagine post-COVID-19 education using technology (Zhao & Watterston, 2021). Due to digital technology's learning affordances, education policies support the integration of digital technology in education (e.g., OECD, 2020). For example, Estonia, a digital forerunner, has set developing and implementing digital pedagogy as a future education goal (EMER, 2020). Digital readiness is high among Estonian teachers (e.g., Karaseva, Pruulmann-Vengerfeldt, & Siibak, 2013, 2017), and the Estonian teacher's professional standards framework emphasizes the use of digital technology in teaching and learning (Pedaste et al., 2019). As Estonia is one of the top-performing education systems (OECD, 2020), it may be beneficial to delve deeper into this context to shed light on the CDTI's effect on learning outcomes.

Estonia's K-12 curriculum gives teachers considerable autonomy in facilitating content (Tire, 2021). This autonomy contributes to teachers' agency (Ketelaar et al., 2013), making way for ownership (Godfrey & Olson, 2019), which is necessary to implement innovations successfully (Ketelaar et al., 2013). To increase ownership, decision-making should include stakeholders, as in Estonia, where teachers have direct input on educational policies (Gabriel et al., 2022). Further, students are increasingly involved in creating educational experiences under the paradigm of contemporary learning (EMER, 2020). In this paradigm, students are at the centre of learning design and instruction to scaffold their agency, keeping in mind the need for independent and self-regulated learners in a rapidly changing, uncertain world (Sum, Chan, & Wong, 2021). Previous research links co-constructivist practices alike to CDTI (Admiraal et al., 2017). Indeed, under the theory of technology-mediated learning, it is crucial to recognize the role played by the stakeholders in influencing the CDTI (Bower, 2019). According to previous studies, teachers are the primary agents determining the CDTI (Liu & Chao, 2017), so studying teacher practices is the key to understanding the phenomenon. A 2013 qualitative study in Estonia identified CDTI teacher CDTI practices practices in relation to subjects and

instructional styles (Karaseva, Pruulmann-Vengerfeldt, & Siibak, 2013). Due to the complex relationships between CDTI practices and learning outcomes, they should be considered interwoven with various factors, including purpose (Fawns, 2022), giving way to an understanding of teachers' pedagogical stance and its role in mediating learning (Aubrey-Smith, 2021). To contribute to the latter, our study seeks to get insight into teachers' CDTI practices and is thus guided by the following research question, what are teachers' digital technology integration practices regarding the use of digital technology and its purpose in the instruction?

2. Background

One of the most used frameworks for describing digital technology integration is the Substitution, Augmentation, Modification, and Redefinition framework (also known by its acronym SAMR), first proposed by Puentedura (2006). For SAMR's refined amalgamation of other frameworks and its popularity among researchers and educators (e.g., Crompton & Burke, 2020; Tlili et al., 2022), we will also base our study on the SAMR framework. SAMR framework describes CDTI in four levels: Substitution, Augmentation, Modification, and Redefinition. Substitution refers to learning activities that can be done with and without technology as it does not afford any functional change in the task, e.g., infinity canvas. Augmentation represents the use of digital technology in which there is a functional improvement in the learning task, e.g., automated control. Modification indicates an alteration of the learning task made possible due to the CDTI, e.g., computer simulations. Redefinition allows for a new learning task that could not be accomplished without integrating digital technology, e.g., hybrid learning.

Despite the model's representation in levels, none of the types of use is superior to one another, as meaningful CDTI depends on the learning objectives (Hamilton, Rosenberg, & Akcaoglu, 2016). As agentic intentions are considered of humans and not technologies (Bower, 2019), meaning digital technologies' affordances and CDTI practices are user-dependent (Aubrey-Smith, 2021), it is important to consider the CDTI's underlying purposes, which helps understand the connection between the CDTI and learning outcomes and variables mediating it. As the aims of the CDTI are not considered in the SAMR model, our study combines the SAMR model with the three ways digital technology enhances teaching and learning processes proposed by Kirkwood and Price (2013) in their meta-analysis. First, operational improvement concerns more flexibility or efficiency, e.g., e-books are faster to navigate than printed books. Quantitative changes refer to increased student involvement, time spent on task, and positive attitude, e.g., gamification. Third, qualitative change indicates facilitating higher-order thinking, critical awareness, and deeper understanding, e.g., using digital technologies as tools for reflection.

