# **Educational Technology & Society**

Published by International Forum of Educational Technology & Society Hosted by National Taiwan Normal University, Taiwan

Educational Technology 8 Society

vol.**26** no.**2** 

Educational Technology & Society has Impact Factor 2.633 and 5-Year impact factor 4.358 according to Thomson Scientific 2021 Journal Citations Report.

# Volume 26 Issue 2

# April 2023

ISSN: 1436-4522 (online) ISSN: 1176-3647 (print) DOI: 10.30191/ETS http://www.j-ets.net/

## **Educational Technology & Society**

An International Journal

#### Aims and Scope

*Educational Technology & Society (ET&S)* is an open-access academic journal published quarterly (January, April, July, and October) since October 1998. *ET&S* has achieved its purposes of providing an international forum for open access scientific dialogue for developers, educators and researchers to foster the development of research in educational technology. Thanks to all the Authors, Reviewers and Readers, the journal has enjoyed tremendous success.

*ET&S* has established a solid and stable editorial office where the Editors-in-Chief aims to promote innovative educational technology research based on empirical inquiries to echo the pedagogical essentials of learning in the real world—lifelong learning, competency-orientation, and multimodal literacy in the 21st century.

*ET&S* publishes research that well bridges the pedagogy and practice in advanced technology for evidence-based and meaningful educational application. The focus of *ET&S* is not only technology per se, but rather issues related to the process continuum of learning, teaching, and assessment and how they are affected or enhanced using technologies. The empirical research about how technology can be used to overcome the existing problems in the frontline of local education with findings that can be applied to the global spectrum is also welcome. However, papers with only descriptions of the results obtained from self-report surveys without systematic or empirical data or any analysis on learning outcomes or processes are not favorable to be included in *ET&S*.

*ET&S* publishes research that fulfills any of the following description:

- Evidence-based Research: Continuous research that is conducted within a sufficient amount of time such that the effectiveness of the research intervention in education can be evaluated and demonstrated convincingly via empirical methods. Preferably, research experience, outcome, and results for a semester long of adopting educational technologies; or an evaluation/experiment that is conducted multiple times and spans over several semesters.
- Emerging Technology: New and emerging technology used in education
- Special Domain Research: Research with specific participants using technology in education, for example, those with privacy issues in special education or those of younger ages like infants and toddlers, who may not form a large pool of participants

#### Founding Editor

Kinshuk, University of North Texas, USA.

#### Journal Steering Board

Nian-Shing Chen, National Taiwan Normal University, Taiwan; Kinshuk, University of North Texas, USA; Demetrios G. Sampson, University of Piraeus, Greece.

#### Editors-in-Chief

Maiga Chang, Athabasca University, Canada; Dirk Ifenthaler, University of Mannheim, Germany; Yu-Ju Lan, National Taiwan Normal University, Taiwan.

#### Associate Editors

Yacine Atif, University of Skövde, Sweden; Howard Hao-Jan Chen, Department of English, National Taiwan Normal University, Taiwan; Hsieh-Jun (Scott) Chen, Graduate Institute of Children's English, National Changhua University of Education, Taiwan; Yannis Dimitriadis, Universidad de Valladolid, Spain; Kyparisia Papanikolaou, School of Pedagogical & Technological Education, Greece; Yun Wen, National Institute of Education, Singapore.

#### **Editorial Board Members**

Ahmed Hosny Saleh Metwally, Northeast Normal University, China; Bernardo Pereira Nunes, The Australian National University, Australia; Ching-sing Chai, The Chinese University of Hong Kong, Hong Kong; David Gibson, Curtin University, Australia; Grace Yue Qi, Massey University, New Zealand; Ig Ibert Bittencourt Santana Pinto, Universidade Federal de Alagoas, Brazil; Jerry Chih-Yuan Sun, National Chiao Tung University, Taiwan; Jie Chi Yang, National Central University, Taiwan; Joice Lee Otsuka, Federal University of São Carlos, Brazil; Kaushal Kumar Bhagat, Indian Institute of Technology, India; Minhong Wang, The University of Hong Kong, Hong Kong; Morris Siu-Yung Jong, The Chinese University of Hong Kong, Hong Kong; Regina Kaplan-Rakowski, University of North Texas, USA; Rita Kuo, New Mexico Tech, USA; Robert Li-Wei Hsu, National Kaohsiung University of Hospitality and Tourism, Taiwan; Rustam Shadiev, Nanjing Normal University, China; Stephen J.H. Yang, National Central University, Taiwan; Tony Liao, NOAA Earth System Research Laboratories, USA; Wen-Ta Tseng, National Taiwan University of Science and Technology, Taiwan; Yanjie Song, Education University of Hong Kong, Hong Kong; Ahmed Tlili, Smart Learning Institute of Beijing Normal University, China; Chia-Wen Tsai, Department of Information Management, Ming Chuan University, Taiwan; Hsueh Chu Rebecca Chen, Education University of Hong Kong, Hong Kong; Nur Hamid, Universitas Islam Negeri Walisongo Semarang, Indonesia; Katrin Saks, University of Tartu, Estonia; Sheng-Shiang Tseng, Tamkang University, Taiwan; Siddharth Srivastava, Indian Institute of Technology Kanpur, India; Li Wang, Open University of China, China; Robin Jocius, University of Texas at Arlington, USA; Dongsik Kim, Hanyang University, South Korea; Chiu-Lin Lai, National Taipei University of Education, Taiwan; Daner Sun, Education University of Hong Kong; Ying-Tien Wu, National Central University, Taiwan; Hui-Chin Yeh, National Yunlin University of Science and Technology, Taiwan.

#### Managing Editor

Sie Wai (Sylvia) Chew, National Taiwan Normal University, Taiwan.

#### **Editorial Assistant**

I-Chen Huang, National Taiwan Normal University, Taiwan.

#### **Technical Manager**

Wei-Lun Chang, National Taiwan Normal University, Taiwan.

#### **Executive Peer-Reviewers**

see http://www.j-ets.net

#### Publisher

International Forum of Educational Technology & Society

#### Host

National Taiwan Normal University, Taiwan

#### **Editorial Office**

c/o Chair Professor Nian-Shing Chen, Institute for Research Excellence in Learning Sciences, Program of Learning Sciences, National Taiwan Normal University, No.162, Sec. 1, Heping E. Rd., Da-an Dist., Taipei City 10610, Taiwan.

#### **Supporting Organizations**

University of North Texas, USA University of Piraeus, Greece

#### Website Maintenance

Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taiwan

#### Abstracting and Indexing

Educational Technology & Society is abstracted/indexed in Social Science Citation Index, Scopus, ACM Guide to Computing Literature, airiti, Australian DEST Register of Refereed Journals, Computing Reviews, Current Contents/Social & Behavioral Sciences, DBLP, DOAJ, Educational Administration Abstracts, Educational Research Abstracts, Educational Technology Abstracts, Elsevier Bibliographic Databases, ERIC Clearinghouse on Information & Technology, Inspec, ISI Alerting Services, JSTOR, PsycINFO, Social Scisearch, Technical Education & Training Abstracts, and VOCED.

#### **Guidelines for authors**

- Submissions are invited in the following categories:
- Peer reviewed publications: Full length articles (up to 8,000 words excluding References and Appendices)
- Special Issue publications

All peer review publications will be referred in double-blind review process by at least two international reviewers with expertise in the relevant subject area.

For detailed information on how to format your submissions, please see: https://www.j-ets.net/author\_guide

For Special Issue Proposal submission, please see: https://www.j-ets.net/journal\_info/special-issue-proposals

#### Submission procedure

All submissions must be uploaded through our online management system (http://www.j-ets.net). Do note that all manuscripts must comply with requirements stated in the Authors Guidelines.

Authors, submitting articles for a particular special issue, should send their submissions according to the channel specified in the Call for Paper of the special issue.

All submissions should be in electronic form. Authors will receive an email acknowledgement of their submission.

The preferred formats for submission are Word document, and not in any other word-processing or desktop-publishing formats. Please place figures and tables in their respective format in the anonymous manuscript along with all appendices (if any).

Please provide following details with each submission in a separate file (i.e., Title Page):

- Author(s) full name(s) including title(s),
- Name of corresponding author,
- Job title(s),
- Organisation(s),
- Full contact details of ALL authors including email address, postal address, telephone and fax numbers.

In case of difficulties, please contact journal.ets@gmail.com.

## **Educational Technology & Society**

Volume 26 Number 2 2023

## **Table of Contents**

## **Editorial Position Paper**

Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions *Gwo-Jen Hwang and Nian-Shing Chen* 

#### **Full Length Articles**

Effects of a Hybrid Training for Plagiarism Prevention Module on Plagiarism-free Academic Writing in Higher Education <i>Yin Zhang, Samuel Kai Wah Chu, Yonghui Liu and Xiaoli Lu</i>	1–18
Using E-textiles to Design and Develop Educational Games for Preschool-aged Children Ersin Kara and Kursat Cagiltay	19–35
Effects of Mobile-Assisted Language Learning on EFL Learners' Listening Skill Development Rui Li	36–49
The Interplay Between Cognitive Load and Self-Regulated Learning in a Technology-Rich Learning Environment <i>Tingting Wang, Shan Li and Susanne Lajoie</i>	50–62
Lag Sequential Analysis for Identifying Blended Learners' Sequential Patterns of e-Book Note-taking for Self-Regulated Learning Christopher C.Y. Yang and Hiroaki Ogata	63–75
Effects of Undergraduate Student Reviewers' Ability on Comments Provided, Reviewing Behavior, and Performance in an Online Video Peer Assessment Activity Liang-Yi Li and Wen-Lung Huang	76–93
Editorial Note	
Creating Computational Thinkers for the Artificial Intelligence Era—Catalyzing the Process through Educational Technology Ahmed Tlili, Daniel Burgos and Chee-Kit Looi	94–98
Special Issue Articles	
Integrating Computational Thinking into Primary and Lower Secondary Education: A Systematic Review Panagiotis Kampylis, Valentina Dagienė, Stefania Bocconi, Augusto Chioccariello, Katja Engelhardt, Gabrielė Stupurienė, Vaida Masiulionytė-Dagienė, Eglė Jasutė, Chiara Malagoli, Milena Horvath and Jeffrey Earp	99–117
A Normative Analysis of the TechCheck Computational Thinking Assessment Emily Relkin, Sara K. Johnson and Marina U. Bers	118–130
Integration of Computational Thinking with Mathematical Problem-based Learning: Insights on Affordances for Learning <i>Zhihao Cui, Oi-lam Ng and Morris Siu-Yung Jong</i>	131–146
The SNS-based E-mentoring and Development of Computational Thinking for Undergraduate Students in an Online Course Yeonju Jang, Seongyune Choi, Seonghun Kim and Hyeoncheol Kim	147–164
Effect of a Reflection-Guided Visualized Mindtool Strategy for Improving Students' Learning Performance and Behaviors in Computational Thinking Development <i>Xiao-Fan Lin, Jing Wang, Yingshan Chen, Yue Zhou, Guoyu Luo, Zhaoyang Wang, Zhong-Mei Liang,</i> <i>Xiaoyong Hu and Wenyi Li</i>	165–180

ISSN 1436-4522 (online) and 1176-3647 (print). DOI 10.30191/ETS. This article of *Educational Technology & Society* is available under Creative Commons CC-BY-NC-ND 3.0 license (https://creativecommons.org/licenses/by-nc-nd/3.0/). For further queries, please contact Editors at ets, editors@gmail.com.

A Framework for Applying Sequential Data Analytics to Design Personalized Digital Game-Based 181–197 Learning for Computing Education *Zhichun Liu and Jewoong Moon* 

ISSN 1436-4522 (online) and 1176-3647 (print). DOI 10.30191/ETS. This article of *Educational Technology & Society* is available under Creative Commons CC-BY-NC-ND 3.0 license (https://creativecommons.org/licenses/by-nc-nd/3.0/). For further queries, please contact Editors at ets,editors@gmail.com.

Hwang, G.-J., & Chen, N.-S. (2023). Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions. *Educational Technology & Society*, 26(2). https://doi.org/10.30191/ETS.202304\_26(2).0014

## Editorial Position Paper: Exploring the Potential of Generative Artificial Intelligence in Education: Applications, Challenges, and Future Research Directions

## Gwo-Jen Hwang<sup>1,2</sup> and Nian-Shing Chen<sup>3\*</sup>

<sup>1</sup>Graduate Institute of Digital Learning and Education, National Taiwan University of Science and Technology, Taiwan // <sup>2</sup>Graduate Institute of Educational Information and Measurement, National Taichung University of Education, Taiwan // <sup>3</sup>Institute for Research Excellence in Learning Sciences, Program of Learning Sciences, National Taiwan Normal University, Taiwan // gjhwang.academic@gmail.com // nianshing@gmail.com \*Corresponding author

**ABSTRACT:** Generative artificial intelligence (GAI) applications, such as ChatGPT (Chat Generative Pretrained Transformer) and Midjourney, have recently attracted much attention from researchers and school teachers. While many people are eager to learn more about GAI applications, some scholars are concerned about the potential misuse of them. It is predicted that the use of GAI applications will increase rapidly in the coming years. Therefore, it is important to consider the challenges and research issues through some concrete application examples of using GAI for education. In this position paper, the authors aim to address these issues from the perspectives of academic research and educational objectives. Along with defining GAI, several illustrative examples of using GAI applications in educational settings are provided. Moreover, potential research issues of GAI-based learning, including research design, relevant learning strategies, research focus, and measuring tools, are discussed. ET&S journal is especially welcoming research on unlocking the potential of GAI for education to realize the two notions of "Knowing [why] is the essential element for learners to have in-depth understanding" and "It is all about prompts: Get rid of the 'search' mindset and use 'programming prompt' instead."

Keywords: Generative artificial intelligence, ChatGPT, Midjourney, Artificial Intelligence in education, Programming prompt

## 1. Challenges of generative AI-based learning

The use of Generative Artificial Intelligence (GAI), such as ChatGPT and Midjourney, has recently garnered a lot of attention. For example, ChatGPT can generate detailed responses to questions related to diverse domains, and Midjourney can create images or drawings based on user's requests or descriptions. The generated text or images are almost indistinguishable from those created by humans, which is both impressive and concerning. While these applications offer the potential for creating quality articles or artworks, their misuse could lead to serious problems in education and social security. (Yang et al., 2021).

From an academic research perspective, several publishers have recently introduced new policies in response to the growing use of GAI applications by authors. For example, Elsevier has developed a new AI author policy to ensure the integrity of the scholarly record and to provide clear guidance to authors, readers, reviewers, and editors of their journals. According to the new policy, authors must declare that they use GAI technologies solely to enhance readability and language, not to replace essential researcher tasks such as data interpretation or drawing conclusions. Authors must supervise and control the use of GAI applications and review and edit the results to ensure accurate reporting. Additionally, as GAI applications serve as research or learning tools in studies, they cannot be listed as co-authors of a paper.

According to Hwang (2014), a crucial application of AI in educational settings is "enabling personalized learning." With the help of AI technologies, it has become possible to provide individual learners with guidance or support based on their learning performances and needs. This is why previous studies on AI in education (AIED) mainly focused on using AI technologies to enhance learners' abilities in memorizing, comprehending, applying, analyzing, and assessing. The emergence of GAI could even shift the educational objective to the highest cognitive level, which is creativity.

Despite the concerns and controversies, there is no doubt that using GAI applications in educational settings is becoming a new trend of technology-enhanced learning (Dehouche, 2021; Dwivedi et al., 2023). Several scholars have pointed out the potential of GAI-based learning, including promoting learners' creativity, critical thinking, and problem-solving performances (Eysenbach, 2023). To remove the barriers of adopting GAI applications in schools, it is important to develop guidelines or templates of GAI-based learning. To further depict the effectiveness and potential of GAI-based learning, more quality studies are required to further depict

the effectiveness and potential of GAI-based learning. In the following sections, several examples and guidelines of using GAI in educational settings are provided. In particular, some useful educational applications of GAI are presented for educators' reference. Following that, several research topics that are worth investigating are provided.

## 2. Roles of generative AI in education

By referring to the AIED model proposed by Hwang et al. (2020), we categorize the roles of GAI in education into several categories, that is, teacher/tutor, student/tutee, learning peer/partner, domain expert, administrator, and learning tool, as shown in Figure 1.



## 2.1. Teacher/Tutor

Traditional AI applications can provide learning guidance, support, and supplementary materials to students based on their learning status. In contrast, GAI applications can go a step further by summarizing learning content and demonstrating problem-solving or artwork creation like an experienced tutor. For instance, if a student struggles to write a program, they can seek help from ChatGPT, which can show them how to write the program.

## 2.2. Student/Tutee

So far, few AI applications that can play the role of a student or tutee. However, GAI applications have great potential to play the roles of a good tutee, as they are able to learn from users. For example, when students are asked to play the role of a teacher, they can train ChatGPT by inputting new knowledge related to a specified topic. As a tutee, ChatGPT is capable of learning from the knowledge provided by the students as well as from those collected from the Internet. As a tutor, students can also assign tasks to GAI applications and specify the rubrics or criteria for evaluating the quality of the tasks. For example, Midjourney can draw pictures to respond to the requests from the students who play the role of a tutor. If the students are not satisfied with the pictures, they can provide comments or detailed descriptions about the artwork they need and ask Midjourney to redraw the pictures. They can also rate the pictures based on the rubrics. By playing the role of a tutor and interacting with the GAI-based tutee, students have the opportunity to learn to perceive things from a tutor's perspective.

#### 2.3. Learning peer/partner

GAI can be a teammate in collaborative learning activities. For example, a teacher might conduct a collaborative learning activity in which a student and ChatGPT work together as a team to complete a project or task related to a specified topic. When presenting the report, each team needs to indicate how they collaborated, as well as the roles of the students and ChatGPT.

#### 2.4. Domain expert

By assigning GAI the role as a domain expert, a Generative Artificial Intelligence (GAI) application can provide solutions to problems encountered by users, who may be tutors or tutees. Additionally, the GAI can engage in discussions with the user about specific issues and offer valuable advice.

#### 2.5. Administrator

In terms of policy-making support, conventional AI applications mainly provide predictions based on the collected data and the adopted reasoning model. Not only can GAI applications, such as ChatGPT, make predictions by analyzing large sets of data, it can also summarize what they have found and present the findings in the form of an abstract or a report with tables and figures. The analytical ability of these GAI applications boosts its practicability and is even more helpful to decision makers.

#### 2.6. Learning tool

It is no doubt that GAI applications are powerful tools for assisting learners to collect and analyze data. GAI applications not only can work with learners but also enable them to focus on critical objectives by sharing their loads. For example, using Midjourney to create artworks enables students to focus on their creative thinking rather than spending time on drawing the details.

## 3. Examples and implications of using ChatGPT in academic research and education

There are several potential applications of using GAI in educational settings. For example, when playing the role of a tutor, GAI can generate lesson plans or be a proofreader for academic writing in language courses. When playing the role of a research/teaching tool, researchers or educators can work with GAI to design question items for assessment or develop rubrics for evaluating students' essays.

#### 3.1. Examples of GAI-based learning

#### 3.1.1. Academic article proofreading

Several GAI applications are capable of proofreading academic writing. From the perspective of learning design, there are three levels of guiding students to work with GAI in academic writing activities. Taking ChatGPT as an example, in Level 1, students only need to use a simple prompt, such as "Proofread my writing" to ask ChatGPT to start proofreading the academic writing, as shown in Figure 2(a).

Figure 2(b) shows the proofreading results at Level 1 provided by ChatGPT. It is important to note that at this level of proofreading request, ChatGPT only provides a revised version of the text without any explanations.

As for Level 2 proofreading, learners need to provide a more concrete prompt that includes the task and the goal. For example, in addition to requesting proofreading of the academic writing, additional instructions can be added to the prompt, such as "make the writing more professional" and "explain why the changes were made" as shown in Figure 3(a).

A more professional version was created by providing ChatGPT with a specific prompt, as shown in Figure 3(b). In addition to proofreading the text, ChatGPT also provided explanations for each change made, as seen in

Figure 3(c). This not only improved the grammar, spelling, and punctuation, but also enhanced the overall flow and structure of the writing.

Figure 2(a). Illustrative exmaple of submitting a Level 1 preooferading request to ChatGPT

## Proofread my writing:

Some guidelines for providing comprehensive prompts to GAI like ChatGPT to get the best responses.

1. Be specific: Describe your goal/question with specific and clear wordings, avoid too general topics or ambiquous words.

2. Be Contextualize: Describe the context where your question is situated or based, the more relevant information about the context provided to GAI, the more accuracy of the generated response you will get.

3. Role play: Telling GAI what role would you want it to play, asking GAI to act on a specific role, the generated output will be more inline/leaning to what that specific role will normally behalf/react to the question you asked.

4. Reciprocal questioning: Instead of keep asking GAI questions in an one-way direction, it is recommended to make it more like a two-way conversational process by asking GAI to raise questions along the way until a final output you satisfied be generated/created.

5. Formatting output: Describe what specific format you would like to see for the GAI generated outputs.

#### Figure 2(b). Level 1 proofread results by ChatGPT

ፊ 🖓 B Here's a revised version of your text with some minor edits: "Here are some guidelines for providing comprehensive prompts to GAI like ChatGPT to get the best responses: 1. Be specific: Describe your goal/question with clear and specific wording, avoiding overly general topics or ambiguous words. 2. Contextualize: Provide relevant information about the context where your question is situated or based. The more contextual information you provide to the GAI, the more accurate the generated response will be. 3. Role play: Tell the GAI what role you would like it to play and ask it to act in a specific role. The generated output will be more in line with how that specific role would typically respond to the question you asked. 4. Reciprocal questioning: Instead of solely asking questions in a one-way direction, encourage a two-way conversational process by asking the GAI to raise questions along the way until a satisfactory output is generated. 5. Formatting output: Describe the specific format you would like to see for the GAI generated outputs."

Can you please improve my original writings to make it look more professional and explain  $\square$  where and why you make those changes?

	Figure $3(b)$ . Level 2 proofread results by ChatGPT		
\$	Here's a revised version of your text with some edits:	ഫ	ጥ
	"Guidelines for Providing Comprehensive Prompts to GAI like ChatGPT to Get the Best Responses:		
	<ol> <li>Be specific: Use specific and clear language to describe your goal or question. Avoid general topics or ambiguous words that could lead to vague responses.</li> <li>Provide Context: Describe the context in which your question is situated or based. The more relevant information you provide to the GAI, the more accurate the generated response will be.</li> </ol>		
	3. Role-Play: Specify the role you want the GAI to play and ask it to act in that role. This will help ensure that the generated output aligns with the expected behavior or response of that specific role.		
	4. Reciprocal Questioning: Encourage a two-way conversational process by asking the GAI to raise questions along the way until a satisfactory output is generated. This will help refine the question and generate a more accurate response.		
	5. Format Output: Specify the format you would like to see for the GAI-generated outputs. This will ensure that the generated output is presented in a manner that is most useful to you."		
	<i>Figure 3(c).</i> ChatGPT's explanations to the Level 2 proofread results		
In this and o	revised version, I made several changes to improve the grammar, sentence struc verall professionalism of the text:	ture	9,

- 1. Improved title: I added a more professional and descriptive title to the guidelines.
- 2. Capitalization: I capitalized the first letter of each word in the title to match standard title capitalization rules.
- 3. Improved phrasing: I revised some of the phrasing to be more concise and professional.
- 4. Corrected typos: I corrected spelling errors, such as "ambiquous" to "ambiguous."
- 5. Added more detail: I added additional detail and explanation to clarify the meaning of each guideline.
- 6. Used active voice: I used active voice throughout the text to make it more engaging and clear.

Level 3 requires the use of a more detailed prompt that includes the task, goal, and output format. For example, in addition to the criteria for "professional" and "explanation," learners can include a description of the desired format, such as "formatted in a table with three columns including the original texts, revised texts, and explanations of changes" as shown in Figure 4(a). In response to this request, ChatGPT would generate a revision summary table, as depicted in Figure 4(b). From the example of Level 3, it is apparent that the GAI-

based learning mode has great potential in promoting students' learning performance and in-depth thinking. As indicated by scholars, knowing [why] is the essential element for learners to have in-depth understanding (Aslan, 2021; Chookaew & Panjaburee, 2022).

## Figure 4(a). Illustrative example of submitting a Level 3 prooferading request to ChatGPT

Can you please improve my original writings to make it look more professional and explain where and why you make those changes? I would also like the output to be formatted in a table with three columns including the original texts, your revised texts and the explanations of changes?

## Figure 4(b). The revision summary table generated by ChatGPT in Level 3 proofreading

Scertainly, I can format the output in a table with three columns as you requested. Here is the revised text with the explanations of the changes:

	Original Text	Revised Text	Explanations of Changes
	Some guidelines for providing comprehensive prompts to GAI like ChatGPT to get the best responses.	Guidelines for Providing Comprehensive Prompts to GAI Like ChatGPT to Get the Best Responses:	Added capitalization and punctuation for title case and changed the sentence structure to make it more clear and concise.
	Be specific: Describe your goal/question with specific and clear wordings, avoid too general topics or ambiquous words.	Be Specific: Use clear and specific language to describe your goal or question. Avoid general topics or ambiguous words that could lead to vague responses.	Reworded for clarity and added capitalization for title case.
	Be Contextualize: Describe the context where your question is situated or based, the more relevant information about the context provided to GAI, the more accuracy of the generated response you will get.	Provide Context: Describe the context in which your question is situated or based. The more relevant information you provide to the GAI, the more accurate the generated response will be.	Reworded for clarity and added capitalization for title case.

## 3.1.2. Content analysis – classifying emotional types

ChatGPT not only can generate learning content for a specified topic, but also can analyze emotional types from the written texts of the story for each paragraph. The following example shows the procedure of creating a story for children and analyzing the emotional type of each paragraph in the story.

Step 1: Provide a very comprehensive prompt to indicate the role of ChatGPT, the goal of the story, target audience and the properties of the generated story, as shown in Figure 5(a). ChatGPT would generate a story following the instructions in the prompt, as shown in Figure 5(b).

Step 2: Instruct ChatGPT to classify the generated story into different emotional types based on the provided coding scheme, with each paragraph classified separately. As shown in Figure 6(a), the emotional types include: (1) Neutral, (2) Happy, (3) Smiling, (4) Excited, (5) Sad, (6) Crying, (7) Depressed, (8) Satisfied, (9) Playful, and (10) Confused. Additionally, request the output to be presented in a table with two columns: the story text and its classified emotional type. The generated table is depicted in Figure 6(b).

Step 3: Ask ChatGPT to generate a more interesting story that covers all the emotional types defined in the coding scheme. An example of such a prompt is shown in Figure 7(a). Following that, ChatGPT generates a new story that meets the criteria, as shown in Figures 7(b) and 7(c).

The same principle and approach can be applied to work with ChatGPT to design better lesson plans, course contents, generate a complete set of question items for quantitative assessment or even develop a rubric for qualitative analysis.

#### Figure 5(a). Example of prompting ChatGPT to generate a story for children

Now, I would like you to play the role of a professional children's book writer and write a story for children between the ages of 5 and 7. The story should be very interesting, with ups and downs to capture the attention and engagement of young readers.

#### *Figure 5(b).* Example of a story generated by ChatGPT

Sure, I would be happy to write a story for children between the ages of 5 and 7. Here's a  $\square \nabla$  story I've come up with:

#### Title: The Magic Marble

Once upon a time, there was a little boy named Jack who loved playing with marbles. He had a special marble that he treasured more than any other. It was a shiny, red marble that glowed in the sunlight. Jack loved it so much that he named it Ruby.

One day, while playing with Ruby, Jack accidentally dropped it into a nearby pond. He searched and searched but couldn't find it. He was heartbroken and missed his precious marble terribly.

That night, Jack had a strange dream. He dreamed that Ruby was alive and had turned into a magical marble. The magical marble had the power to make things happen, but only if Jack found it again.

The next morning, Jack woke up and went to the pond. He searched and searched again, but still couldn't find Ruby. Just as he was about to give up, he spotted a glimmer of red at the bottom of the pond.

With all his strength, Jack dove into the pond and retrieved Ruby. As soon as he held it in his hand, he felt a jolt of magic run through his body. Suddenly, Ruby began to glow and Jack felt a surge of energy.

#### Figure 6(a). Example of prompting ChatGPT to have the generated story classified into different emotional types

Thanks. I would like the generated story to be classified into different emotional types based on the coding scheme I provided, with each paragraph classified separately? The emotional types include: 1. Neutral, 2. Happy, 3. Smiling, 4. Excited, 5. Sad, 6. Crying, 7. Depressed, 8. Satisfied, 9. Playful, and 10. Confused. The output should be presented in a table with two columns: the story text and its classified emotional type.

Figure 6(b). Generated table that classifies the story into different emotional types

\$

Sure, I can classify the story I wrote into different emotional types based on your coding scheme. Here's a table with the story text and its corresponding emotional type for each paragraph:

Story Text	Emotional Type
Once upon a time, there was a little boy named Jack who	Neutral
that he treasured more than any other. It was a shiny	Нарру
One day, while playing with Ruby, Jack accidentally dropped it into a nearby pond. He searched and searched but couldn't find it. He was heartbroken and missed his precious marble terribly.	Sad
That night, Jack had a strange dream. He dreamed that Ruby was alive and had turned into a magical marble. The magical marble had the power to make things happen, but only if Jack found it again.	Playful
The next morning, Jack woke up and went to the pond. He searched and searched again, but still couldn't find Ruby. Just as he was about to give up, he spotted a glimmer of red at the bottom of the pond.	Excited
With all his strength, Jack dove into the pond and retrieved Ruby. As soon as he held it in his hand, he felt a jolt of magic run through his body. Suddenly, Ruby began to glow and Jack felt a surge of energy.	Нарру
He soon discovered that the magical marble could make things happen, but only if he wished for it with all his heart. With Ruby's help, Jack went on many exciting adventures. He wished for a giant slide to appear in his backyard, a rocket ship to take him to the moon, and even a friendly dragon to fly him to far-off lands.	Excited
However, Jack soon realized that the magic marble had a price. Every time he used it, he felt weaker and weaker. He realized that the magic was coming from within him, and that he needed to be careful with his wishes.	Sad
In the end, Jack made one final wish. He wished for the magic marble to go back to being a simple, shiny red marble. As soon as he made the wish, the magic marble disappeared, and Jack felt a surge of energy and strength return to him.	Satisfied
From then on, Jack treasured Ruby even more, knowing that it was the love and appreciation he had for it that made it so special. And although he could no longer make magical wishes, Jack knew that his love for Ruby was the greatest magic of all.	Smiling

Figure 7(a). Prompt to ask ChatGPT to regenerate the story to cover all the emotional types

Based on the analyzed emotional types shown in the table, some of the emotion types in the coding scheme were not covered in your original story. Can you rewrite your original story so that all types of emotion will be included in the new story?

#### *Figure* 7(b). The new story generated in Step 3 (Part 1)

Sure, I'd be happy to rewrite the story to include all the types of emotions in your coding  $\square$   $\square$  scheme. Here's a new version of the story:

Title: The Magic Marble

B

Once upon a time, there was a little boy named Jack who loved playing with marbles. He had a special marble that he treasured more than any other. It was a shiny, red marble that glowed in the sunlight. Jack loved it so much that he named it Ruby.

One day, while playing with Ruby, Jack accidentally dropped it into a nearby pond. He searched and searched but couldn't find it. He was heartbroken and missed his precious marble terribly. (Sad)

That night, Jack had a strange dream. He dreamed that Ruby was alive and had turned into a magical marble. The magical marble had the power to make things happen, but only if Jack found it again. (Playful)

The next morning, Jack woke up and went to the pond. He searched and searched again, but still couldn't find Ruby. Just as he was about to give up, he spotted a glimmer of red at the bottom of the pond. (Excited)

With all his strength, Jack dove into the pond and retrieved Ruby. As soon as he held it in his hand, he felt a jolt of magic run through his body. Suddenly, Ruby began to glow and Jack felt a surge of energy. (Happy)

#### *Figure* 7(c). The new story generated in Step 3 (Part 2)

He soon discovered that the magical marble could make things happen, but only if he wished for it with all his heart. With Ruby's help, Jack went on many exciting adventures. He wished for a giant slide to appear in his backyard, a rocket ship to take him to the moon, and even a friendly dragon to fly him to far-off lands. (Excited)

However, Jack soon realized that the magic marble had a price. Every time he used it, he felt weaker and weaker. He realized that the magic was coming from within him, and that he needed to be careful with his wishes. (Sad)

In the end, Jack made one final wish. He wished for the magic marble to go back to being a simple, shiny red marble. As soon as he made the wish, the magic marble disappeared, and Jack felt a surge of energy and strength return to him. (Satisfied)

From then on, Jack treasured Ruby even more, knowing that it was the love and appreciation he had for it that made it so special. And although he could no longer make magical wishes, Jack knew that his love for Ruby was the greatest magic of all. (Smiling)

As Jack grew up, he kept Ruby safe in a special place where he could always see it. Whenever he looked at the marble, he felt a warm and happy feeling in his heart. (Happy)

Years went by, and Jack grew old. One day, as he was looking at Ruby, he realized that the time had come to pass it on to someone else. He carefully wrapped Ruby in a box and gave it to his granddaughter, Emily. (Neutral)

## 3.1.3. Question item generation and rubric development

Assessment plays an important role in teaching and leanring. A well designed assessment could help teachers identify learners' status and problems, and hence proper assistance or guidance can be provided. A frequently adopted approach for assessment is using multiple-choice items to test learners' knowledge. Figure 8(a) shows an example of prompting ChatGPT to generate a set of multiple-choice items related to the generated story, which will be used in a story reading class for my students; moreover, after the class, the students will take a reading comprehension test about the story and write an essay to describe their feedback and feelings towards the story. Figure 8(b) shows five of the generated items.

Figure  $\delta(a)$ . Example of prompting ChatGPT to generate test items related to the generated story

I am going to use this story in a story reading class for my students. After the class, the students will take a reading comprehension test about the story and write an essay to describe their feedback and feelings towards the story. Could you please help me create an assessment test to assess the students' reading comprehension? The test should cover different aspects of the story with a total of 10 questions.

#### *Figure 8(b).* The test items generated by ChatGPT (Showing the first five items only)

- Sure, I'd be happy to help you create a reading comprehension test for your students. Here's a sample test with 10 questions:
  - 1. What was the name of the little boy in the story?
    - a) Jake
    - b) Jack
    - c) Jim
  - 2. What was the name of the little boy's favorite marble?
    - a) Sapphire
    - b) Diamond
    - c) Ruby
  - 3. How did the little boy feel when he lost his special marble?
    - a) Happy
    - b) Sad
    - c) Confused
  - 4. What did the little boy dream about after he lost his special marble?
    - a) A unicorn
    - b) A talking animal
    - c) A magical marble
  - 5. What did the little boy find at the bottom of the pond?
    - a) His lost shoe
    - b) A treasure chest
    - c) His special marble

Another form of assessing learner's competences and skills is to engage them in completing a project or report, or responding to some open-ended questions. In this form of assessment, a rubric is needed to evaluate their outcomes. For many teachers, developing a rubric is a challenging task. In the following, an example of

prompting ChatGPT to generate a quality rubric is given. In the rubric development process, the first step is to use a prompt to guide ChatGPT to ask the questions necessary to develop a suitable rubric for evaluating the quality of students' essays. It is important to guide ChatGPT to ask one specific question at a time until it has enough information to generate the rubric, as shown in Figure 9(a).

Following the prompt, ChatGPT asks a series of questions to collect information to develop the rubric. Figures 9(b), 9(c), 9(d), 9(e), 9(f), and 9(g) show the series of questions asked following the initial prompt to identify the aspects (dimensions) for assessing the essays, length limitations of the essays, criteria included in each aspect, proficiency levels for each criterion, descriptors included in each level of proficiency in each criterion, and examples for the descriptors in each level.

After collecting the necessary information, ChatGPT generates the rubric for assessing the essays. Figures 10(a), 10(b), and 10(c) show three (i.e., Characters, Plot and Setting) of the aspects in the rubric as well as the proficiency levels (i.e., Excellent, Good, Fair, or Poor) and descriptors for each level of proficiency in each criterion. In this example, the rubric is developed through the collaboration of ChatGPT and a human teacher.

*Figure 9(a).* Example of prompting ChatGPT to ask the questions necessary to develop a suitable rubric

Regarding the rubric for evaluating students' essays, please ask me the questions necessary to develop a suitable rubric for evaluating the quality of students' essays. Please ask one specific question at a time until you have enough information to generate the rubric.

*Figure 9(b).* ChatGPT's question regarding the aspects for evaluating the essays

- Sure, I'd be happy to help you develop a rubric for evaluating the  $\Box \nabla$  quality of your students' essays. Here's the first question:
  - 1. What are the main aspects or elements of the story that the students should focus on in their essays?

Characters, Plot, Setting, Theme and Emotional impact

Figures 9(c). ChatGPT's question regarding the length limitation of the essays

- Great, thank you for that information. Here's the next question:
  - 2. What is the desired length or word count for the students' essays?

800 words

*Figure 9(d).* ChatGPT's question regarding the criteria included in each aspect

Mank you for that. Here's the next question:

3. What specific criteria would you like the rubric to include for each aspect or element of the story? For example, for the Characters aspect, would you like the rubric to evaluate the students' ability to describe the characters' personalities, motivations, or relationships?

## Q.

B

## Just need to include personalities and relationships

Figure 9(e). ChatGPT's question regarding the proficiency levels for each criterion

- Got it, thank you. Here's the next question:
  - 4. What are the different levels of proficiency that you would like to evaluate in each criterion of the rubric? For example, for the criterion of "Describing Characters' Personalities," would you like the rubric to have levels such as "Excellent," "Good," "Fair," and "Poor?"

## Four levels sound good enough.

#### Figure 9(f). ChatGPT's question regarding the descriptors for each level of proficiency in each criterion

Great, thank you. Here's the next question:

5. What are the specific descriptors or characteristics that you would like to include for each level of proficiency in each criterion? For example, for the criterion of "Describing Characters' Personalities," what would be the characteristics of an essay that would receive an "Excellent" rating, as opposed to a "Good" or "Fair" rating?



## *Figure* 9(g). ChatGPT's question regarding the examples of descriptors for each level of proficiency in each criterion

Understood. However, those characteristics are not specific descriptors for each level of proficiency. Could you please provide specific examples of descriptors for each level of proficiency in each criterion, such as for "Excellent," "Good," "Fair," and "Poor?" This will help me to create a more specific rubric.

\$

Excellent: Originality and creativity with in-depth analysis. Good: Use of evidence and critical thinking critique. Fair: Organization and structure with a clear writing style. Poor: None of the above characteristics.