3. Methodology

3.1 Sample

The study has a purposeful sample of Estonian 3rd, 6th and 9th-grade teachers of Estonian, mathematics, and natural sciences. The teachers were chosen based on the profiles of the schools they worked at. The schools were selected purposefully to have a broad representation of schools considering the following: students' academic and digital competence tests, the school's self-assessment of digital competence and the school's result in the school satisfaction survey. Teachers' participation in the study was voluntary. In total, 93 teachers, of whom 24% were 3rd grade, 39% 6th grade, 24% 9th grade and 14% 6th and 9th-grade teachers from 13 different schools, participated in the study. Amongst the total, 76% of 6th and 9th-grade teachers, 28% were Estonian language teachers, 44% mathematics and 28% natural sciences teachers. Usually, the same 3rd-grade teacher teaches the Estonian language, mathematics, and natural sciences in Estonia.

3.2 Data Collection

Seven purposefully trained researchers observed in-class lessons and interviewed teachers who used digital technology in their lessons. Each subject (Estonian, math, natural sciences) lesson was observed once in each grade (3rd, 6th, 9th). The lessons were chosen randomly because we were interested in how digital technology is used in the everyday classroom, not in a specifically prepared situation of CDTI. In total, 169 lessons were observed, of which 82% included CDTI. These lessons included 269 different learning activities with the use of digital technology. During observations, we noted down descriptions of who used what digital technology and how. We conducted immediately after the observations semi-structured interviews soliciting answers to a question about why digital technology was used in the lessons. During the interviews, meaningful utterances were noted down for coding later.

3.3 Data Analysis

The data were analyzed under the SAMR model and the categories of purposes of digital technology integration (i.e., operational improvement, quantitative and qualitative change). Besides these deductively coded dimensions, researchers were opened to other categories that might emerge inductively as the theoretical underpinnings of our study had not been previously used in combination in the K-12 context in Estonia. Each teacher's observational and interview data were used to create a string of binary codes of the considered dimensions to describe the teacher's practice. Code 0 was attributed when the dimension was not present, and code 1 when it was present. Descriptions of events relating to the use of digital technology were noted during observations and then encoded according to the SAMR framework. Relevant utterances of the interviews were noted down and then coded according to Kirkwood & Price's (2013) categories for using digital technology to support teaching and learning: operational, quantitative, and qualitative. For inter-rater reliability, we used Fleiss Kappa, which allows considering more than two raters. The Kappa was 0.79, indicating acceptable reliability.

To identify types of teacher practices in the use of digital technology, we ran hierarchical clustering using Ward's method to identify clusters in which the dispersity within a cluster would be minimal but the diversity between the clusters maximal. The number of clusters and cluster centroids were determined from the dendrogram and further optimized with K-means cluster analyses. Clusters were validated by first checking clusters' stability by replicating the analysis by randomly dividing the sample into two halves and considering clusters' meaningfulness through ANOVA analysis, indicating variation between sample means and eta squares, showing which variables distinguish the clusters the most. The identified clusters were described and compared using means to find the statistically and meaningfully best fitting cluster model. All the statistical tests were done using the program SPSS version 28.0.

4. Results

We performed a deductive analysis of teachers' CDTI practices in relation to the use of digital technology based on the SAMR framework and purpose under the technology-enhanced learning and teaching categories proposed by Kirkwood and Price (2013). All the sub-dimensions of use and purpose were present in our data. Furthermore, we identified one additional category, namely sustainability. Although all the sub-dimensions were present, the level of their occurrence varied considerably. Regarding digital technology, uses on the Substitution and Augmentation level were markedly more prominent than on the Modification and Redefinition level. Considering the purposes of digital technology integration, categories of technology-enhanced teaching were considerably less common than categories of technology-enhanced learning.

Based on hierarchical clustering of teachers' CDTI practices and cluster comparison by ANOVA analysis, we identified four different teachers' practices, named as follows: facilitating, motivating, deepening, and introducing. All four practices differ in their use and purpose, while differences in the latter seem the most prominent. Differentiations relating to use appear mainly on the Augmentation level but somewhat also in use on the Substitution level. The objectives of CDTI differ mainly regarding student-centred purposes (see Figure 1).