*Figure 10(a).* The Characters aspect of the generated rubric

Thank you for that information. Based on your responses, here is a sample rubric for evaluating your students' essays:

Criteria	Excellent	Good	Fair	Poor
Characters	The essay shows	The essay shows a	The essay shows	The essay lacks an
	an original and	good use of	an organization	original and
	creative approach	evidence and	and structure with	creative approach
	in analyzing the	critical thinking in	a clear writing	in analyzing the
	personalities and	analyzing the	style in analyzing	personalities and
	relationships of	personalities and	the personalities	relationships of
	the characters.	relationships of	and relationships	the characters.
	The descriptions	the characters. The	of the characters.	The descriptions
	are in-depth and	descriptions are	The descriptions	are superficial,
	detailed, providing	clear and well-	are adequate,	providing little
	a clear	supported,	providing a basic	understanding of
	understanding of	providing a good	understanding of	the characters and
	the characters and	understanding of	the characters and	their motivations.
	their motivations.	the characters and	their motivations.	The essay
	The essay	their motivations.	The essay	demonstrates a
	demonstrates a	The essay	demonstrates a	poor
	deep	demonstrates a	basic	understanding of
	understanding of	good	understanding of	the characters and
	the characters and	understanding of	the characters and	their development
	their development	the characters and	their development	throughout the
	throughout the	their development	throughout the	story.
	story.	throughout the	story.	
		s Ġ Regenerate res	ponse	

the story's themesunderstanding ofunderstanding ofand symbolism.and symbolism.the story's themesthe story's themes	Plot	The essay shows an original and creative approach in analyzing the plot of the story. The descriptions are in-depth and detailed, providing a clear understanding of the major events and their significance. The essay demonstrates a deep understanding of the story's themes and symbolism.	The essay shows a good use of evidence and critical thinking in analyzing the plot of the story. The descriptions are clear and well- supported, providing a good understanding of the major events and their significance. The essay demonstrates a good understanding of the story's themes	The essay shows an organization and structure with a clear writing style in analyzing the plot of the story. The descriptions are adequate, providing a basic understanding of the major events and their significance. The essay demonstrates a basic understanding of the story's themes	The essay lacks an original and creative approach in analyzing the plot of the story. The descriptions are superficial, providing little understanding of the major events and their significance. The essay demonstrates a poor understanding of the story's themes and symbolism.
--	------	--	---	--	--

## *Figure 10(b).* The Plot aspect of the generated rubric

## *Figure 10(c)*. The Setting aspect of the generated rubric

Setting	The essay shows an original and creative approach in analyzing the setting of the story. The descriptions are in-depth and detailed, providing a clear understanding of the physical and emotional context of the story. The essay demonstrates a deep understanding of the setting's	The essay shows a good use of evidence and critical thinking in analyzing the setting of the story. The descriptions are clear and well- supported, providing a good understanding of the physical and emotional context of the story. The essay demonstrates a good understanding of	The essay shows an organization and structure with a clear writing style in analyzing the setting of the story. The descriptions are adequate, providing a basic understanding of the physical and emotional context of the story. The essay demonstrates a basic understanding of the setting's	The essay lacks an original and creative approach in analyzing the setting of the story. The descriptions are superficial, providing little understanding of the physical and emotional context of the
	understanding of the setting's significance to the story.	good understanding of the setting's significance to the	understanding of the setting's significance to the story.	
		story.		

#### **3.2. Implications and suggestions**

The above examples deliver an important message that learning higher level GAI usage would generally benefit the learners more. The examples also imply that learners benefit from ChatGPT based on the way they interact with it. Five learner-GAI interactive levels are listed as follows:

- Level 1 None: The learner always waits for the teacher or others' commands or instructions.
- Level 2 A little: The learner always asks wrong questions.
- Level 3 Average: The learner knows how to ask the right questions.
- Level 4 A lot: The learner knows how to ask the right questions in logical sequences using a conversational approach.
- Level 5 Super: The learner treats ChatGPT as a teammate and always works together as a good teammate with ChatGPT.

Accordingly, here are some guidelines for providing comprehensive prompts to GAI like ChatGPT to get the best responses:

- Be specific: Describe your goal/question with specific and clear wording. Avoid using too general topics or ambiguous words.
- Provide context: Describe the context where your question is situated or based. The more relevant information about the context provided to GAI, the more accurate the generated response will be.
- Role-play: Tell GAI what role you want it to play, and ask it to act as a specific role. The generated output will align with how that specific role would typically react to the question you asked.
- Reciprocal questioning: Instead of asking GAI questions in a one-way direction, it is recommended to make it more like a two-way conversational process by asking GAI to raise questions along the way until a final output that you are satisfied with is generated.
- Output formatting: Describe the specific format you would like to see for the GAI-generated outputs.

## 4. Potential research topics of generative AI in education

The diverse roles that GAI can play in educational settings have attracted attention from researchers in both the fields of computer science and educational technology. On the other hand, scholars and school teachers raise several concerns regarding the use of GAI in school settings. That is, GAI in education is not only an interesting issue, but also a challenging one. This implies that GAI in education research has great potential in the coming years since people are eager to know how to use it and what could happen. To this end, the present study proposes the following potential research issues of GAI in education:

(1) *Proposing GAI-based educational models or frameworks*. Individual roles of GAI in education, as addressed in Figure 1, could be further extended by taking into account different educational objectives, computer technologies or devices, research foci, and pedagogical theories to establish GAI-based educational models or frameworks. The models or frameworks could be a good reference for guiding researchers or instructors to implement GAI-based learning strategies, teaching plans or research designs.

(2) Investigating the effects of GAI-based learning on students' learning performance and behavioral patterns. Despite the fact that some researchers and educators worry about the possible misuse of GAI applications by students, the powerful GAI functions are likely to engage students in totally different learning modes, in particular, in creative learning tasks. That is, properly adopting GAI applications in learning designs could shift technology-enhanced learning to a different level, in which students can fully exploit their creativity and application capabilities to create artworks, solve problems, or complete projects with the assistance from GAI. Therefore, it is important and challenging to explore the effective ways of using GAI in educational settings by conducting experimental studies to address this issue.

(3) *Exploring the effectiveness of GAI-based learning from diverse angles.* In addition to comparing the performances of the students learning with GAI and those with the conventional approach, it is also important to probe the effectiveness of GAI-based learning by taking different factors or issues into considerations. For example, it is interesting to investigate the impacts of GAI-based learning on the performances of learners with different cognitive styles or knowledge levels. It is also interesting to reconsider several application domains in which researchers seldom take into account when trying to applying conventional technologies for educational purposes, such as arts, music, and design courses. It is also worth investigating whether the use of GAI could improve students' creative thinking and performances. With the new functions provided by GAI, it is possible to conduct those learning designs that cannot be implemented using conventional AI technologies.

(4) *Revisiting the roles of pedagogical theories by taking the use of GAI in education into account*. It is obvious that GAI provides people with a totally different perspective of what technologies can do. As addressed by Hwang et al. (2020), different educational technologies generally imply angles of perceiving pedagogies. The new features and functions of GAI reveal that educational contexts and objectives need to be perceived from a totally different perspective. Be taking the features of GAI into account, it is possible that those well-known pedagogical theories can be interpreted in a different and innovative way.

(5) Incorporating effective learning strategies into GAI-based learning activities. The effectiveness of GAIbased learning heavily depends on how learners use this new technology to complete their learning tasks. Without proper guidance, bad ways of applying GAI could happen. For example, students could completely rely on ChatGPT to write a report by sending a request to it. To engage students in learning with GAI in a good way, incorporating effective learning strategies into the learning design is very important. In the above example of asking students to write a report using ChatGPT, a possible strategy could be "video sharing," which requires students to record and share how they work with ChatGPT to write their report. Using the peer-assessment strategy, which requests the students to provide ratings and comments to peers' reports based on the rubrics provided by the teacher, could be a good follow-up task of the video sharing activity. Other learning strategies or tools, such as concept mapping and gamification, could also be good choices to facilitate GAI-based learning.

(6) *Examining the effects of different roles played by GAI on students' learning performances.* As mentioned above, GAI applications could play the role of a tutor, tutee or learning partner. It is interesting to investigate the best way of using GAI in different applications domain by comparing the different roles played by GAI.

(7) Constructing ethical guidelines and examples for applying GAI applications to educational settings. From the concerns raised by researchers and school teachers, it is apparent that the issues of using GAI in education is relevant not only to the effectiveness of this approach in promoting learners' performances, but also to the possible ethical problems caused by implementing it in educational settings (Zohny et al., 2023). Without clear guidelines, improper use of GAI could lead to plagiarism or authorship problems. Allowing students to use GAI to complete learning tasks could deliver the wrong messages to them that asking others (e.g., ChatGPT) to do homework or projects for them is acceptable. Therefore, it is essential to take the ethical issues into account when prompting the use of GAI in educational settings.

(8) *Investigating the impacts of diverse Human-GAI collaborative modes*. As mentioned above, GAI could play the role of a tutor, tutee or partner. This implies that the role of treating AI as an intelligent tutor in traditional AIED research could now be very different owing to the presence of GAI. That is, the relationships between learners/instructor and GAI need to be reconsidered. It is possible that, in educational settings, GAI is more like a partner rather than a tutor. In addition, how to use GAI as a tutee to benefit human learners, who play the role of a tutor, remains an open issue.

(9) Exploring the effective ways of employing and evaluating the effectiveness of multimodal GAI applications in education. GAI technologies have advanced from version 1.0 to 4.0. During the process, the functions of a GAI application have significantly changed. For example, GAI 1.0 applications generate text outputs based on users' text and image inputs (e.g., ChatGPT 4.0), GAI 3.0 applications generate text and image outputs based on users' text and image inputs, while GAI 4.0 applications are able to generate any combinations of text, image, and video outputs based on the text and video inputs provided by the users. This raises an interesting question: what are the potential applications and research issues of multimodal GAI applications?

(10) *The fusion of various GAI tools together for personalized learning*. Investigating the integration of AI-generated learning content, AI-driven pedagogical and learning strategies, and AI-simulated educational avatars to create adaptive learning systems that cater to individualized learning experiences. For example, the guidance and support designed in previous studies (Chu et al., 2021) for helping individual learners based on their learning obstacles and special needs, those design mechanisms can be empowered by integrating pedagogical GAI tools.

(11) Integrating embodiment features into GAI by combining GAI with educational robots to promote active social learning. Based on embodied cognition theory and active social learning theory, effective learning requires learners to immerse themselves in a contextually meaningful learning environment and engage in learning activities that involve social interactions using both the brain and body. It is recommended that researchers who have developed learning applications using educational robots (Cheng et al., 2021; Cheng et al., 2023) could explore the possibility of integrating GAI tools using their APIs to make the educational robots behave smarter and more human-like.

## 6. Conclusions

The presence of GAI has shifted the paradigm of AI in education. It is crucial for researchers and educators to perceive GAI from a completely different angle from conventional AI, chatbots, and information systems: (1) GAI is not just an intelligent system that provides personalized guidance or support; (2) GAI is not just a chatbot that can interact with learners via a natural language interface; (3) GAI is not just an application enabling users to search for information. From the examples given in this paper, it is suggested that researchers and educators get rid of the "search" mindset and use "programming prompt" instead to work with GAI applications. The "programming prompt" refers to the ability to guide GAI applications to complete tasks following a sequence of logical instructions. A well-designed set of prompts will guide ChatGPT to perform a quality task. This implies that, from the perspective of education, fostering teachers' and students' competences of "programming prompts" would significantly affect the quality of GAI-based teaching and learning, including the quality of learning content, learning designs, and assessment designs for teachers, and the learning outcomes of students.

Under the notion of programming prompt, there are four categories of programming prompts including (1) Conversational prompts, (2) Content analysis prompts, (3) Coding prompts, and (4) Multimodal prompts. The application examples shown in this paper are examples of conversational prompts. Researchers are encouraged to further explore the design principles and development guidelines to help learners master the knowledge and skills about these four programming prompts to better utilize the benefits that GAI can offer for education.

Although ChatGPT is used to demonstrate various strategies of using GAI in educational settings, the concepts and strategies introduced in this paper can be applied to other GAI applications. It is believed that, in the coming years, more and more GAI applications will be developed and introduced to people and will replace most of the existing computer applications. Therefore, the authors of this paper would like to advise researchers and educators to seriously perceive this new trend of technology-supported education and learn how to correctly and wisely use this new form of computer applications in promoting the educational quality.

## Acknowledgement

This research was supported by the National Science and Technology Council, Taiwan under project numbers MOST 111-2410-H-003 -028 -MY3, 109-2511-H-003-053-MY3 & MOST 111-2410-H-011 -007 -MY3. This work was financially supported by the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

#### References

Aslan, A. (2021). Problem-based learning in live online classes: Learning achievement, problem-solving skill, communication skill, and interaction. *Computers & Education*, 171, 104237. https://doi.org/10.1016/j.compedu.2021.104237

Cheng, Y. W., Wang, Y., Yang, Y. F., Yang, Z. K., & Chen, N. S. (2021). Designing an authoring system of robots and IoTbased toys for EFL teaching and learning. *Computer Assisted Language Learning*, 34(1-2), 6-34, https://doi.org/10.1080/09588221.2020.1799823

Cheng, Y. W., Wang, Y., Cheng, Y. J., & Chen, N. S. (2023). The Impact of learning support facilitated by a robot and IoTbased tangible objects on children's game-based language learning. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2022.2152053

Chu, H. C., Chen, J. M., Kuo, F. R., & Yang, S. M. (2021). Development of an adaptive game-based diagnostic and remedial learning system based on the concept-effect model for improving learning achievements in mathematics. *Educational Technology & Society*, 24(4), 36-53.

Chookaew, S., & Panjaburee, P. (2022). Implementation of a robotic-transformed five-phase inquiry learning to foster students' computational thinking and engagement: A Mobile learning perspective. *International Journal of Mobile Learning and Organisation*, *16*(2), 198-220. https://doi.org/10.1504/IJMLO.2022.121888

Dehouche, N. (2021). Plagiarism in the age of massive generative pre-trained transformers (GPT-3). *Ethics in Science and Environmental Politics*, 21, 17-23.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities,

challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Eysenbach, G. (2023). The Role of ChatGPT, generative language models, and artificial intelligence in medical education: A conversation with ChatGPT and a call for papers. *JMIR Medical Education*, 9(1), e46885. https://doi.org/10.2196/46885

Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in Education. *Computers and Education: Artificial Intelligence*, 1, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Yang, S. J., Ogata, H., Matsui, T., & Chen, N. S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, 2, 100008. https://doi.org/10.1016/j.caeai.2021.100008

Zohny, H., McMillan, J., & King, M. (2023). Ethics of generative AI. Journal of Medical Ethics, 49(2), 79-80.

## Effects of a Hybrid Training for Plagiarism Prevention Module on Plagiarism-free Academic Writing in Higher Education

## Yin Zhang<sup>1\*</sup>, Samuel Kai Wah Chu<sup>2</sup>, Yonghui Liu<sup>3</sup> and Xiaoli Lu<sup>4</sup>

<sup>1</sup>Department of Education, Ocean University of China, Qingdao, P. R. China // <sup>2</sup>Faculty of Education, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, P. R. China // <sup>3</sup>College of Engineering, Ocean University of China, Qingdao, P. R. China // <sup>4</sup>School of Mathematical Sciences & Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai, P. R. China // zhangyinouc@sina.com // samchu@hku.hk // liuyonghui@ouc.edu.cn // xllu@math.ecnu.edu.cn

\*Corresponding author

(Submitted February 5, 2022; Revised July 19, 2022; Accepted August 18, 2022)

**ABSTRACT:** Previous research has looked into educational approaches to prevent plagiarism in academic writing, yielding insights into how plagiarism can be avoided. However, plagiarism remains a major problem in the education sector. We designed a training module that includes a customised Online Scaffolding Writing System (OSWS) to help faculty teach undergraduates how to avoid committing plagiarism in their academic writing. A quasi-experimental design was used to analyse the plagiarism-related perceptions and behavioural changes of 121 undergraduate students and to test the effects of the new module on students' academic writing. The experimental group performed significantly better than the control group in terms of decreasing the extent of plagiarism in their writing (with a mean decrease from a moderate to minor level of plagiarism), and improving their writing quality (with a mean increase of 18 percentage points in writing scores). Furthermore, more than 95% of the students in the experimental group and their instructor reported that they valued the benefits of adopting the training module in class, and almost 90% of them expressed high levels of satisfaction with the learning they had obtained from the OSWS. This study also provides insights into how the new training module can be implemented across disciplines.

Keywords: Plagiarism, Hybrid training, Academic writing, Online Scaffolding Writing System (OSWS)

## **1. Introduction**

Plagiarism is the act of appropriating others' ideas, language or writing without proper acknowledgement (Vessal & Habibzadeh, 2007) and is a major problem in higher education (Eaton, 2021). The increasing prevalence of plagiarism on campus endangers the academic integrity of educational institutions and poses a threat to the quality of higher education (Hopp & Speil, 2021). As such, many institutions around the world have adopted various policies to punish those who commit plagiarism; these policies include informal or formal warnings, grade penalties, suspension, or expulsion (Tremayne & Curtis, 2021). However, plagiarism prevention in higher education is markedly different from that in other fields, and it is unwise to punish student plagiarists without educating them on the topic (Mphahlele & McKenna, 2019). Furthermore, empirical studies have demonstrated that punitive policies work by instilling in students the fear of being caught, but fail to help students learn from their mistakes (Parks et al., 2018). Studies have also provided robust evidence that undergraduates typically commit plagiarism unintentionally and that the expulsion of student plagiarists could deprive them of the opportunity to be educated about plagiarism prevention (Pecorari & Shaw, 2018; Zhang & Tang, 2021).

Thus far, researchers have generally agreed that all relevant stakeholders should participate in efforts to curb plagiarism (Uzun & Kilis, 2020), and a substantial amount of empirical studies have produced a rich array of evidence to support the design of instructional materials, instruments, and strategies for plagiarism prevention in higher education (Lee et al., 2016; Tindall & Curtis, 2020). Information and communication technologies (ICTs) have also been harnessed to address plagiarism, such as through the use of plagiarism detection software (PDS) and/or ICT-supported anti-plagiarism instruction (Batane, 2010).

However, the increase in the number of reported plagiarism cases on campuses worldwide indicates that the problem of plagiarism has not yet been solved (Roller, 2018). Researchers have discovered that previous educational approaches have focused only on the act of plagiarism and have neglected students' learning needs for anti-plagiarism instruction (Pàmies et al., 2020). Some researchers have suggested approaches that they consider more responsive to students' learning needs, such as teaching students about plagiarism prevention in academic writing to facilitate plagiarism-related learning conditions and experiences (Hu & Lei, 2016). According to Hofer et al. (2012), plagiarism is a typical threshold concept in students' academic writing practise. That is, students' poor understanding of plagiarism may affect their critical analysis and understanding of the

literature and thereby hamper their writing; by contrast, students' rich understanding of plagiarism could give them a transformative understanding of academic writing and help them to construct individual knowledge on and skills in plagiarism-free academic writing. Therefore, teaching plagiarism prevention is an essential component of teaching academic writing.

However, there are gaps in this area that need to be addressed. Although researchers have high expectations of academic writing instruction, there is a lack of empirical research exploring how to orchestrate the teaching of plagiarism prevention in academic writing contexts. Moreover, the literature has rarely discussed the possibilities afforded by ICT in scaffolding plagiarism prevention instruction and learning (instead of detecting plagiarism). To fill these research gaps, we designed a Hybrid Training for Plagiarism Prevention (HTPP) module applicable to academic writing, where "hybrid" refers to a combination of face-to-face and online teaching. The online teaching component of HTPP is supported by a customised ICT tool, i.e., the Online Scaffolding Writing System (OSWS). We performed a quasi-experimental study to determine the ability of this new module to help undergraduate students generate plagiarism-free academic writing and to determine how useful the students and their instructor found the new module: experimental group students finished writing assignments with the proposed module, whereas the control group students finished the same writing assignments without the module. The following research questions (RQs) were addressed in this study.

RQ1: What are the effects of the HTPP module on the students' perceptions of plagiarism?

- RQ2: What are the effects of the HTPP module on the students' writing performance, in terms of their writing quality and the level of plagiarism in their writing?
- RQ3: What are the students' and their instructor's perceptions of the HTPP module?

## 2. Literature review

The following paragraphs provide an overview of previous research on plagiarism prevention in higher education and anti-plagiarism instruction in academic writing, which illustrates the rationale of the current study.

#### 2.1. Subverting plagiarism in higher education

Plagiarism is a serious problem among university students worldwide (Roller, 2018). Researchers from different disciplines have investigated this phenomenon and have recommended various approaches to prevent it (de Maio et al., 2020; Parks et al., 2018). This study classifies the current instructional solutions as follows: detect to punish, detect to evaluate, and educate to learn.

The "detect to punish" solution is frequently used by many universities when dealing with plagiarism. For example, according to a survey of 93 institutions in the UK, 143 students were expelled from campus because of plagiarism (Attwood, 2008). In 2019, two of China's most prestigious universities, Tsinghua University and Peking University, announced that they would expel students for plagiarism (Xinhua, 2019). However, some researchers have argued that the dismissal of student plagiarists might not be an appropriate solution (Schinkel, 2015); some have even suggested that such a harsh punishment might have an overall negative impact (Davies & Howard, 2016), a view that has been supported by empirical findings. For instance, Abasi and Graves (2008) found that some students over-cited in their manuscripts, i.e., included a string of references for each sentence, in the hope of reducing the likelihood that their work would be flagged for plagiarism. This indicates how the fear of punishment may push students to focus on strategies for escaping punishment rather than on improving their ability to write plagiarism-free pieces.

"Detect to evaluate" is another common approach that is adopted by faculty members to help them distinguish between students' original contributions and borrowed ideas. For instance, Mostert and Snowball (2013) reported that PDS may help detect plagiarism and that it provides evidence to faculty members to allow them to take instructional measures. However, others argued that the functional design of PDS may be faulty in that it uses text matching to identify plagiarism, which decreases its validity in interpreting the extent of plagiarism in a written piece (Mphahlele & McKenna, 2019).

Unlike the above two plagiarism-prevention methods, which are reactive, the "educate to learn" approach addresses the problem in a proactive manner. It consists of information science instructors providing information ethics courses for students (Liu & Yang, 2012) and instructors assigned to writing centres educating students about procedures related to anti-plagiarism (Chu et al., 2021). Blum (2011) stated that plagiarism is a mere

symptom of a larger educational problem, which, if so, means it is essential to prioritise the educate to learn approach. From this, it can be inferred that the educational problem underlying plagiarism deserves due attention from researchers to improve the efficacy of anti-plagiarism measures.

#### 2.2. Teaching anti-plagiarism in academic writing

Academic writing is at the heart of university undergraduates' study programmes. Students may encounter various difficulties related to academic writing, among which plagiarism is the most common (Löfström et al., 2017). In a survey conducted by the International Center for Academic Integrity across five American universities, 25.1% of undergraduates admitted using unauthorised electronic resources when completing their papers or other written assignments (International Center for Academic Integrity, n.d.). There is also concern that the number of plagiarism cases may continue to increase if students do not receive timely and appropriate instructional support (Harris et al., 2020). The increasing incidence of plagiarism highlights the urgent need to teach students how to avoid committing plagiarism in academic writing. Such instruction can usually be provided by writing instructors, other faculty members, or a collaboration of faculty members and librarians (Awasthi, 2019).

Plagiarism prevention instruction in academic writing comprises two parts: instruction on plagiarism and instruction on how to cite sources and thus avoid committing plagiarism when writing (Pecorari & Petrić, 2014). Educational interventions may include discussions of plagiarism cases from specific disciplines (Brown & Janssen, 2017) or instructions regarding the writing skills needed to avoid plagiarism (Du, 2019). However, some limitations remain unaddressed. For instance, in one study, although most students could define plagiarism, few could identify plagiarised texts (Leung & Cheng, 2017). This is attributable to faculty's one-sided understanding of plagiarism (Greenberger et al., 2016). Plagiarism prevention has been thought to be easily taught in a traditional in-class way by faculties (Myers, 2018); however, it is difficult to teach because it requires an understanding of why students plagiarise in addition to how they perceive and cope with plagiarism (Peled et al., 2019).

A few studies have highlighted some of the additional limitations to current anti-plagiarism instruction from faculties' perspectives. For example, a multi-institution writing project was conducted to help writing faculty deliver instruction on plagiarism-free writing to American undergraduates (Jamieson, 2017). However, although the project urged writing faculty to care about each students' behavioural characteristics in writing practise and provide corresponding face-to-face instruction, some faculty were reluctant to do so because they felt it was too time-consuming and labour-intensive. Moreover, some writing professionals and subject experts have reported that their teaching of writing skills to prevent plagiarism mainly relied on their own writing experiences, which indicated that the quality of this teaching largely depended on these instructors' own knowledge and experiences (Tomaš, 2010; Huang, 2017).

The literature review presented above provides several key insights and highlights research gaps. First, students cannot easily gain practical knowledge about plagiarism if their faculty has a one-sided understanding of plagiarism and adopts the traditional in-class approach of instruction. Second, traditional instruction on plagiarism prevention in academic writing is labour-intensive and time-consuming. Third, previous instruction has been designed for students in classroom settings, which limits the time students have to practise their anti-plagiarism skills. To fill these gaps, the HTPP module was designed and tested in the current study to help on-campus instructors teach students how to produce plagiarism-free academic writing.

#### 2.3. Theoretical framework of the HTPP module

The features of Teaching for Understanding, Hybrid Learning, and Group and Learning Dynamics theories were incorporated into an integrated conceptual framework for the HTPP module (Figure 1). These three theories are highly related to constructivist learning, which underpins the core mission of the HTPP module, i.e., to facilitate students' construction of knowledge on plagiarism-free academic writing by engaging them in instructional activities (Figure 1, middle panel). The instructional activities specified in the concave-cornered rectangles were conducted online, whereas those specified in the round-cornered rectangles were conducted in class.

Teaching for Knowing theory focuses on rote learning, whereas Teaching for Understanding theory focuses on improving students' understanding from the level of remembering to the level of performance (Wiske & Breit, 2013). Our literature review on plagiarism instruction in academic writing shows that for plagiarism to be

eliminated, students' understanding of plagiarism at the practical performance level needs to be prioritised. Thus, our use of Teaching for Understanding theory in the current study was justified.





Once a learning objective is clearly identified, instructional activities need to be carefully designed to intensify instruction (Hutchison & Woodward, 2014). Hybrid learning has three key features: intensified instruction, improved flexibility, and enhanced cost-effectiveness (Law et al., 2019). These features match the instructional needs of the HTPP module. Therefore, a customised ICT tool—OSWS—was developed by the first author to support the construction of a hybrid learning environment with three instructional purposes: to intensify plagiarism instruction by supporting faculty in their supervision of students' writing processes and provision of timely feedback; to improve flexibility by providing students with enhanced temporal and geographic flexibility to complete their academic writing tasks and communicate with peers; and to enhance cost-effectiveness by delivering plagiarism instruction within disciplinary courses and scaffolding students' learning about plagiarism during their course writing assignments.

Task cooperation may reduce academic dishonesty by strengthening conscientiousness among students (Peled et al., 2019), which supports the adoption of peer interactions in the design of the HTPP module. Moreover, we used Group and Learning Dynamics theory to explore how interactions between students can be facilitated in a hybrid learning space. It has been reported that when students successfully collaborate with one another in hybrid learning contexts, the success of their interactions is attributable to both positive group dynamics (e.g., reflection and feedback) and positive learning dynamics (e.g., the building of ideas and meta-communication) (AlSheikh & Iqbal, 2019).

## 3. Methodology

## 3.1. Instructional and experimental procedures

#### 3.1.1. OSWS supports the hybrid instructional procedure

The theoretical framework of the HTPP module was used to design instructional strategies and activities to meet the intended learning objective: to give students' an adequate understanding of plagiarism. Peer review and peer discussions are complementary peer interaction strategies for academic writing instruction (To & Carless, 2016), and a combination of written peer review and oral peer discussions can facilitate deep interaction between peers by enabling reviewers and reviewees to clarify their writing and receive feedback (Hadwin et al., 2017). In the current study, we employed written peer review and oral peer discussions as the key scaffolding strategies: the students were expected to identify plagiarism and poor writing in their peers' writing during peer review and exchange ideas to solve the identified problems during peer discussion. Compared with face-to-face peer review, online peer review provides greater feedback and improves writing performance to a greater extent (Awada & Diab, 2021). Therefore, we conducted online written peer review supported by OSWS.

The HTPP module was designed and introduced to both the students and their instructor during the academic writing process in their subject courses. When the students were assigned a writing task, they prepared their writing based on the academic material provided by their course instructor. After submitting their first drafts, they followed the instructional procedure of the HTPP module to revise their writing. First, the students learned about the process and criteria of peer review by reviewing three examples of writing with varying levels of plagiarism, after which they compared their review results with those of the instructor. Then, the students were divided into groups and reviewed their group members' writings. According to a previous study, each peer-review group should include three to four students (Reinholz, 2016); thus, students were asked to review the writing of two of their group members and evaluate the writing using the criteria introduced in the peer-review training. Second, after receiving peer and instructor reviews on their writing, the students were given a chance to discuss face-to-face with their peers in class regarding the plagiarism problems identified in the peer reviews. Third, the instructor provided 1 h of instruction based on the plagiarism detected by the OSWS and the students' concerns about plagiarism and writing problems. Finally, the students reflected on own writing, revised it if necessary, and submitted it for assessment.



Online Scaffolding Writing System	
<b>教学系统设计</b> (Course: In へ人主页 ▶ 課程 ▶ 我的課程 ▶ 数学系统设计 导航 (Course navigation)∋□ へ人主页	structional design) 前言 (Preface)
<ul> <li>网站首页</li> <li>网站页面</li> <li>我的课程</li> <li>课程</li> <li>求的课程</li> </ul>	<ul> <li>欢迎同学们未到"在线学术写作系统"完成学术写作任务。学术写作是一项重要的学习活动,它不仅需要你理解学科知识,开展对于学术话题的批判性思考,而且需要你在综合评价现有相关文献的基础上通过文本创作呈现出对于某学术主题的分析和评论。</li> <li>* 学习活动即将开始,你做好准备了吗? (Brief introduction)</li> </ul>
<ul> <li>▶ 教学系统设计</li> <li>▶ 2018春季课程</li> </ul>	■ 写作资源和写作要求 (Writing resources and requirements)
系統管理 □ 3 ▼ 课程管理 将我加入此课程 → 题库	<ul> <li>「「」」「「」」」」」</li> <li>「」」「」」」」</li> <li>「」」「」」」」</li> <li>「」」「」」」」</li> <li>「」」」」</li> <li>「」」」」」</li> <li>「」」」」</li> <li>「」」」</li> <li>「」」</li> <li>「」</li> <li>「」」</li> <li>「」</li> <li>「」</li> <li>「」</li> <li>「」</li> <li>「」</li> <li>「」</li> <li>「」</li></ul>
▶ 网站管理	抄發分析模块 (Fiagianisin analysis)

Aside from the face-to-face peer discussions and in-class lectures, most of the learning activities took place in an online learning environment supported by the OSWS, which is a key component of the HTPP module. We had

considered whether current ICT tools could meet the above-mentioned training needs. Previously created ICT tools can be used for plagiarism detection (e.g., CrossCheck and WCopyFind), writing practise (e.g., Criterion and WriteToLearn), or both (e.g., Turnitin and Glatt) (Liu et al., 2013). However, users of plagiarism-detection tools may receive similarity scores for their writing but receive no feedback on how to improve their writing (Zhang et al., 2019). Similarly, users of writing tools may practise their writing without learning how to correctly incorporate aspects from sources into writing. In addition, because these tools are tailored for self-learning, users may be limited by the lack of an opportunity to learn from others. The OSWS was developed by the first author as a solution to the above-mentioned limitations, the term "scaffolding" in the full-form of the OSWS indicates that the system was designed to support students to practise plagiarism-free academic writing.

The OSWS is based on Moodle, a widely used open-source learning content management system. The online learning activities were arranged based on the timeline of the instructional procedure. Some of the features of the OSWS were designed based on Moodle's pre-designed functions; for example, the "written peer review" function was built using the "workshop" activity in Moodle (see the icon in the red rectangle in Figure 2). The OSWS also includes customised functions, such as the "plagiarism analysis" tool. This function was established to help instructors identify plagiarism in students' writing.

#### 3.1.2. Experimental procedure

This 8-week-long quasi-experiment was performed in the spring of 2018. Figure 3 presents the experimental procedure of this study. Before the writing assignments, two groups of students completed a pre-activity questionnaire on their perceptions of plagiarism. The students were then assigned a writing task and instructed to write pieces based on the provided academic material. Once the students had finished the first draft of their writing, each group revised their first drafts in different ways: the experimental group followed the HTPP module procedure (which consists of an online written peer review, face-to-face peer discussions, and in-class lectures), whereas the control group followed the conventional approach, i.e., they received instructional feedback from their instructor regarding plagiarism and other problems in their first draft. After both groups had revised their first drafts and submitted them for assessment, they took a post-activity questionnaire to examine possible perceptional changes during the experiment as well as perceptional differences on plagiarism between the two groups. Each groups' writing was assessed and compared in terms of two aspects—its level of plagiarism and its quality—to reveal improvements in students' writing performance. At the end of the experiment, the experimental group students completed a feedback questionnaire, and 50% of the students (randomly selected) attended an interview to share their views on the usefulness of the HTPP module.



#### **3.2. Study participants**

The participants were recruited using convenience sampling (Creswell, 2012). The target population comprised first-year undergraduates from two different classes (mean age, 19 years) and their instructor Ms Z (This is a pseudonym to ensure the instructor's anonymity). from the education department of a public university in Chinese Mainland. The students' demographic data are shown in Table 1. The sample comprised 121 participants, with 66 in the experimental group and 55 in the control group. The participants were segregated based on sex [97 (80.2%) girls, 24 (19.8%) boys], and the sex distribution in each group was similar. Moreover, both groups had few prior experiences of learning about plagiarism. The HTPP module was introduced to the experimental group as an academic writing project tied to writing assignments in a subject course called "instructional design." By contrast, the control group undertook the same writing assignments but without using the module.

Table 1. Participants' demographic data			
Sex	Experimental Group (%)	Control Group (%)	Total (%)
Male	13 (10.7)	11 (9.1)	24 (19.8)
Female	53 (43.8)	44 (36.4)	97 (80.2)
Anti-plagiarism learning experience			
During university education	0 (0)	1 (0.8)	1 (0.8)
Before entering university	0 (0)	3 (2.5)	3 (2.5)
None	66 (54.6)	51 (42.1)	117 (96.7)
Total	66 (54.6)	55 (45.4)	121 (100)

#### **3.3. Instruments**

The ability of the HTPP module to prevent students from committing plagiarism in their academic writing was examined by comparing the experimental group students' perceptional and behavioural changes regarding plagiarism-free academic writing with those of the students in the control group. Moreover, the students' and their instructor's views on the HTPP module were assessed by analysing data from the feedback questionnaires and interviews.

#### 3.3.1. Perceptions of plagiarism questionnaire

The questionnaire was based on the "knowledge and attitudes to plagiarism" questionnaire developed by Lee et al. (2016) and assessed 13 items. The first three items were students' self-evaluations of their abilities to write without committing plagiarism, while the remaining 10 items were students' self-evaluations of their abilities to distinguish plagiarism in various scenarios. All of the questions were checked by the second author and an academic writing expert to ensure the validity. Cronbach's alpha was 0.74, indicating that there was reasonable internal consistency between the items (Robinson et al., 1991).

#### 3.3.2. Feedback questionnaire

The feedback questionnaire comprised two parts. One part sought the students' views on the effectiveness of the HTPP module, whereas the other sought their views on the usefulness and usability of OSWS. The first part was based on a questionnaire developed by Lee et al. (2016), whereas the second part was adapted from a questionnaire developed by Liu et al. (2013). The feedback questionnaire contained 19 questions, which were checked by the second author and an academic writing expert to ensure validity. The Cronbach's alpha was 0.93 for the HTPP module and 0.91 for the OSWS, demonstrating good internal consistency (Robinson et al., 1991).

#### 3.3.3. Writing assignment

The effects of the HTPP module on the students' behaviour were examined by analysing two dimensions of their submitted writing: its quality and its level of plagiarism. The requirements for the writing assignment were based on Lu's (2013) writing assignment design. The writing quality was evaluated using a four-dimensional rubric based on the studies of Liu et al. (2013) and Choi (2012) (detailed information is provided in Appendix A). The students' writing was independently rated by the first author (rater 1) and the students' instructor (rater 2); the

Pearson correlation co-efficient (r = 0.83) was calculated to confirm the presence of inter-rater reliability (Benesty et al., 2009).

The extent of plagiarism in writing is usually evaluated based on two widely used plagiarism assessment measures—the degree of similarity and the number of consecutively copied words. However, both measures have some limitations. The validity of the degree of similarity measure is often questioned, and many universities (e.g., University College London, 2019) consider any degree of similarity to be unacceptable. The number of consecutively copied words is frequently used by research associations to evaluate plagiarism; however, the threshold for determining plagiarism varies widely under different academic conditions (Masic, 2012). Therefore, in this study, the extent of plagiarism in writing was evaluated using a newly designed plagiarism assessment scale (see Table 2). The scale was customised based on the plagiarism assessment criteria of Chu et al. (2021) and Yeung et al. (2018); this scale had been used by a group of university students over an academic year, and its validity was carefully examined by the authors of the present study and an academic writing expert. The first author (rater 1) and a research assistant (rater 2) independently rated the levels of plagiarism, and Spearman's co-efficient (r = 0.93) was calculated to confirm the presence of inter-rater reliability. The level of plagiarism in each piece of writing was determined by averaging its two ratings.

Level	Label	Description
Level 1	None	No plagiarism.
Level 2	Minor	• Copying a block of text, which is greater than <i>x</i> and less than <i>y</i> Chinese characters, from a source, rearranging its phrases, adding words and replacing words with synonyms, and not providing a citation.
Level 3	Moderate	• Copying a block of consecutive Chinese characters, which is greater than <i>x</i> and less than <i>y</i> , from a source without providing a correct quotation, or
		• Copying a block of text of over <i>y</i> Chinese characters from a source without providing a citation but providing a reference at the end of the work.
Level 4	Serious	• Copying more than <i>y</i> consecutive Chinese characters from a source without providing a correct quotation, or
		• Copying a block of text of over <i>y</i> Chinese characters from a source without providing a citation.

*Note.* The minimal phrase match is denoted by x, which was six Chinese characters for student writing, based on previous research (Kostoff et al., 2006); the mean score of the most consecutively copied words in each piece of writing is denoted by y, which was calculated to be 85 Chinese characters in the experiment (using the plagiarism analysis module in the OSWS).

#### 3.3.4. Interviews

The student interviews were conducted according to student interview procedure of Lu (2013); thus, 50% of the students (n = 33) were randomly selected and invited to interviews at the end of the experiment. However, because several students were reluctant to attend the interview, convenience sampling was used to invite students, which resulted in 23 volunteering to be interviewed. In addition, the instructor was interviewed to collect her instructional experiences in using the HTPP module, with the interview based on the faculty interview procedure of Grigg (2016). For the convenience of the interviewees, all interviews were conducted at the interviewees' university after the experiment was completed.

#### 3.4. Data analysis

We used different statistical tests to analyse the quantitative data. If the data met the requirements for normal distribution and the variance homogeneity conditions, an independent samples *t*-test was used to compare the data of the experimental group with that of the control group; if not, the Mann–Whitney *U* test was used to make this comparison (Hoy & Adams, 2015).

The content analysis method was used to analyse the interview data. To ensure the reliability of the interviews, 50% of the interview data were separately coded by the first author and a research assistant into themes and sub-themes, and active discussions were conducted until the inter-rater agreement reached a level of 83%, indicating a satisfactory reliability (Stemler, 2004). Then, the remaining part of the interview data were coded by the first author alone.

## 4. Results

#### 4.1. Changes in students' perceptions of plagiarism

Data on students' perceptions of plagiarism were collected before and after the writing assignment. The data collected from the pre-activity questionnaires revealed that the two groups had similar views on plagiarism. Moreover, although most students had no prior experience learning about plagiarism prevention, they provided high ratings to some items. For instance, most students believed that they had "a good understanding of the importance of avoiding plagiarism," and the mean rating for the item was above 5. In addition, many of the students could easily identify typical explicit examples of plagiarism, such as "adding others' writing into my writing assignments without acknowledging the source," which had a mean rating of above 4.5.