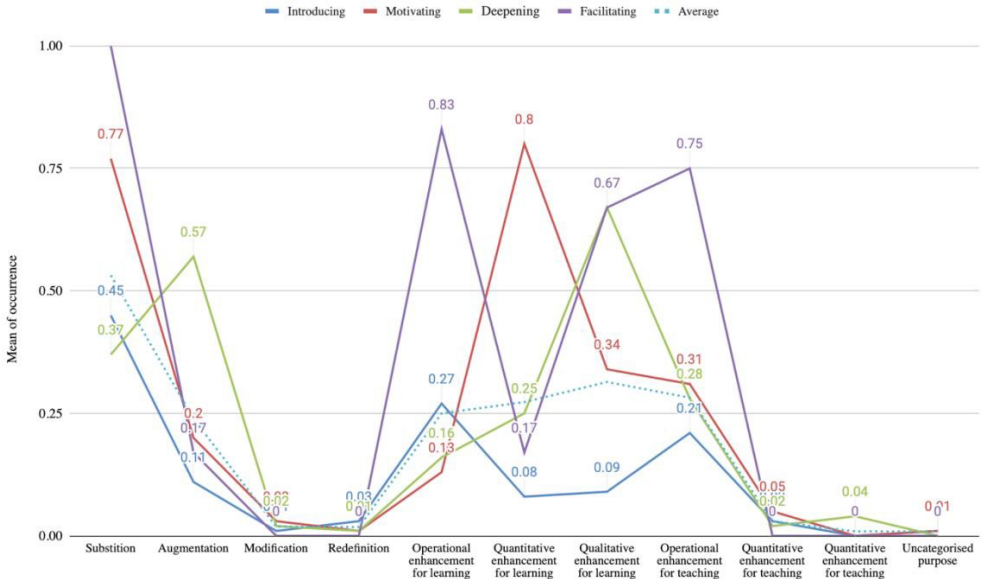


Figure 1. Teacher practices in classroom digital technology integration.

The CDTI is the most varied amongst deepening practices, using it for Substitution and Augmentation, the latter being most prominent. Deepening focuses mainly on CDTI's different affordances for the qualitative enhancement of learning. Facilitating, motivating, and introducing practices use technology mostly, although with different regularities, for Substitution, but their purposes differ. More specifically, facilitating stands out from the others regarding the purposes of operational enhancement of teaching and learning and qualitative enhancement for students, i.e., facilitating the teaching-learning processes while keeping in mind in-depth learning. On the other hand, motivating practices focus mainly on quantitative enhancement for students, such as motivation and engagement. The use of digital technology is the least common practice among introducing, and it is primarily without a specific aspect of enhancement in mind when doing so.

The variable that distinguishes the clusters the most is a quantitative enhancement for learning (60.3%), followed by qualitative enhancement for learning (46.7%), use of digital technology on the Augmentation level (41.9%) and operational improvement for learning (20.3%). The values of these variables were also statistically significantly different in describing variation between means based on ANOVA. A somewhat prominent distinguisher (12.5%), operational improvement for teaching is close to being a statistically significant variable in variations. Digital technology uses on the levels of Modification and Redefinition, quantitative and qualitative enhancements of teaching, and uncategorized enhancement, to distinguish the clusters minimally (between 1.1% and 6.9%). These variables were also not statistically significantly different in describing variations between clusters. Reasons for this might be their low occurrences in the data set (less than 2.5% of the cases). Hence, these variables were left out for describing and comparing clusters.

5. Discussion & Conclusion

Our study aimed to identify types of Estonian basic education teachers' classroom digital technology integration practices considering the use of digital technology under the SAMR model (Puentedura, 2006) and CDTI's purpose in the instruction as suggested by Kirkwood and Price (2013). From the data, an additional purpose for the CDTI emerged, namely, sustainability, aligning with the transformative concept, e.g., technology-mediated, sustainable education necessary for supporting continual coexistence with other ecosystems (Burns, 2018). Although all the sub-dimensions of the SAMR framework were noted, uses on the Modification and Redefinition level of CDTI were far less

common than uses on the Substitution and Augmentation level. The latter is a possible indicator that Estonian teachers use CDTI mainly to replicate and supplement existing practices rather than transform them.

Within these general tendencies regarding teachers' CDTI practices, our study identified four practices: introducing, facilitating, motivating, and deepening. In introducing, facilitating, and motivating, CDTI is mainly used for Substitution, although with different regularities; in deepening practices, CDTI is also used for Augmentation, yielding in more varied use of CDTI. Deepening practices focus more on qualitative enhancement, e.g., facilitating the teaching-learning processes and learning opportunities for students. Qualitative enhancement is also a significant catalyst for practices of facilitating alongside a more protruding focus on operational enhancement. Motivating practices focus on quantitative enhancement, e.g., motivating and engaging. Introducing practices where technology is used in less than half of the events has no strong specific purpose for CDTI.