Table 3. Perceptions of plag	iarism in the post-act	ivity questionnaire		
Questionnaire items	Mean	Mean (SD)		
	Experimental group $(N = 61)$	Control group $(N = 43)$	р	
(1) I am capable of identifying plagiarism- involving cases.	5.11 (0.99)	5.28 (0.63)	.588	
(2) I am capable of avoiding plagiarism.	4.28 (0.93)	4.30 (0.99)	.454	
(3) I have a good understanding of the importance of avoiding plagiarism.	4.13 (0.94)	4.44 (0.85)	.045*	
<ul><li>(4) Adding others' writing into my writing assignment without acknowledging the source is plagiarism.</li></ul>	4.57 (1.27)	4.49 (1.18)	.548	
(5) Paraphrasing others' writing, and adding it to my writing assignment without acknowledging the source is plagiarism.	4.92 (1.22)	4.53 (1.20)	.060	
(6) Incorporating others' ideas (not writing) into my writing assignment without acknowledging the source is plagiarism.	4.33 (1.59)	2.91 (1.41)	$.000^{*}$	
(7) Incorporating teachers' course materials into my writing assignment without acknowledging the source is plagiarism.	4.56 (1.46)	4.09 (1.41)	.053	
<ul><li>(8) Using my previous writing assignment for the current one without acknowledging the source is plagiarism.</li></ul>	4.93 (1.15)	4.51 (1.26)	.053	
<ul><li>(9) Introducing Internet materials without source information in my writing assignment and submitting it in my name is plagiarism.</li></ul>	5.10 (1.01)	5.09 (0.72)	.482	
(10) Inserting an Internet picture into my writing assignment without acknowledging the source is plagiarism.	4.18 (1.59)	3.51 (1.45)	.024*	
(11) Incorporating data that has no author's information into my writing assignment without acknowledging the source is plagiarism.	4.07 (1.52)	3.33 (1.39)	.014*	
(12) Inserting an author's conclusion into my writing assignment without acknowledging the source is plagiarism.	4.67 (1.11)	4.19 (1.20)	.042*	
(13) Even though I have helped my friend finish his/her writing assignment, it is plagiarism if I copy his/her manuscript and hand it in under my name.	4.92 (1.23)	4.91 (0.97)	.504	

*Note.* Seventeen students failed to complete the questionnaire, so 104 completed questionnaires were obtained. The ratings are based on a 6-point Likert scale ranging from 1(*strongly disagree*) to 6 (*strongly agree*). \*p < .05.

Some significant differences were found when the two groups' perceptions of plagiarism were compared again at the end of the experiment (see Table 3). Compared with the control group students, the experimental group students provided higher self-ratings for 10 of the 13 items, four of which had statistical differences (see Q6,

Q10, Q11, and Q12 in Table 3). The comparison revealed that many of the experimental group students could identify typical implicit plagiarism problems, such as "incorporating others' ideas into my writing assignments without acknowledging the source." However, it was surprising that both groups had lower self-ratings on their understanding of the importance of avoiding plagiarism after the activity than before (see Q3 in Table 3) and that the self-ratings of the experimental group were even lower than those of the control group in the post-activity questionnaire.

#### 4.2. Students' behavioural changes in response to plagiarism

Before the HTPP module instructional intervention was delivered, two dimensions of the students' written pieces were analysed: their extent of plagiarism and their quality of writing. Most had serious plagiarism problems: the mean level of plagiarism was about Level 3, indicating a moderate level of plagiarism based on the Plagiarism Assessment Scale (Table 2), and the mean value of the most consecutively copied words was 85 Chinese characters. In addition, the students' writing quality was poor, with a mean writing quality score of 58 points based on the four-dimensional rubric (Appendix A). After the intervention, the experimental group significantly improved in terms of both their plagiarism level and writing quality: the mean plagiarism was Level 2, and the mean writing quality score of 76 points. The Mann–Whitney U test and independent samples t-test results showed there were significant differences between the groups in terms of their pieces' level of plagiarism and writing quality (see Table 4).

Measure Mann-Whitney U test Mean (SD) Experimental group Control group Uр (N = 46)(N = 64)I. Extent of plagiarism (1) Level of plagiarism 2.34 (1.04) 3.00 (1.05) 974 .002\* Independent t-test t .022 (2) Most consecutively copied 52.25 (46.42) 74.33 (5.057) -2.337 Chinese characters *II. Writing quality* Mann-Whitney U test Uр 1.217.5 (1) Assignment response 17.81 (4.89) 16.25 (3.06) .114 (2) Coherence and cohesion 22.58 (3.67) 19.08 (2.90) 607.5  $.000^{*}$ (3) Vocabulary and language use 17.50 (3.33) 1,166.5 .059 19.22 (4.18) 1,109 .025\* (4) Citation 16.41 (8.44) 14.40 (6.92) 76.02 (16.58) 67.23 (13.85) Total

Table 4. Students' writing performance after the intervention (the HTPP module)

*Note.* Eleven students failed to submit their writing, and 110 pieces of writing were collected. p < .05.

#### 4.3. Students' and instructor's feedback about the HTPP module

The HTPP module was presented to the experimental group students as an academic writing project. At the end of the experiment, the students' opinions regarding their learnings from the project were solicited. Based on the data collected from the feedback questionnaire, most of the students (> 95%) indicated a high level of satisfaction with the HTPP module, with all ratings above the mid-point (3.5) on the 6-point Likert scale (Table 5). Many of the students considered the module useful for improving their ability to identify plagiarism, avoid plagiarism, and become aware of the importance of avoiding plagiarism (see the high mean scores for O1, O2, and O3). Moreover, most of the students believed that their enhanced knowledge regarding paraphrasing. summarising, synthesising, and in making citations, facilitated their ability to avoid plagiarism (see the high rates of agreement for Q9, Q10, and Q11). Although the students' scores for the usefulness and usability of the OSWS were lower (see Q12–Q19) than their ratings on the effectiveness of the module, more than four out of five students valued the learning gained from using the OSWS. Most of the students considered that the OSWS was easy to use (see Q15) and were satisfied with the usefulness of the peer review process for facilitating communication between peers and decreasing plagiarism (see Q18 and Q19).

<i>Table 5.</i> Student feedback on the HTPP Module		
Questionnaire items	Mean (SD) (N = 64)	Proportion of agreement
I. The effectiveness of the Academic Writing (AW) project <sup>#</sup>		
Understanding Plagiarism		
(1) I am more capable of identifying plagiarism cases after completing the AW project.	4.81 (.84)	94.9%
(2) I am more capable of avoiding plagiarism after completing the AW project.	4.97 (.74)	98.3%
(3) I have a better understanding of the importance of avoiding plagiarism after	5.03 (.74)	96.6%
completing the AW project.		
Understanding academic writing		
(4) I am more capable of expressing others' ideas in my own words (i.e., paraphrasing) after working on the AW project.	4.61 (.70)	96.6%
(5) I am more capable of presenting the key information as a concise statement (i.e., summarizing) after working on the AW project.	4.53 (.73)	96.6%
(6) I am more capable of distinguishing paraphrasing, summarizing and patch- writing after working on the AW project.	4.59 (.65)	98.3%
(7) I am more capable of integrating several source materials with my own ideas (i.e., synthesizing) after working on the AW project.	4.68 (.68)	96.6%
<ul><li>(8) I am more capable of producing proper citations after working on the AW project</li></ul>	4.81 (.68)	98.3%
Others		
<ul> <li>(9) Due to gaining a better understanding of plagiarism by completing the AW project. I am more capable of avoiding plagiarism in my work.</li> </ul>	4.80 (.81)	94.9%
<ul> <li>(10) Due to gaining an enhanced ability to use the skills of paraphrasing, summarizing and synthesizing by completing the AW project, I am more capable of avoiding plagiarism in my work.</li> </ul>	4.76 (.68)	98.3%
(11)Due to gaining an enhanced ability to create proper citations by completing the AW project, I am more capable of avoiding plagiarism in my work.	4.78 (.62)	98.3%
II. Perceived usefulness and usability of the Online Scaffolding Writing System		
(12) Use of the online writing system has stimulated my interest in writing	4.39 (1.03)	88.1%
<ul><li>(13) Use of the online writing system has enhanced my engagement in writing assignments.</li></ul>	4.49 (.92)	88.1%
(14) I would like to use the online writing system in other courses.	4.54 (.82)	94.9%
(15) The online writing system is easy to use in general.	4.76 (.70)	96.6%
(16) The benefits of using the online writing system outweigh its technical challenges for users.	4.39 (.97)	84.7%
(17) The online writing system helps me to achieve my learning goals.	4.37 (.96)	88.1%
(18) In the online writing system, peer review is helpful for exchanging views on plagiarism and source use with peers.	4.80 (.81)	94.9%
<ul><li>(19) In the online writing system, peer review is effective at decreasing plagiarism in academic writing</li></ul>	4.71 (.89)	91.5%

*Note.* <sup>#</sup>The HTPP module was introduced to students as an "academic writing project." The ratings are based on a 6-point Likert scale ranging from 1(*strongly disagree*) to 6 (*strongly agree*). "Proportion of agreement" refers to the number of responses expressing agreement (i.e., *slightly agree*, *agree*, or *strongly agree*), compared with the total number of responses (i.e., 64). Two students failed to submit their feedback questionnaires; the number of questionnaires collected was 64.

Twenty-three students and their instructor were interviewed to further explore their opinions on learning with the HTPP module. According to Creswell (2012), interview data analysis comprises several steps that include having a general idea about an interview transcription, coding the transcription, listing codes, and reducing the codes into several themes. The student and instructor interview data were analysed following these steps, and some key

themes emerged (Table 6). The numbers in the column of "students' feedback" are the numbers of interviewees who held positive or negative perceptions of the stated theme.

All the students expressed satisfaction with their learning experiences from using the HTPP module (see Item 1 in Table 6), and most were satisfied with the peer interaction section that facilitated anti-plagiarism behaviour and academic writing (see Item 2 and Item 3). One student (S10) mentioned that she might not have received a chance to learn about plagiarism avoidance and academic writing if she had not been involved in this academic writing project (i.e., the HTPP module). As to what extent they believed that the HTPP module had effects on their plagiarism-free academic writing, 19 interviewees responded that the proportion was more than 70% and its positive effects were mainly on helping them know how to conduct academic writing without plagiarism (see Item 4). However, some students remained concerned about the long-term effects of the HTPP module. One student (S13) remarked, "I feel that my capability of avoiding plagiarism hasn't been greatly improved [by completing the HTPP module]. It is just the beginning and I need more similar writing practise to make greater improvement."

Themes	Students' feedback		
	Positive	Negative	Codes
(1) Learning	23	0	(a) More aware of the importance of anti-plagiarism than
experience with the	(100%)	(0%)	before
HTPP module			(b) Learned how to identify and avoid committing plagiarism
			(c) Improved writing skills
(2) Anti-plagiarism	23	0	(a) Peer interaction
enabling factors of	(100%)	(0%)	(b) Learning resources in the OSWS, especially the three
the HTPP module			pieces of writing with varying levels of plagiarism and
			source acknowledgment
			(c) The course instructor's instruction
(3) Academic writing	23	0	(a) Peer interaction
enable factors of	(100%)	(0%)	(b) Learning resources in the OSWS, especially the three
the HTPP module			writings with varying quality in plagiarism and source use
			(c) The course instructor's instruction
(4) Effects of the	19	4	(a) Improved knowledge of how to generate plagiarism-free
HTPP module on	(87%)	(13%)	academic writing
plagiarism-free			(b) Improved knowledge of how to avoid committing
academic writing			plagiarism
			(c) Understand what constitutes academic writing
(5) Perceptions of	23	0	(a) Objective
undergraduates'	(100%)	(0%)	• Plagiarism is getting serious and it needs be treated
learning about anti-			seriously
plagiarism			• Lack of skills in generating plagiarism-free academic
			writing
			(b) Subjective
			• Serious plagiarism detected in one's dissertation will
			affect one's graduation (according to the university's
			policy on dissertations)

Table 6. Student interview data

The interview with the instructor provided deeper insights into the effects of the HTPP module on facilitating plagiarism-free academic writing. Ms Z remarked, "With the help of the hybrid training module, I not only know exactly how serious the plagiarism problems are in students' writing but can also supervise and provide instructional support during their writing process." Clearly, the introduction of the hybrid training module strengthened the instructor's confidence in delivering plagiarism-free academic writing instruction and her understanding of students' learning needs in relation to plagiarism prevention.

## **5.** Discussion

This study provided robust evidence of the benefits of the HTPP module in facilitating students to generate plagiarism-free academic writing. This success supports a previous study's claims on the pedagogical rationale of developing anti-plagiarism training modules for students and faculty members (Michalak et al., 2018).

#### 5.1. Effects on the students' perceptions of plagiarism

Before the instructional intervention, the students' perceptional baseline was established by comparing the groups' perceptions towards plagiarism. Both groups found it easy to identify the typical features of explicit plagiarism but many students from both groups struggled to distinguish implicit plagiarism in various contexts. The students' learning from the HTPP module were found to contribute more to enhancing their awareness of implicit plagiarism rather than their awareness of explicit plagiarism. This indicates that students' personal experiences related to anti-plagiarism are foundational for acquiring plagiarism awareness, as reported by Peled et al. (2019), and that the effects of the HTPP module on the students' acquisition of knowledge about explicit and implicit plagiarism were uneven.

#### 5.2. Effects on the students' writing, particularly its level of plagiarism and quality

The students' learnings of plagiarism prevention were implemented into their academic writing practise, and the experimental group students were more successful in reducing their writing's level of plagiarism and improving its quality than the control group students. This supports earlier claims regarding the need for systematic instruction on academic writing to improve students' comprehension of plagiarism prevention (Pecorari & Shaw, 2018).

Moreover, although the level of plagiarism in the experimental group students' writings decreased from Level 3 to Level 2 and the average number of the most consecutively copied Chinese characters decreased from 85 to 52, plagiarism remained. This shows that instructors should use the HTPP module for a sufficient length of time to improve their students' ability to generate plagiarism-free writing. That is, the knowledge and skills related to anti-plagiarism are not obtained in one session; they are obtained gradually via a continual process of instructional guidance (Patak et al., 2021).

#### 5.3. Differences and similarities between students' writing performance and perceptions of plagiarism

Analysis of the students' perceptional and behavioural data revealed consistency between the students' perceptions and behavioural performances in one dimension of the intervention but an inconsistency between these in another dimension. According to Hecht et al. (2001), the relationship between perception and behaviour is complicated and thus requires close investigation, and a clear cause-and-effect relationship may not be apparently guaranteed. Similarly, analysing the link between the students' perceptions and behaviours could reveal their learning needs and preferences regarding the HTPP module.

Most of the experimental group stated that they were more capable of writing from sources and avoiding plagiarism after the intervention than before (see Table 5 and Table 6), which was in line with their statistical data showing greater improvements in writing performance than the control group students (see Table 4).

Compared with the writing of the control group, after the intervention the writing of the experimental group had a lower level of plagiarism and was of higher quality (see Table 4). However, this behavioural trend did not correspond to the students' perceptional changes; i.e., the self-evaluation score of the experimental group regarding having 'a good understanding of the importance of avoiding plagiarism' (see Q3 in Table 3) was lower than those of the control group after the intervention, and the experimental group's post-intervention self-evaluation score was also lower than its pre-intervention self-evaluation score for this item. Some researchers have suggested that people's awareness of their changing behaviours may lead them to change their perceptions (Cheng et al., 2019; Festinger, 1962). Thus, the experimental group students may have originally overestimated their understanding of plagiarism before the intervention. Therefore, during the writing process, their high self-evaluation scores may have been challenged by their realisation that they had committed plagiarism in their drafts. This might have led them to change their perceptions of their own knowledge regarding anti-plagiarism behaviour. Because students' self-evaluations regarding plagiarism-free academic writing can be altered, we believe that students' self-evaluation scores on their understanding of plagiarism will increase if they are given more opportunities to practise academic writing using the HTPP module.
# 6. Conclusion

We used a quasi-experimental design to investigate the effects of the HTPP module on students' plagiarism-free academic writing. Some important findings were obtained. First, the HTPP module was found to effectively improve the students' perceptions of plagiarism and their ability to generate plagiarism-free academic writing. Second, both the students and the instructor expressed their satisfaction about the use of the HTPP module in the course. Third, the students' perceptions of plagiarism were consistent with their anti-plagiarism behavioural performance in some dimensions but not in others. These findings shed light on the utility of the HTPP module in helping undergraduates to generate plagiarism-free academic writing. This should enhance instructors' understanding of and confidence in adopting the HTPP module in their course designs. Moreover, our findings – particularly those related to the learning experiences of the students and the instructional experiences of the instructor – support the implementation of this innovative module in various disciplinary and cultural contexts.

Our study had several limitations that should be noted, including a limited number of participants and the short duration of the experiment. Thus, in future work, more students should be evaluated over a longer duration to determine their long-term developmental progress in learning about plagiarism prevention from the HTPP module. It would also be valuable to investigate the effects of the HTPP module on the writing skills of participants of various ages. Moreover, the way of inquiry of plagiarism-free academic writing in different disciplines can greatly affect the results of a study. Therefore, there is a need to examine HTPP module-based instructional strategies that are suitable for scenarios in various disciplines and for students' various learning modes.

## Acknowledgements

This paper is the phased research results of Shandong Social Science Planning Research Project [Project approval number: 20CJYJ15].

## References

Abasi, A. R., & Graves, B. (2008). Academic literacy and plagiarism: Conversations with international graduate students and disciplinary professors. *Journal of English for Academic Purposes*, 7(4), 221-233.

AlSheikh, M. H., & Iqbal, M. Z. (2019). Student perceptions regarding group learning activities in a hybrid medical curriculum. *Medical Science Educator*, 29(4), 1221-1228.

Attwood, R. (2008, June 12). 143 students expelled for plagiarism. Times Higher Education. https://www.timeshighereducation.com/news/143-students-expelled-for-plagiarism/402351.article?site=cn

Awada, G. M., & Diab, N. M. (2021). Effect of online peer review versus face-to-face peer review on argumentative writing achievement of EFL learners. *Computer Assisted Language Learning*. https://doi.org/10.1016/j.compedu.2019.103642

Awasthi, S. (2019). Plagiarism and academic misconduct: A Systematic review. *DESIDOC Journal of Library & Information Technology*, 39(2), 94-100.

Batane, T. (2010). Turning to Turnitin to fight plagiarism among university students. *Educational Technology & Society*, 13(2), 1-12.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*. Springer. https://doi.org/10.1007/978-3-642-00296-0\_5

Blum, S. D. (2011). My word!: Plagiarism and college culture. Cornell University Press.

Brown, N., & Janssen, R. (2017). Preventing plagiarism and fostering academic integrity: A Practical approach. *Journal of Perspectives in Applied Academic Practice*, 5(3), 102-109.

Cheng, E. W. L., Chu, S. K. W., & Ma, C. S. M. (2019). Students' intentions to use PBWorks: A factor-based PLS-SEM approach. *Information and Learning Sciences*, *120*(7/8), 489-504. https://doi.org/10.1108/ILS-05-2018-0043

Choi, Y. H. (2012). Paraphrase practices for using sources in L2 academic writing. English Teaching, 67(2), 51-79.

Chu, S. K. W., Li, X., & Mok, S. (2021). UPCC: A Model of plagiarism-free inquiry project-based learning. *Library & Information Science Research*, 43(1), 101073. https://doi.org/10.1016/j.lisr.2021.101073

Creswell, J. W. (2012). Educational research: Planning, conducting, and evaluating quantitative and qualitative research. Pearson Education.

Davies, L. J. P., & Howard, R. M. (2016). Plagiarism and the internet: Fears, facts, and pedagogies. In *Handbook of academic integrity* (pp. 591-606). Springer Singapore.

de Maio, C., Dixon, K., & Yeo, S. (2020). Responding to student plagiarism in Western Australian universities: The Disconnect between policy and academic staff. *Journal of Higher Education Policy and Management*, 42(1), 102-116.

Du, Y. (2019). Evaluation of intervention on Chinese graduate students' understanding of textual plagiarism and skills at source referencing. *Assessment & Evaluation in Higher Education*, 45(1), 14-29.

Eaton, S. E. (2021). Plagiarism in higher education: Tackling tough topics in academic integrity. ABC-CLIO.

Festinger, L. (1962). A Theory of cognitive dissonance. Stanford University Press.

Greenberger, S., Holbeck, R., Steele, J., & Dyer, T. (2016). Plagiarism due to misunderstanding: Online instructor perceptions. *Journal of the Scholarship of Teaching and Learning*, *16*(6), 72-84.

Grigg, A. (2016). *Chinese international students' and faculty members' views of plagiarism in higher education* (Publication No. 10010905) [Doctoral thesis, Walden University]. ProQuest Dissertations and Theses Global. https://scholarworks.waldenu.edu/cgi/viewcontent.cgi?article=3153&context=dissertations

Hadwin, A. F., Järvelä, S., & Miller, M. (2017). Self-regulation, co-regulation and  $\in$  shared regulation in collaborative learning environments. In D. Schunk & B. Zimmerman (Eds.), *Handbook of self-regulation of learning and performance* (2nd, pp. 83-106). Routledge.

Harris, L., Harrison, D., McNally, D., & Ford, C. (2020). Academic integrity in an online culture: Do McCabe's findings hold true for online, adult learners? *Journal of Academic Ethics*, *18*(4), 419-434.

Hecht, H., Vogt, S., & Prinz, W. (2001). Motor learning enhances perceptual judgment: A Case for action-perception transfer. *Psychological Research*, 65(1), 3-14.

Hofer, A. R., Townsend, L., & Brunetti, K. (2012). Troublesome concepts and information literacy: Investigating threshold concepts for IL instruction. *Portal: Libraries and the Academy*, 12(4), 387-405.

Hopp, C., & Speil, A. (2021). How prevalent is plagiarism among college students? Anonymity preserving evidence from Austrian undergraduates. *Accountability in Research*, 28(3), 133-148.

Hoy, W. K., & Adams, C. M. (2015). Quantitative research in education: A Primer. Sage Publications.

Hu, G., & Lei, J. (2016). Plagiarism in English academic writing: A Comparison of Chinese university teachers' and students' understandings and stances. *System*, 56, 107-118.

Huang, J. C. (2017). What do subject experts teach about writing research articles? An Exploratory study. *Journal of English for Academic Purposes*, 25, 18-29.

Hutchison, A., & Woodward, L. (2014). A Planning cycle for integrating digital technology into literacy instruction. *The Reading Teacher*, 67(6), 455-464.

International Center for Academic Integrity. (n.d.). *Facts and statistics*. https://academicintegrity.org/resources/facts-and-statistics?highlight=WyJzdGF0aXN0aWNzII0=

Jamieson, S. (2017). The Evolution of the citation project: Lessons learned from a multi-year, multi-site study. In T. Serviss, & S. Jamieson (Eds.), *Points of departure: Rethinking student source use and writing studies research methods* (pp. 33-61). Utah State University Press.

Law, K. M., Geng, S., & Li, T. (2019). Student enrollment, motivation and learning performance in a blended learning environment: The Mediating effects of social, teaching, and cognitive presence. *Computers & Education*, *136*, 1-12.

Lee, C. W. Y., Chu, S. K. W., Cheng, J. O. Y., & Reynolds, R. (2016). Plagiarism-free inquiry project-based learning with UPCC pedagogy. *Proceedings of the Association for Information Science and Technology*, 53(1), 1-11.

Leung, C. H., & Cheng, S. C. L. (2017). An Instructional approach to practical solutions for plagiarism. *Universal Journal of Educational Research*, 5(9), 1646-1652.

Liu, C. J., & Yang, S. C. (2012). Applying the practical inquiry model to investigate the quality of students' online discourse in an information ethics course based on Bloom's teaching goal and Bird's 3C model. *Computers & Education*, 59(2), 466-480.

Liu, G. Z., Lo, H. Y., & Wang, H. C. (2013). Design and usability testing of a learning and plagiarism avoidance tutorial system for paraphrasing and citing in English: A Case study. *Computers & Education*, 69, 1-14.

Löfström, E., Huotari, E., & Kupila, P. (2017). Conceptions of plagiarism and problems in academic writing in a changing landscape of external regulation. *Journal of Academic Ethics*, 15(3), 277-292.

Lu, H. C. (2013). Developing Taiwanese EFL undergraduate students' knowledge of plagiairism avoidance and enhancing their English paraphrasing and citation skills by using an online writing tutorial system [透過線上寫作系統來發展台灣大學生的反抄襲知識、英文改寫能力與引用技巧] (Unpublished master thesis). National Cheng Kung University, Taiwan.

Masic, I. (2012). Plagiarism in scientific publishing. Acta Informatica Medica, 20(4), 208-213.

Maxwell, A., Curtis, G. J., & Vardanega, L. (2006). Plagiarism among local and Asian students in Australia. *Guidance & Counselling*, 21(4), 210-215.

Michalak, R., Rysavy, M., Hunt, K., Smith, B., & Worden, J. (2018). Faculty perceptions of plagiarism: Insight for librarians' information literacy programs. *College & Research Libraries*, 79(6), 747. https://doi.org/10.5860/crl.79.6.747

Mostert, M., & Snowball, J. D. (2013). Where angels fear to tread: Online peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education*, 38(6), 674-686.

Mphahlele, A., & McKenna, S. (2019). The Use of Turnitin in the higher education sector: Decoding the myth. *Assessment & Evaluation in Higher Education*, 44(7), 1079-1089.

Myers, C. S. (2018). Plagiarism and copyright: Best practices for classroom education. *College & Undergraduate Libraries*, 25(1), 91-99.

Pàmies, M. D. M., Valverde, M., & Cross, C. (2020). Organising research on university student plagiarism: A Process approach. *Assessment & Evaluation in Higher Education*, 45(3), 401-418.

Parks, R. F., Lowry, P. B., Wigand, R. T., Agarwal, N., & Williams, T. L. (2018). Why students engage in cyber-cheating through a collective movement: A Case of deviance and collusion. *Computers & Education*, 125, 308-326.

Patak, A. A., Wirawan, H., Abduh, A., Hidayat, R., Iskandar, I., & Dirawan, G. D. (2021). Teaching English as a foreignlanguage in Indonesia: University lecturers' views on plagiarism. *Journal of Academic Ethics*, 19(4), 571-587.

Pecorari, D., & Petrić, B. (2014). Plagiarism in second-language writing. Language Teaching, 47(3), 269-302.

Pecorari, D., & Shaw, P. (2018). Student plagiarism in higher education: Reflections on teaching practice. Routledge.

Peled, Y., Eshet, Y., Barczyk, C., & Grinautski, K. (2019). Predictors of academic dishonesty among undergraduate students in online and face-to-face courses. *Computers & Education*, 131, 49-59.

Reinholz, D. (2016). The Assessment cycle: A Model for learning through peer assessment. Assessment & Evaluation in Higher Education, 41(2), 301-315.

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (1991). Criteria for scale selection and evaluation. *Measures of Personality and Social Psychological Attitudes*, 1(3), 1-16.

Roller, M. (2018, June 19). Oxford course tackles plagiarism. The Times. https://www.thetimes.co.uk/article/oxford-university-launches-anti-plagiarism-course-to-tackle-record-cheating-xtfm8z77l.

Schinkel, A. (2015). Education and ultimate meaning. Oxford Review of Education, 41(6), 711-729.

Stemler, S. E. (2004). A Comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4), 1-19.

Tindall, I. K., & Curtis, G. J. (2020). Negative emotionality predicts attitudes toward plagiarism. *Journal of Academic Ethics*, 18(1), 89-102.

To, J., & Carless, D. (2016). Making productive use of exemplars: Peer discussion and teacher guidance for positive transfer of strategies. *Journal of Further and Higher Education*, 40(6), 746-764.

Tomaš, Z. (2010). Addressing pedagogy on textual borrowing: Focus on instructional resources. Writing and Pedagogy, 2(2), 223-250.

Tremayne, K., & Curtis, G. J. (2021). Attitudes and understanding are only part of the story: Self-control, age and self-imposed pressure predict plagiarism over and above perceptions of seriousness and understanding. *Assessment & Evaluation in Higher Education*, 46(2), 208-219.

University College London. (2019). *Plagiarism: A Definition*. University College London. https://www.ucl.ac.uk/students/exams-and-assessments/plagiarism

Uzun, A. M., & Kilis, S. (2020). Investigating antecedents of plagiarism using extended theory of planned behavior. *Computers & Education*, 144, 103700. https://doi.org/10.1016/j.compedu.2019.103700

Vessal, K., & Habibzadeh, F. (2007). Rules of the game of scientific writing: Fair play and plagiarism. *The Lancet*, 369(9562), 641. https://doi.org/10.1016/S0140-6736(07)60307-9

Wiske, M. S., & Breit, L. (2013) Teaching for understanding with technology. Jossey-Bass.

Xinhua. (2019, April 12). China's top university to expel students who plagiarize. Xinhua. http://www.xinhuanet.com/english/2019-04/12/c\_137971212.htm

Yeung, A. H., Chu, C. B., Chu, S. K. W., & Fung, C. K. (2018). Exploring junior secondary students' plagiarism behavior. *Journal of Librarianship and Information Science*, 50(4), 361-373.

Zhang, N., & Tang, K. Y. A. (2021). The Reasons and countermeasures of Chinese College students' unintentional plagiarism. In 2021 7th International Conference on Education and Training Technologies (pp. 153-157). ICETT.

Zhang, Y., Xiuhan, L., Zainuddin, Z., & Chu, S. K. (2019, January). Using pedagogical intervention with ICT to minimize student plagiarism [Paper presentation]. ASIS&T Asia-Pacific Regional Conference 2019, Phnom Penh, Cambodia.

	Assignment response	Coherence and	Vocabulary and	Citation
Excellent (20~25 points)	<ul> <li>The writing meets all of the assignment requirements.</li> <li>The writer projects a clear stance with supporting ideas.</li> <li>The summary includes all of the important aspects of the sources and displays the writer's full understanding of the sources.</li> </ul>	<ul> <li>The writer presents ideas in clear and logical sequence.</li> <li>The writer uses cohesive devices effectively.</li> <li>The writer uses paragraphing appropriately.</li> </ul>	<ul> <li>The writer         paraphrases and             summarizes texts in             his/her own way             which is totally             different from the             sources.     </li> <li>The writer uses a             wide range of             vocabulary naturally.</li> </ul>	The citations and references are presented correctly.
Good (13 ~19 points)	<ul> <li>The writing meets all of the assignment requirements, although some requirements are less fulfilled than others.</li> <li>The writer projects a stance, although the conclusion is not clearly stated.</li> <li>The summary includes some important aspects of the sources and displays the writer's good understanding of the sources, although some aspects are not accurately addressed.</li> </ul>	<ul> <li>The writer generally presents ideas in a logical sequence.</li> <li>The writer uses cohesive devices to connect sentences, but some cohesive devices are used erroneously.</li> <li>The writer uses paragraphing but not always correctly.</li> </ul>	<ul> <li>The writer paraphrases and summarizes texts in his/her own way, although sometimes uses similar phrases.</li> <li>The writer uses a wide range of vocabulary but there are some inaccuracies.</li> </ul>	There are citations and references, but some are presented incorrectly.
Average (6 ~12 points)	<ul> <li>The writing meets the assignment requirements only partially.</li> <li>The writer projects a stance but it is not clearly stated.</li> <li>The summary includes few important aspects of the sources and displays the writer's limited understanding of the sources.</li> </ul>	<ul> <li>The writer presents ideas in sequence but this lacks logic.</li> <li>The writer uses a few cohesive devices but does so repeatedly or incorrectly.</li> <li>The writing is not paragraphed.</li> </ul>	<ul> <li>The paraphrases and summaries contain few of the writer's own sentence structures or expressions, and contain several phrases that are similar to those in the sources.</li> <li>The writer repeatedly uses a limited range of vocabulary.</li> </ul>	Either the citations or the references are missing.
Poor (1~5 points)	<ul> <li>The writing barely meets the assignment requirements.</li> <li>The writer doesn't project a stance.</li> <li>The summary doesn't include any important aspects of the sources and displays no sign of the writer's understanding of the sources.</li> </ul>	• The writer has very little control of organizational features.	<ul> <li>The paraphrase and summary display the writer's use of source language with little modifications.</li> <li>The writer uses an extremely limited range of vocabulary.</li> </ul>	Both the citations and the references are missing.

# Appendix A. The four-dimensional writing rubric

Very poor The author has not handed in the assignment or merely copied words from original passage. (0 point)

*Note.* The rubric is based on the writing assignment rubric of Lu (2013) and the rubric on paraphrasing of Choi (2012).

# Using E-textiles to Design and Develop Educational Games for Preschoolaged Children

# Ersin Kara<sup>1\*</sup> and Kursat Cagiltay<sup>2,3</sup>

<sup>1</sup>Middle East Technical University, Turkey // <sup>2</sup>Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey // <sup>3</sup>The Digital Economy Research Center, Azerbaijan State University of Economics (UNEC), Baku, Azerbaijan // ersinkara07@gmail.com // cagiltay@gmail.com

\*Corresponding author

(Submitted November 28, 2021; Revised August 28, 2022; Accepted September 19, 2022)

**ABSTRACT:** This paper reports on the design and development of educational games and materials that utilize affordable e-textile technology. The researchers employed a design-based approach whereby preschool children used three e-textile materials in two cycles to inform on the development of interactive materials from ordinary objects and bodily interactive games. The study's data were collected and analyzed according to the design-based research framework through iterative cycles of interviewing, video recording, and note-taking. The paper describes the characteristics, pros, and cons of e-textiles and what to consider when using them to create interactive educational materials for preschool-aged children.

Keywords: E-textile, Wearable technology, Preschool education, Design-based research (DBR), Executive functions

# 1. Introduction

The educational value of e-textiles has been on the agenda for researchers for some time (e.g., Fields et al., 2021; Peppler & Glosson, 2013). Most e-textiles are easy to employ within the preschool context, enabling the transformation of ordinary objects, toys, and clothing into digitally interactive materials. Developing interactive materials from the objects already familiar to children, such as toys and clothes, helps children to become more easily familiarized with them (Vega-Barbas et al., 2015). More importantly, as the current study will exemplify, e-textile technology enables the development of bodily interactive games (e.g., Doménech et al., 2018) which have, among other treatments, been shown to cultivate Executive Functions (EFs) in children (e.g., Best, 2012; Gao et al., 2019; Rafiei Milajerdi et al., 2021; Xiong et al., 2019) and young people (López-Serrano et al., 2021).

Although other technologies in the market can provide some of the same functionalities that e-textiles are claimed to serve (i.e., bodily interactions and maintaining the connection with the real world during use) such as Microsoft Kinect and HoloLens, these technologies are more expensive, allow for little to no modification as they have readily-available hardware, and may not be considered appropriate for preschool-aged learners. Further, developing games or educational materials with these types of technology may be far more complicated than developing games with e-textiles. We contend that bodily interactive games that harvest the power of e-textile technologies can help enhance EF skills and investigate EF intervention characteristics stated in the literature (e.g., Diamond & Ling, 2016; Gashaj et al., 2021; Rafiei Milajerdi et al., 2021). However, the design issues related to e-textiles have only been briefly addressed in the literature, and as such there is no guidance for researchers or educators to refer to when seeking to utilize e-textiles. This gap is considered even wider when we look at the design of e-textiles for educational purposes, and specifically for younger-aged children.

The current study employed a design-based research (DBR) to discover how e-textile technology can be utilized to develop educational games for preschool-aged children. The study aimed to contribute to the literature, and specifically to the field of human-computer interaction, by reporting on the outcomes of a DBR project undertaken with preschool children and teachers over two iterative cycles. The study further puts forward a suggested guide for the deployment of e-textiles when developing interactive educational e-textiles for preschool-aged children, and the pertinent design issues that should be considered.

# 1.1. Literature review

The term electronic textiles or "e-textiles" refers to a wide variety of electronic components such as mainboards, sensors, and conductive materials that are incorporated into clothing and wearable accessories that are aimed to handle specific tasks such as sensing physical properties and communication (Coccia et al., 2021; Ismar et al.,

2020). E-textile applications extend the functionality of electronic systems, enabling them to become portable, customized, and ubiquitous (Komolafe et al., 2019). Being malleable and in harmony with the body, e-textiles can help to diversify tangible experiences and encourage bodily interaction (Ugur Yavuz et al., 2021). The main e-textile components used to develop materials in the current study are presented in Figure 1.



Figure 1. E-textile components used in the study

*Note.* Components by Adafruit are used: (a) programmable microcontroller for governing other electronic components and workflow, (b) Bluetooth LE for facilitating wireless communication, (c) color sensor for detecting the colors of objects, (d) accelerometer/gyroscope/magnetometer for detecting motion, (e) Neopixel for emitting different colors, (f) lux sensor for detecting light levels, and (g) conductive thread for connecting components of a circuit.

A wide variety of applications have utilized e-textile materials. Although the purposes of e-textile-based artifacts may vary, as Chen et al. (2021) indicated, e-textile design is an interdisciplinary process that integrates both engineering and textile design. It also requires decision-making at various levels of production, realizing that functionality goes hand in hand with expertise in creating e-textile materials, as well as considering the personal preferences and cultural expectations of the target group (Kafai & Peppler, 2014). Studies that discuss e-textile design are still emerging, and the available information addresses several general areas of concern. For example, Almusawi et al. (2021) explored physical education teachers' perceptions regarding wearable use during physical education, and revealed that the teachers perceived comfort (e.g., convenience, inclusion, and wearability), safety, and customizability (e.g., esthetics) as important. Working with children between the ages of 1 and 7 years old, Honauer et al. (2019) explained the characteristics of digitally interactive soft toys; stating that they should be durable and washable, have simple and obvious interactions, and include tactile feedback. They also added that actions should produce familiar and immediate effects. Nonnis and Bryan-Kinns (2019) considered intuitive use and durableness in their material design process. Similar issues that have also been considered in other studies are esthetics, function, and wearer acceptance (Balestrini et al., 2014); lightness and durability, combining multiple senses, instantaneous feedback mechanisms, and real-life movement-based interaction (Norooz et al., 2015); providing feedback, incorporating multiple senses (i.e., lights, sounds, and vibrations), hiding and protecting the battery, children's motivation to play with the item, understandability, intuitiveness, and comfort (Vega-Barbas et al., 2015).

Numerous materials with distinct qualities and purposes can be developed using e-textiles. However, the guiding frameworks for e-textile design in terms of educational purposes are yet to be established. The current study aims to advance the e-textile design literature by reporting on the design considerations that may contribute to the formation of a reliable framework.

# 2. Research design

The study employed Design-Based Research (DBR) to investigate the design issues related to the production of e-textile materials. DBR can be defined as "an emerging paradigm for the study of learning in context through the systematic design and study of instructional strategies and tools" (The Design-Based Research Collective, 2003, p. 5). Iterative processes should lead to design principles, which can then be employed by other researchers or practitioners (Amiel & Reeves, 2008). The current study consisted of two iterative cycles that worked on four bodily interactive e-textiles. The purpose of Cycle 1 was to reveal the characteristics of the e-textiles utilized in the study, and to determine material revisions through implementations and interviews. Cycle 2 mainly included testing the effectiveness of each applied revision by using the same methods as applied in Cycle 1, and then

concluding the applicable design principles (see Figure 2). Three of the designed games were aimed to be used with two wearable e-textile materials (i.e., a belt bag and a thigh band) that could interact with a tablet computer, whilst one game was played with a non-wearable standalone material. The study was completed over a 4-year period. Significant dates during the study's lifecycle are provided in the Appendix (see Figure A1).



#### Figure 2. Iterations and outcomes of the study

#### 2.1. Research questions

The following were formed as the main research questions of the study:

- What are the affordances and constraints of e-textiles for developing bodily interactive games for preschoolaged children?
- What are the design principles and the lessons learned that guide the design and development of bodily interactive e-textiles for preschool-aged children?

### 2.2. Participants

Criterion sampling (Patton, 1990) was used to select 19 preschool-aged children between 48 and 72 months old from two different preschools. Six teachers from Cycle 1, together with three from Cycle 2 who were also present in Cycle 1, participated in the study. The teachers were selected from three different preschools. An academic specializing in Early Childhood Education (ECE) also participated in the study. Both the teachers and the academic were selected according to convenience sampling. Demographic information of study's participants is presented in Table 1.