Hence, these practices differ in the purpose and use of digital technology in instructional processes, while the first is more protruding in differentiating the clusters. This discrepancy may be explained by purposes serving as a more robust insight into teachers' pedagogical stance than mere uses (Aubrey-Smith, 2021); however, a connection between those has also been noted (e.g., Admiraal et al., 2017). These results suggest that teachers' knowledge and skills alone may not be enough to support meaningful technology use in learning. Teachers also need to have a clear understanding of the purposes of CDTI. The latter is not easily developed in traditional in-service training but needs to be discussed among groups of teachers as they act as primary agents in CDTI (Liu & Chao, 2017).

Further, it is worth noting that regarding the purposes of the CDTI, student-centred purposes were considerably more prominent than teacher-centred purposes, implying that Estonian teachers' practices, and thus possibly their pedagogical stances (Aubrey-Smith, 2021), focus on students. Furthermore, as the teachers who focused more on student-centred purposes were the teachers with more frequent or varied practices of CDTI, our results align with previous findings that indicate a connection between teachers' technology practices and co-constructivist teaching approach, i.e., approach where learning builds on a dialogue between teachers and students or peers (Admiraal et al., 2017), which might be an indicator of the relative success behind Estonian educational system, taken that Estonia excels in academic achievement in international tests like PISA (OECD, 2020).

The results of our study might serve as valuable information to support stakeholders in developing and implementing CDTI. More specifically, the practical implications of our study are threefold. First, to support meaningful and effective CDTI, we suggest offering tailored support and training to the teachers considering their prominent type of CDTI practice rendering, thus the support most beneficial for them. Second, considering teachers' practices may give way for more meaningful dialogue between teachers and developers of educational technology so that this technology would be knowledgeably designed for empowering teachers and students in the teaching-learning processes whilst being considerate of relevant challenges in the domain, including teacher's practices and pedagogical stances that mediate these. Third, the results of our study may serve as a starting point for studying the impact of different CDTI practices on students' learning outcomes. Therefore, it constitutes commencement for research into the advocation for meaningful uses of digital technology where the use is planned based on the learning objectives while considering the ways digital technology can facilitate the learning process for different kinds of students.

Finally, we also acknowledge some limitations in our study. First, our study did not consider possible reasons behind the identified practices, which could support the meaning-making of the practice and thus also the teachers' pedagogical stance. For example, we could consider the following factors influencing the CDTI practices: contextual variables and teacher-specific characteristics such as perceptions of the CDTI. Second, the types of practices were identified based on means. Yet, within the practice types, there may be teachers who are not comparable to the other teachers in the same cluster regarding all the variables. Therefore, future research is needed to collect further data on possible related factors of the identified practices to underpin sense-making and increase the trustworthiness of the results by conducting cluster analysis and latent profile analysis, combining it with member checking.

6. Acknowledgements

This study has been supported by the DIGIVARA5 project "The Effect of Using Digital Learning Materials for Learning and Teaching in the Context of Estonian Basic School (1.05.2020 − 30.04.2023)" financed by the Ministry of Education and Research in Estonia.

References

Admiraal, W., Louws, M., Lockhorst, D., Paas, T., Buynsters, M., Cviko, A., ... & Kester, L. (2017). Teachers in school-based technology innovations: A typology of their beliefs on teaching and technology. Computers & Education, 114, 57–68.

Aubrey-Smith, F. (2021). An Exploration Of The Relationship Between Teachers' Pedagogical Stance And The Use Of ICT In Their Classroom Practice (Doctoral dissertation, The Open University).

Burns, H. (2018). Thematic analysis: Transformative sustainability education. Journal of Transformative Education, 16(4), 277–279.

Bower, M. (2019). Technology-mediated learning theory. British Journal of Educational Technology, 50(3), 1035–1048.

Crompton, H., & Burke, D. (2020). Mobile learning and pedagogical opportunities: A configurative systematic review of PreK-12 research using the SAMR framework. Computers & Education, 156.