Table 1. Demographic information of participants									
Cycle	No. of teachers	Average teaching experience	No. of children	Average age of children					
		(years)		(months)					
Cycle 1	6	6.6	10	58					
Cycle 2	3	9.6	9	67					

#### 2.3. Development of games and materials

Structuring games (i.e., mechanisms, rules, game flow, interaction ways) is a challenging process that requires several issues to consider. In the current study, we started by listing the capabilities of e-textile components, which enabled the development of draft game ideas (e.g., a Neopixel's ability to emit more than one color would enable producing numerous color sequences). Simultaneously, EF assessment tools and ECE curriculum

objectives and practices were also examined. In the current study, we aimed to address three core EFs, inhibition or inhibitory control, working memory, and cognitive flexibility (Center on the Developing Child at Harvard University, 2011; Diamond, 2012). Each proposed game idea was then evaluated considering resources (i.e., cost, time, and technical feasibility) and then the most applicable were decided upon with the approval of the ECE academic.

The digital games were developed using Adobe Animate CC 2015 and the connection between wearables and the tablet computer was established by Bluetooth. The digital games were designed so as to provide both visual and auditory reinforcement, and also verbal feedback and instruction (except *Light Order* game since the material did not possess that ability). Verbal feedback and instructions were prepared with the help of a participant ECE teacher.

# 2.3.1. Light Order game (LO)

The first game idea involved remembering a color sequence and then repeating it without further sight of the initial sequence. Neopixels, which can be programmed to emit different colors, were used to create the color sequences. A color sensor was then used to detect the responses of the user. An image from the second version of the developed material is shown in Figure 3. The game requires children to use their working memory skills.

Figure 3. Child playing LO game (second version)



# 2.3.2. Follow Pattern game (FP)

The second game idea involved viewing a pattern (e.g., top-right image in Figure 4), recognizing the pattern structure at that time, and then repeating the pattern through sitting and standing movements after the initial pattern was no longer visible. The game aimed to address the working memory skills of the user. A 9-DOF IMU placed on a thigh band was used to detect the user's body movements, which is considered a feasible way to detect sitting and standing movements. An image from the second version of the developed material is shown in Figure 4.

*Figure 4*. Child playing FP game (second version, screenshot displayed top-right)



# 2.3.3. Do as I Say/Do game (DISD)

The third game to be evaluated was named *Do as I Say/Do*, for the sake of simplicity. The game required the user to manage their focus of attention on one type of stimuli whilst ignoring another (i.e., inhibitory control). We evaluated different game ideas and ECE objectives with a participant preschool teacher and decided upon a game that combined two games that are commonly played with preschool-aged children, *Camel-Dwarf* and *Do as I Say Not as I Do* games. In the traditional *Camel-Dwarf* game, children are expected to stand up when they hear the word *Camel* and sit down upon hearing the word *Dwarf*. The new game is very similar to the *Do as I Say Not as I Do* game, in which the teacher provides auditory instruction and provides a visual distractor, or vice versa. However, in the new game, visual distractors are confined to sitting and standing, whilst the auditory stimuli are limited to the words *Camel* and *Dwarf*. An image from the second version of the developed materials is shown in the Appendix (see Figure A2).

# 2.3.4. Object Sorting (OS)

The fourth game was inspired by two cognitive flexibility tests, Dimensional Change Card Sort (DCCS; Zelazo, 2006) and Wisconsin Card Sorting Test (WCST; Grant & Berg, 1948). The purpose of the game was to sort the object at the center of the screen to the right-hand or left-hand side of the screen according to the feedback received. An interactive belt bag with two light sensors functioning as buttons was used to sort the objects on a tablet computer screen. An image from the second version of the developed materials is shown in Figure 5.



As previously s	tated, ECE	objectives	were	also	considered	whilst	developing	the	games.	Table	2	shows	the
objectives that ea	ach game ac	ldresses.											

Table 2. Games and ECE objectives addressed

Game	ECE Curriculum objective	Objective domain(s)	EF domain
Light Order	1. Gives attention to an object/situation/event	1, 2: Cognitive	Working
	2. Remembers what is perceived		memory
Follow Pattern	1. Gives attention to an object/situation/event	1, 2, 3: Cognitive	Working
	2. Remembers what is perceived		memory
	3. Creates patterns with objects		
Do as I Say/Do	1. Gives attention to an object/situation/event	1: Cognitive	Inhibition
	2. Comprehends the meaning of what is listened	2: Language	
	to/watched	development	
Object Sorting	1. Gives attention to an object/situation/event	1, 2: Cognitive	Cognitive
	2. Groups objects or assets according to their		flexibility
	properties		

### 2.4. Data collection and analysis

DBR interventions generally adopt mixed methods that utilize diverse means and procedures (Anderson & Shattuck, 2012). The current study adopted interviewing, video recording, and note-taking as the data collection methods. Using multiple data collection methods enabled the triangulation of findings by comparing the data obtained from each method. Additionally, the study's iterative structure made it possible to interview the same participants at different stages in the process, facilitating the triangulation of data from multiple sources (Merriam, 2009). The iterative structure of the study necessitated that, following the first cycle, the researchers analyzed all the data, made decisions about each digital game, revisions and enhancements to the e-textile materials, and their application.

#### 2.4.1. Interviews

DBR involves collaboration between researchers and practitioners to bring solutions to education problems (Anderson & Shattuck, 2012). In terms of the current study, an academic from ECE field and preschool teachers were interviewed iteratively as part of the collaborative approach to the study. During semi-structured interviews, the participants were each asked to state their opinions regarding the following: (a) the game designs, (b) the strengths and weaknesses of each design, (c) how the materials and games could be enhanced so as to make them more appropriate to the preschool children's developmental level, and (d) their opinions regarding utilizing these games and materials in ECE. Interviews with the participant ECE teachers were conducted after they had either observed the children engaging with the games and materials or had watched footage of children playing the games. The purpose was to allow the ECE teachers to observe real-world interactions instead of them trying to imagine how children would engage with the proposed games and materials. Also, the ECE teachers were each provided with the actual e-textile materials and digital games so that they could use them at their will to gain adequate insight into the game itself and the e-textile material mechanism employed. Prior to analyzing the interview data, audio recordings of the interviews were transcribed, and reliability was also addressed according to the intercoder agreement protocol as suggested by Creswell and Poth (2016).

#### 2.4.2. Video recordings

Design-based studies generally require the handling of an extensive amount of data such as video recordings (Collins et al., 2004). Erickson (2006) suggested video recording as a method that could be used to explore learner-instructional material interaction. In the current study, the implementations applied with the participant preschool children were video-recorded (68 sessions in total) while the children played the developed educational games and interacted with the e-textile materials they incorporated. This method of recording is termed *observational recording*, in which the subjects of a study are recorded while they are busy with the activity of interest (Penn-Edwards, 2004). The filming was made from a fixed point using a tripod so as not to distract the children. The participant preschool children played the games individually in a vacant classroom because each game required their full attention. The average game durations for Cycle 1 and Cycle 2 were 13/19 mins for the *LO*, 11/17 mins for the *FP*, 5/7 mins for the *DISD*, and 6/9 mins for the *OS*, respectively. As Erickson (2006) noted, video analysis requires developing methods to attend to phenomena of interest through multiple viewings. In the current study we analyzed all the video data extensively and for several times in order to detect issues regarding gameplay, e-textile use, and game-e-textile-child interaction.

The inductive approach, in which footage is analyzed exploratively without following a sound theory (Derry et al., 2010), was applied in the transcription of the video recordings. The researchers' instructional design background and experience with e-textile projects, as well as the e-textile literature, guided the selection of instances to be coded, such as situations that prevented the preschool children from playing the games as intended, the children's prominent behaviors (e.g., casual movements made whilst game playing), technical problems and limitations, and features that supported ease of use of each game and material. While the nature of Cycle 1 of the analysis was explorative, Cycle 2's analysis was largely confirmative. In other words, the findings from the video analysis of Cycle 1 guided the video analysis in Cycle 2 through enabling comparison of whether phenomena observed in Cycle 1 were also present in Cycle 2. The video analysis process applied can be summarized as follows:

- (1) Situations of interest were transcribed together with timestamps in Cycle 1 and Cycle 2.
- (2) Transcriptions were analyzed according to qualitative content analysis (Merriam, 2009) with MAXQDA software used in Cycle 1 and Cycle 2. This included two steps:
  - Open coding a method in which the coder tries to identify any potentially valuable data chunks.

- Analytical coding grouping codes under subcategories, and then structuring the main categories.
- (3) Observation checklists with comment areas were prepared based on the content analysis in Cycle 1 (e.g., items on the OS checklist included: a) The child applied too much pressure on, hit, or rubbed the belt bag, and b) The child had difficulty in covering the circles on the belt bag). The checklists prepared in Cycle 1 were revised based on the content analysis in Cycle 2 (e.g., the LO checklist was revised to include the item "The child had difficulty in understanding the negative feedback sound produced by the Piezo Buzzer element," since the first version of the material was unable to provide a negative feedback sound).
- (4) Another researcher watched and coded each video recording (n = 4) using checklists created in Cycle 1 and Cycle 2. The process included corroborating the behavior (agreeing or disagreeing with commenting) and categorization.
- (5) The researchers compared and discussed the codes they had assigned. A shared understanding of the data was achieved in both cycles.

## 2.4.3. Design notes

One of the research goals of the current study was to reveal the affordances of the e-textile components used. For this purpose, the researcher who designed the e-textile materials took notes whilst they were developed. Design notes included coding requirements (e.g., basic or complex) and practical knowledge of the electronic components such as their ideal working conditions and environmental factors deemed pertinent to their application.

# **3.** Findings

This DBR study revealed some useful findings and lessons with regards to the development of materials for preschool-aged children using e-textile technology. Four categories emerged from the qualitative data analysis: (a) technical issues; (b) usability; (c) visual design; and (d) perceived usefulness and instructional quality.

## 3.1. Technical issues of e-textile components

E-textiles can be used to develop interactive interfaces and wearables through components enabling communication, detecting physical inputs such as light, color, motion, and producing stimuli such as light and sound. However, they also posit various technical limitations and challenges. A summary of the technical issues related to e-textiles used in the current study are presented in Table 3.

Table 3. Summary of the technical issues encountered					
Component	Function	Issues	Indication		
Bluetooth module	Communication between e-textile material and tablet computer	Pros: Provides movement flexibility.	Children played the games while doing casual movements or moving according to the background music's rhythm.		
		Cons: Requires advanced	Communication delays affected		
Color sensor	Detecting colors as answers	algorithms.Pros: Everyday objects can be turned into interactive tools.Cons: Requires preliminary work to define color ranges for the objects to be used.Position and distance of objects to the color sensor and ambient lighting affect detection.	Squares and hands made of felt were used in this project. Color range of felt objects were explored under different conditions (e.g., various distances, positions, and lighting).		
Conductive thread	Used to create circuits	Pros: Enables flexible and washable circuits. Can be sewn onto fabric. Cons: High resistance. Vulnerable to short circuiting.	<ul><li>Circuits in the <i>LO</i> used conductive thread sewn onto felt.</li><li>Creating circuits required delicate work and consideration to avoid short</li></ul>		

		Less reliable than insulated wire for data transmission Susceptible to friction.	circuits. Isolated wires were more stable and reliable in designs using BLE connections. Friction wore off threads and caused connection problems.
Light sensor	Functioned as buttons	<ul><li>Pros: Easy interaction.</li><li>Basic coding.</li><li>Cons: Shadows may trigger unintended activation.</li><li>Ambient lighting may affect algorithms.</li></ul>	<ul><li>Children had no difficulty covering the light sensors placed on the belt.</li><li>Light sensors of belt bag v1 had frames that left a small gap for diodes on the sensors, which sometimes led to false activation since the frames cast shadows when the children moved.</li><li>Algorithms were tested and updated according to ambient lighting.</li></ul>
Neopixel	Used to emit different colors	<ul> <li>Pros: Easy to manage.</li> <li>Versatile as emits numerous colors.</li> <li>Cons: Non-primary colors (i.e., not red, green, or blue) may be less precise to the human eye.</li> <li>High-RGB value colors are eyestraining and appear whiteish.</li> <li>Low-RGB values do not mix sufficiently to produce intended color.</li> </ul>	<ul><li>Six Neopixels were easily used in the <i>LO</i> game material to create color sequences.</li><li>Children had difficulty identifying non-primary colors.</li><li>Experiments showed that very high and very low RGB values did not work as intended.</li></ul>
9-DOF IMU	Used to detect sitting and standing motions	<ul> <li>Pros: Versatile as detects various positional data.</li> <li>Can be used to detect body movement.</li> <li>Cons: May require complex algorithms and calculations, especially in projects requiring combined measurements.</li> </ul>	<ul><li>The developed thigh band used a 9- DOF IMU to detect sitting and standing motions.</li><li>A thigh band was designed since it enabled detecting motions with relatively simple algorithms.</li></ul>
Piezo buzzer element	Used to give auditory stimuli in the <i>LO</i> game	Pros: Easy to use and code. Cons: Inability to produce advanced or verbal sounds. Sounds produced may be too similar.	Simple algorithms were written to use the component. The researchers gave verbal feedback in the <i>LO</i> game Children confused different positive reinforcements given for various purposes.

#### **3.2.** Usability of the developed materials

The study's results showed that while designing educational e-textiles, four issues should be considered to increase their usability and effectiveness: (a) Intuitiveness of the interactions; (b) size of the materials; (c) perceived responsiveness; and (d) stability and sturdiness.

# 3.2.1. Intuitiveness of the interactions

The processes of deciding on games and materials included considering interaction styles being intuitive. We established in-game action-response consistency (e.g., children sat to imitate the sitting position). Interaction types were putting felt objects on the color sensor (LO), sitting and standing to imitate those positions (FP), conforming to rules by sitting or standing (DISD), and covering the light sensor on the left or right to sort falling

objects (OS). Video recording analysis showed that the children engaged with the developed e-textiles without difficulty.

# 3.2.2. Size of the materials

The sizing of materials is dependent on their being wearable or not. Four of the six teachers in the study expressed that non-wearable materials should be sufficiently large to allow for easy interactions, and that children should be able to use it on their own. One teacher put it this way: "I think the [level-setting] button is small... I do not want everything to be under the control of the teacher. First, the teacher teaches, then we [teachers] help the children. After that, children start to do it by themselves." Therefore, we redesigned considering the feedback received. Figure A3 and Figure A4 in the Appendix illustrate the dimensions of the first and second versions of the LO game material. Six children in Cycle 2 were asked to use the level-setting button and all of them used it with ease.

When it comes to wearables, however, the size suggestion issue is reversed because wearables affect mobility and comfort. We developed two forms of wearables, a thigh band and a belt bag. The size of the belt bag did not cause any problems and was kept the same in both cycles (see Figure A7). However, the thigh band was changed considerably in Cycle 2 (see Figure A5 and Figure A6), and which resulted in fewer problems (i.e., shifting/slipping due to loose clothes worn by some children) noted after making the band more compact, and no further or repeat mobility difficulties were observed. Additionally, two of the participant teachers commented on the second version thigh band and confirmed its size to be more appropriate.

## 3.2.3. Perceived responsiveness

The way that interaction took place between the children and the materials they used affected the perceived responsiveness of the materials. Data extracted from the video recordings revealed that the children applied too much pressure on the sensors, hit, or rubbed them when they received delayed or no feedback at all. In two of the games (LO and OS), the children exhibited these behaviors. In the LO, the color sensor was not found to work as intended, mainly due to the incorrect placement of objects on the sensor, and in the OS, the children considered the material to be unresponsive due to receiving feedback one second after a light sensor having been covered.

### 3.2.4. Sturdiness and stability

Analyses of the video recordings and interviews conducted with the teachers revealed that both the wearable and non-wearable materials need to be sturdy and stable. Sturdiness is deemed a critical aspect of materials in this context considering the target group's age-specific characteristics. E-textile materials can consist of small electronic components that could inadvertently be swallowed or become lodged in a child's windpipe. The electronic components used in the initial prototype materials were considered vulnerable since they were left exposed. The material prepared for the *LO* game had the electronic platform and the circuits uncovered, the 9-DOF IMU and the electronic platform in the second version of the thigh band were left unprotected, and the light sensors of the first version of the belt bag were framed but also left uncovered. One of the teachers noted that the material should be intact, sturdy, and that it should be hard to tear off any component. The teacher stated that; *"Another thing is that no parts of the material should be able to fall off or disintegrate. They [the small pieces on the material] are very prone to being swallowed by young children."* Another teacher highlighted the curiosity of preschool-aged children: *"Keep in mind that, after a while, children will often become curious about the material and will naturally insert their fingers into it and attempt to break it."* 

Several design changes were applied to ensure that the materials used in Cycle 2 were robust and safe. These considerations are listed as follows:

- Strengthening the textile base that contains electronic components (e.g., five felt layers were used for the UFO material).
- Hiding adjunct components (i.e., electronic platform, conductive thread-made circuits, Bluetooth module, battery, Piezo buzzer element, wires, some sensors) in the design. Additional textile layers (e.g., felt) or plastic covers could be options.
- Protecting electronic components that children interact with (e.g., light sensors framed by felt circles and covered with a transparent film).

Stability refers to a wearable material's ability to maintain its position as intended. The thigh band, for example, required improvements in order for it to safely remain in place on the thigh without slipping or shifting position. In Cycle 1, the thigh band continually slipped, which interfered with two of the games (*FP* and *DISD*). In Cycle 2, the thigh band was subsequently minimized and the surface in contact with the wearer's clothing was redesigned (i.e., hot glue was used to increase friction at the contact surface). In this way, the redesigned version in Cycle 2 reportedly presented fewer stability problems.

#### 3.3. Visual design of the developed materials

The study's findings indicated that visual design is a critical component of e-textile development. The common areas where issues related to visual design are: Esthetics, having a context or theme, and using visual elements as cues.

#### 3.3.1. Esthetics

Esthetics refers to the appearance of a material and relates to its appeal to and perceived pleasure of its users. One of the teachers referred to the characteristics of preschool-aged children and indicated that, "Since we are working with children, the material [LO v1] could be improved visually." Another teacher recommended improving the thigh band visually, expressed that; "If you wanted to place the material on the market, a visually improved material would be better. It seems more like a piece of a machine right now." We subsequently applied two improvements to the materials:

- Hiding adjunct components also provided a simplified visual design.
- Colors used were reconsidered (color palette was used to select colors).

Figure 6 shows the developed materials from the first and second iterations.



#### Figure 6. Visual improvements on both versions of e-textiles

#### 3.3.2. Having a context or theme

The participant teachers and the academic suggested that the *LO* game material would be enhanced if it had some aspect of visual integrity. In other words, using components in a way that represents a certain context or has a particular story can make a material appear more attractive. In the first version of the *LO* material, for example, a teacher liked the idea of putting lights inside the bus and using them as windows. For the second design, the academic demonstrated how the material could be shaped like a UFO, adding a space-themed felt for the background, and using hand-shaped stick-on felt shapes. After seeing the second version of the material, one of the teachers commented that, *"I like the design of the material; it is cute. Space is an engaging subject for children. When they see the UFO, they will like it."* Adopting a theme or context allowed for the storification of

the game rules. The participant academic and one of the teachers composed stories for the *LO* game. The teacher's story was about aliens who would climb aboard the UFO, but in a certain order. The lights on the UFO therefore show the correct order, whilst the hand-shaped felt objects represent the aliens' hands. Our experiences with the children also showed that adopting a theme or context helped to make explaining games that much easier, and were perceived as being more attractive.

## 3.3.3. Visual elements as cues

Visuals can contribute to gameplay in good designs, and equally make a game more complex in poor designs. In both versions of the *LO*, several of the children started from the right-hand side in verbalizing the color sequence although the game actually required the opposite. One of the teachers suggested using visual cues such as redesigning the game so that the lights start from the left and turn on one by one, saying: *"The lights can turn on one by one to show in which order the colors will be kept in mind."* She also noted that an arrow sign that points to the starting point of the Neopixel row could be added.