EMER (Estonian Ministry of Education and Research) (2020). Education strategy 2021-2035 [Web page]. Retrieved May 31, 2022, from https://www.hm.ee/sites/default/files/haridusvaldkonna_arengukava_2035_kinnittaud_vv_eng_0.pdf

Fawns, T. (2022). An Entangled Pedagogy: Looking Beyond the Pedagogy—Technology Dichotomy. Postdigital Science and Education, 1–18.

Gabriel, F., Marrone, R., van Sebille, Y., Kovanovic, V., & de Laat, M. (2022). Digital education strategies around the world: practices and policies. Irish Educational Studies, 41(1), 85–106.

Godfrey, L., & Olson, C. B. (2019). Agency as the achievement of reform ownership. English Teaching: Practice & Critique, 18(2), 249–262.

Hamilton, E. R., Rosenberg, J. M., & Akcaoglu, M. (2016). The Substitution Augmentation Modification Redefinition (SAMR) Model: a Critical Review and Suggestions for its Use. TechTrends, 60(5), 433–441.

Karaseva, A., Pruulmann-Vengerfeldt, P., & Siibak, A. (2013). Comparison of Different Subject Cultures and Pedagogical Use of ICTs in Estonian Schools. Nordic Journal of Digital Literacy, 8(3), 157–171.

Karaseva, A., Pruulmann-Vengerfeldt, P., & Siibak, A. (2017). Relationships between in-service teacher achievement motivation and use of educational technology: case study with Latvian and Estonian teachers. Technology, Pedagogy and Education, 27(1), 33–47.

Ketelaar, E., Koopman, M., den Brok, P. J., Beijaard, D., & Boshuizen, H. P. A. (2013). Teachers' learning experiences in relation to their ownership, sense-making and agency. Teachers and Teaching, 20(3), 314–337.

Kirkwood, A., & Price, L. (2013). Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know? A critical literature review. Learning, Media and Technology, 39(1), 6–36.

Liu, Q., & Chao, C. C. (2017). CALL from an ecological perspective: How a teacher perceives affordance and fosters learner agency in a technology-mediated language classroom. ReCALL, 30(1), 68–87.

OECD 2020. PISA 2018 Results (Volume V): Effective Policies, Successful Schools, Paris: OECD Publishing.

Pedaste, M., Leijen, L., Poom-Valickis, K., & Eisenschmidt, E. (2019). Teacher professional standards to support teacher quality and learning in Estonia. European Journal of Education.

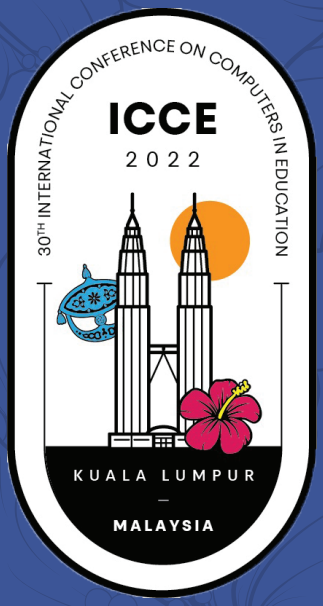
Puentedura, R. (2006). Transformation, technology, and Education [Web page]. Retrieved May 31, 2022, from http://hippasus.com/resources/tte/puentedura_tte.pdf

Sum, C., Chan, I., & Wong, H. (2021). Ready to learn in an uncertain future: ways to support student engagement. Accounting Research Journal, 34(2), 169–183.

Tire, G. (2021). Estonia: A Positive PISA Experience. In N. Crato (Ed.), Improving a Country's Education (pp. 101–120). Springer, Cham.

Tlili, A., Padilla-Zea, N., Garzón, J., Wang, Y., Kinshuk, K., & Burgos, D. (2022). The changing landscape of mobile learning pedagogy: A systematic literature review. Interactive Learning Environments, 1–18.

Zhao, Y., & Watterston, J. (2021). The changes we need: Education post COVID-19. Journal of Educational Change, 22(1), 3–12.



ORGANIZED BY

THE ASIA-PACIFIC SOCIETY FOR COMPUTERS IN EDUCATION (APSCE)

HOSTED BY

FACULTY OF EDUCATIONAL STUDIES, UNIVERSITI PUTRA MALAYSIA, MALAYSIA

SUPPORTED BY

Malaysia Convention & Exhibition Bureau

Meet in Malaysia

Malaysia Truly Asia