The designs ensured that children could see and realize each of the interactive components. Frames used for sensors worked as visual clues since they showed where to cover or place the felt objects. Visuals can also be used to promote the correct attachment of wearables.

# 3.4. Perceived usefulness and instructional quality of e-textiles and games

The participant teachers also mentioned how they perceived the instructional quality of each of the games and materials during their interviews.

## 3.4.1. Active participation and motivation

The teachers stated the features that they liked in each game. For the LO, one of the teachers noted that the game being electronic and having lights could help to draw the attention of the children. For the FP, another of the teachers stated that typically children's focus will be drawn to a computer or television, and then passively watch whatever is playing on the screen. The same teacher expressed her appreciation that the FP did not appear to isolate students from their surroundings, stating that:

I especially like this in the material: Normally, children focus on something on the computer or television and watch what is playing [...] However, in this activity, children wear a thigh band and follow the instructions provided while watching. That is, children do not just sit and watch. It is very nice; indeed, I wish more children could be provided with this kind of game.

Another teacher stated that bodily interaction can also attract the attention of children, which can make games appear more entertaining to them.

#### 3.4.2. Creativity of the materials

The teachers stated that the materials were both different and creative. One of the teachers indicated her views on the LO game, saying; "I think that the idea of detecting colors is both different and creative." Another teacher specified that the use of body movements to create patterns in the FP was considered different, noting that; "Actually, movement is nice because the shape, color, and number patterns are common. This is a bit different, but it can be diversified." Finally, a third teacher indicated her views on the DISD game, saying that "The logic of the game is excellent, and I think there is nothing similar on the market. I have not seen anything like this."

# 4. Discussion and conclusion

In this study, DBR processes were conducted with the aim to reveal the design principles and lessons learned while designing educational e-textiles for preschool-aged children, as well as the affordances and constraints of e-textiles. The study's results suggest that e-textiles can be adopted for the development of games and materials aimed at preschool-aged children; however, several considerations should be made besides concerning children's

developmental characteristics: (a) Technical capabilities of electronic components; (b) usability of developed materials; and, (c) visual design features. Also revealed were the participant preschool teachers' perceptions towards e-textile materials and educational games through game-material systems exemplified in the current study.

Although e-textile technology can be seen as a way of developing new intuitive means of interaction (De La Guia et al., 2016), it is still not considered to be that mature and posits certain technical limitations (Kan & Lam, 2021). Therefore, knowing component capabilities and their limitations plays an essential role in any materials development (McCann et al., 2005). For example, while Bluetooth technology can be used to provide interactions in e-textile projects (Gonçalves et al., 2018), the use of conductive thread may result in short circuits (Peppler & Glosson, 2013). To give another example, knowing that color sensor readings may be liable to change according to distance, area of view, and environment lighting (Earl, 2013) can be beneficial to the utilization of the component. The current study also reports additional issues with regards to the use of e-textile components in the study.

Game mechanisms and the design of materials are highly dependent on selected e-textile components. Two features of e-textile components are prominent. First, e-textile technology provides many different interaction options. This level of flexibility provides interactions that are more prone to "natural play behavior" (Rosales et al., 2015, p. 47). Second, e-textile technology allows accessories to function beyond their authentic purposes (Rosales et al., 2015) and enables turning everyday objects into interaction means and thus, increases intuitiveness, familiarity, and friendliness (Vega-Barbas et al., 2015). The perceptions of the participant teachers also highlighted the potential of e-textiles being used creatively to increase children's motivation, and in helping facilitate a more active level of participation.

Usability is critical in e-textile design (Fernández-Caramés & Fraga-Lamas, 2018). Supporting natural (Peppler & Danish, 2013) play-like actions, e-textiles also enable intuitive interactions, which is also important for usability as noted by Kan and Lam (2021). Material size is an important consideration, especially when developing size-suitable materials to be worn by preschool children that need to be readily adjustable, easy to use, and comfortable. According to Wright and Keith (2014), a wearable device should be comfortable by its very definition. However, being small in size can also be a problem if young children are going to be using the material themselves, as Kazemitabaar et al. (2017) stated. As the fine motor skills of children have yet to mature at the preschool age; the objects and components that they interact with should be sized accordingly. Another issue regarding usability is the perceived responsiveness of the materials and thus the degree of children's perceptions regarding the material's ability to produce a timely response. Any design or technical issues that produce response delays can lead to decreased motivation in child users (Kara & Cagiltay, 2020); this in turn can lead to repeated actions, applying more force to the material than necessary, or extended action duration (Dakova & Dumont, 2014). Finally, sturdiness and stability are the key features that any e-textiles should offer. The materials should be durable (Ismar et al., 2020; Kan & Lam, 2021; Kazemitabaar et al., 2017) in terms of both safety and sustainability, and stable so as not to disrupt the flow of the game.

The visual/esthetic design of e-textile materials is another factor to be considered (Chen, 2020; Kan & Lam, 2021; McCann et al., 2005). Clothing forms part of human fashion, and esthetics should therefore be considered together with functionality (Kafai & Peppler, 2014). Several guiding principles can be emphasized regarding designing visually pleasing e-textile materials:

- The esthetic design of the material should be simple, and if possible, non-interactive components (e.g., electronic platforms and the power source) should be hidden within the design. Honauer et al. (2019) also recommend hiding all electronics within the materials.
- The placement of components should be appropriate for the intended purpose of the game. For example, if children need to see Neopixels to play the game, they should be apparent in the design.
- Appropriate colors that do not interfere with the gameplay (i.e., in desktop games) should be utilized in the construction of materials.
- An exciting, stimulating context or a theme for desktop games can help to make games more attractive and can therefore facilitate game introduction (e.g., rules) through supporting storification.
- Visuals on e-textile materials can be designed in such a way that they contribute to the gameplay itself (e.g., indicating game rules or correctly attaching a wearable e-textile).

E-textiles can be used to create educational materials that are tailored to meet the needs of preschoolers by considering the various design principles and issues highlighted in the current study. Preschool teachers' attitudes toward the use of e-textile technologies in preschool education seem to support this idea. The principles and issues presented in this study may also apply to similar tangible preschool materials. Although the designing

and developing of educational e-textile materials require certain technical skills (e.g., Peppler & Danish, 2013), enabling various bodily interactions, the ability to enrich everyday objects with interactivity, and maintaining a physical world connection whilst engaging with them seem to present significant advantages of utilizing e-textile materials.

Whilst investigating the instructional effectiveness of the developed e-textile materials was beyond the scope of the current study, future studies could aim to explore this issue.

# Acknowledgment

We would like to thank Dr. Elif Manuoğlu for her assistance on this study.

# References

Almusawi, H. A., Durugbo, C. M., & Bugawa, A. M. (2021). Innovation in physical education: Teachers' perspectives on readiness for wearable technology integration. *Computers and Education*, 167, Article 104185. https://doi.org/10.1016/j.compedu.2021.104185

Amiel, T., & Reeves, T. C. (2008). Design-based research and educational technology: Rethinking technology and the research agenda. *Educational Technology & Society*, 11(4), 29-40. https://www.j-ets.net/collection/published-issues/11\_4

Anderson, T., & Shattuck, J. (2012). Design-based research: A Decade of progress in education research? *Educational Researcher*, 41(1), 16-25. https://doi.org/10.3102/0013189X11428813

Balestrini, M., Hernández-Leo, D., Nieves, R., & Blat, J. (2014). Technology-supported orchestration matters: Outperforming paper-based scripting in a jigsaw classroom. *IEEE Transactions on Learning Technologies*, 7(1), 17-30. https://doi.org/10.1109/TLT.2013.33

Best, J. R. (2012). Exergaming immediately enhances children's executive function. *Developmental Psychology*, 48(5), 1501-1510. https://doi.org/10.1037/a0026648

Center on the Developing Child at Harvard University. (2011). Building the brain's "air traffic control" system: How early experiences shape the development of executive function: Working Paper No. 11. http://www.developingchild.harvard.edu/

Chen, A. (2020). The Design and creation of tactile knitted e-textiles for interactive applications. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction* (pp. 905-909). ACM. https://doi.org/10.1145/3374920.3374959

Chen, A., Tan, J., & Henry, P. (2021). E-textile design through the lens of affordance. *Journal of Textile Design Research and Practice*, 9(2), 164-183. https://doi.org/10.1080/20511787.2021.1935110

Coccia, A., Amitrano, F., Donisi, L., Cesarelli, G., Pagano, G., Cesarelli, M., & D'Addio, G. (2021). Design and validation of an e-textile-based wearable system for remote health monitoring. *ACTA IMEKO*, *10*(2), 220-229. https://doi.org/10.21014/acta\_imeko.v10i2.912

Collins, A., Joseph, D., & Bielaczyc, K. (2004). Design research: Theoretical and methodological issues. *Journal of the Learning Sciences*, 13(1), 15-42. https://doi.org/10.1207/s15327809jls1301\_2

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches* (4<sup>th</sup> ed.). Sage.

Dakova, S., & Dumont, N. (2014). An Overview of textile interfaces. https://hci.rwth-aachen.de/index.php?option=com\_attachments&task=download&id=1176

De La Guia, E., Camacho, V. L., Orozco-Barbosa, L., Brea Lujan, V. M., Penichet, V. M. R., & Lozano Perez, M. (2016). Introducing IoT and wearable technologies into task-based language learning for young children. *IEEE Transactions on Learning Technologies*, 9(4), 366-378. https://doi.org/10.1109/TLT.2016.2557333

Derry, S. J., Pea, R. D., Barron, B., Engle, R. A., Erickson, F., Goldman, R., Hall, R., Koschmann, T., Lemke, J. L., Sherin, M. G., & Sherin, B. L. (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences*, *19*(1), 3-53. https://doi.org/10.1080/10508400903452884

Diamond, A. (2012). Activities and programs that improve children's executive functions. *Current Directions in Psychological Science*, 21(5), 335-341. https://doi.org/10.1177%2F0963721412453722

Diamond, A., & Ling, D. S. (2016). Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not. *Developmental Cognitive Neuroscience*, *18*, 34-48. https://doi.org/10.1016/j.dcn.2015.11.005 Doménech, J., Ferri, J., Costa, R., Oliveira, P., Grilo, A., Cardon, G., DeSmet, A., Schwarz, A., Stragier, J., Pomazanskyi, A., & Danilins, J. (2018). SmartLife – Exergames and smart textiles to promote energy-related behaviours among adolescents. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11243 LNCS, 288-293. https://doi.org/10.1007/978-3-030-02762-9\_31

Earl, B. (2013, May 21). Adafruit color sensors. Adafruit. https://learn.adafruit.com/adafruit-color-sensors?view=all#arduino-code

Erickson, F. (2006). Definition and analysis of data from videotape: Some research procedures and their rationales. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Handbook of complementary methods in education research* (pp. 177-205). Erlbaum. https://doi.org/10.4324/9780203874769

Fernández-Caramés, T. M., & Fraga-Lamas, P. (2018). Towards the internet-of-smart-clothing: A Review on IoT wearables and garments for creating intelligent connected e-textiles. *Electronics*, 7(12), Article 405. https://doi.org/10.3390/electronics7120405

Fields, D. A., Kafai, Y. B., Morales-Navarro, L., & Walker, J. T. (2021). Debugging by design: A Constructionist approach to high school students' crafting and coding of electronic textiles as failure artefacts. *British Journal of Educational Technology*, 52(3), 1078-1092. https://doi.org/10.1111/bjet.13079

Gao, Z., Lee, J. E., Zeng, N., Pope, Z. C., Zhang, Y., & Li, X. (2019). Home-based exergaming on preschoolers' energy expenditure, cardiovascular fitness, body mass index and cognitive flexibility: A Randomized controlled trial. *Journal of Clinical Medicine*, 8(10), Article 1745. https://doi.org/10.3390/jcm8101745

Gashaj, V., Dapp, L. C., Trninic, D., & Roebers, C. M. (2021). The Effect of video games, exergames and board games on executive functions in kindergarten and 2nd grade: An Explorative longitudinal study. *Trends in Neuroscience and Education*, 25, Article 100162. https://doi.org/10.1016/j.tine.2021.100162

Gonçalves, C., da Silva, A. F., Gomes, J., & Simoes, R. (2018). Wearable e-textile technologies: A Review on sensors, actuators and control elements. *Inventions*, *3*(1), Article 14. https://doi.org/10.3390/inventions3010014

Grant, D. A., & Berg, E. (1948). A Behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, *38*(4), 404-411. https://doi.org/10.1037/h0059831

Honauer, M., Moorthy, P., & Hornecker, E. (2019). Interactive soft toys for infants and toddlers – Design recommendations for age-appropriate play. In *CHI PLAY 2019 - Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (pp. 265-276). ACM. https://doi.org/10.1145/3311350.3347147

Ismar, E., Kurşun Bahadir, S., Kalaoglu, F., & Koncar, V. (2020). Futuristic clothes: Electronic textiles and wearable technologies. *Global Challenges*, 4(7), Article 1900092. https://doi.org/10.1002/gch2.201900092

Kafai, Y. B. & Peppler, K. A. (2014). Transparency reconsidered: Creative, critical, and connected making with e-textiles. In M. Ratto & M. Boler (Eds.), *DIY citizenship: Critical making and social media* (pp. 179-188). MIT Press.

Kan, C., & Lam, Y. (2021). Future trend in wearable electronics in the textile industry. *Applied Sciences*, *11*(9), Article 3914. https://doi.org/10.4324/9781315232140-14

Kara, N., & Cagiltay, K. (2020). Smart toys for preschool children: A Design and development research. *Electronic Commerce Research and Applications*, *39*(2), Article 100909. https://doi.org/10.1016/j.elerap.2019.100909

Kazemitabaar, M., McPeak, J., Jiao, A., He, L., Outing, T., & Froehlich, J. E. (2017). MakerWear: A Tangible approach to interactive wearable creation for children. In *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 133-145). ACM. https://doi.org/10.1145/3025453.3025887

Komolafe, A., Torah, R., Wei, Y., Nunes-Matos, H., Li, M., Hardy, D., Dias, T., Tudor, M., & Beeby, S. (2019). Integrating flexible filament circuits for e-textile applications. *Advanced Materials Technologies*, 4(7), Article 1900176. https://doi.org/10.1002/admt.201900176

López-Serrano, S., Ruiz-Ariza, A., De La Torre-Cruz, M., & Martínez-López, E. J. (2021). Improving cognition in school children and adolescents through exergames. A Systematic review and practical guide. *South African Journal of Education*, *41*(1), Article 1838. https://doi.org/10.15700/saje.v41n1a1838

McCann, J., Hurford, R., & Martin, A. (2005). A Design process for the development of innovative smart clothing that addresses end-user needs from technical, functional, aesthetic and cultural viewpoints. In *Proceedings – Ninth IEEE International Symposium on Wearable Computers* (pp. 70-77). IEEE. https://doi.org/10.1109/ISWC.2005.3

Merriam, S. B. (2009). Qualitative research: A Guide to design and implementation. Jossey Bass.

Nonnis, A., & Bryan-Kinns, N. (2019). Mazi: A Tangible toy for collaborative play between children with autism. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children* (pp. 672-675). ACM. https://doi.org/10.1145/3311927.3325340

Norooz, L., Mauriello, M. L., Jorgensen, A., McNally, B., & Froehlich, J. E. (2015). Body Vis: A New approach to body learning through wearable sensing and visualization. In *Proceedings of the 33<sup>rd</sup> ACM Conference on Human Factors in Computing Systems* (pp. 1025-1034). ACM. https://doi.org/10.1145/2702123.2702299

Patton, M. Q. (1990). Qualitative evaluation and research methods (2nd ed.). Sage.

Penn-Edwards, S. (2004). Visual evidence in qualitative research: The Role of videorecording. *The Qualitative Report*, 9(2), 266-277. http://nsuworks.nova.edu/tqr/vol9/iss2/5

Peppler, K., & Danish, J. A. (2013). E-textiles for educators: Participatory simulations with e-puppetry. In L. Buechley, K. Peppler, M. Eisenberg, & Y. Kafai (Eds.), *Textile messages: Dispatches from the world of e-textiles and education* (pp. 133-141). Lang.

Peppler, K., & Glosson, D. (2013). Stitching circuits: Learning about circuitry through e-textile materials. *Journal of Science Education and Technology*, 22(5), 751-763. https://doi.org/10.1007/s10956-012-9428-2

Rafiei Milajerdi, H., Sheikh, M., Najafabadi, M. G., Saghaei, B., Naghdi, N., & Dewey, D. (2021). The Effects of physical activity and exergaming on motor skills and executive functions in children with autism spectrum disorder. *Games for Health Journal*, *10*(1), 33-42. https://doi.org/10.1089/g4h.2019.0180

Rosales, A., Sayago, S., & Blat, J. (2015). Beeping socks and chirping arm bands: Wearables that foster free play. *Computer*, 48(6), 41-48. https://doi.org/10.1109/MC.2015.168

The Design-based Research Collective. (2003). Design-based research: An Emerging paradigm for educational inquiry. *Educational Researcher*, *32*(1), 5-8. https://doi.org/10.3102%2F0013189X032001005

Ugur Yavuz, S., Veske, P., Scholz, B., Honauer, M., & Kuusk, K. (2021). Design for playfulness with interactive soft materials: Description document. In R. Wimmer (Ed.), *Proceedings of the 15th International Conference on Tangible, Embedded, and Embodied Interaction* (Article 67). ACM. https://doi.org/10.1145/3430524.3442702

Vega-Barbas, M., Pau, I., Ferreira, J., Lebis, E., & Seoane, F. (2015). Utilizing smart textiles-enabled sensorized toy and playful interactions for assessment of psychomotor development on children. *Journal of Sensors*, 2015, Article 898047. https://doi.org/10.1155/2015/898047

Wright, R., & Keith, L. (2014). Wearable technology: If the tech fits, wear it. *Journal of Electronic Resources in Medical Libraries*, 11(4), 204-216. https://doi.org/10.1080/15424065.2014.969051

Xiong, S., Zhang, P., & Gao, Z. (2019). Effects of exergaming on preschoolers' executive functions and perceived competence: A Pilot randomized trial. *Journal of Clinical Medicine*, 8(4), Article 469. https://doi.org/10.3390/jcm8040469

Zelazo, P. D. (2006). The Dimensional Change Card Sort (DCCS): A Method of assessing executive function in children. *Nature Protocols*, *1*(1), 297-301. https://doi.org/10.1038/nprot.2006.46

# Appendix A



Figure A2. Child playing DISD game (second version, screenshot displayed top-right)





*Figure A4.* LO game material dimensions (second version) **70 cm** -----









Figure A7. Belt bag dimensions (second version only differed by frame size)



# Effects of Mobile-Assisted Language Learning on EFL Learners' Listening Skill Development

# Rui Li

School of Foreign Languages, Hunan University, Hunan, China // liruidianzi@hotmail.com

(Submitted March 24, 2022; Revised August 16, 2022; Accepted August 18, 2022)

**ABSTRACT:** Although an increasing number of studies have focused on the use of mobile-assisted language learning (MALL) for English as a foreign language (EFL) learners' listening skill development, there is a lack of comprehensive meta-analysis regarding the effect sizes of these studies. To fill the gap, 20 selected experimental studies involving 1218 participants were included for a meta-analysis based on the proposed inclusion and exclusion criteria. Results showed that the overall effect size was moderate-to-large, g = 0.792, 95% CI [0.536, 1.047], suggesting that MALL for EFL learners' listening skill development is more effective than traditional methods. Regarding moderators for the overall effect, different moderator effects of educational levels, software types, control conditions, intervention settings, measured outcome types and intervention durations were reported. Specifically, educational levels were found to be a significant moderator, while software types, control conditions, intervention settings, measured outcome types and intervention durations were not the significant moderators. The implications for practice were discussed as well.

**Keywords:** English as a foreign language (EFL), Evidence-based applied linguistics (EBAL), Listening skill, Meta-analysis, Mobile-assisted language learning (MALL)

# 1. Introduction

The development of English as a foreign/second language (EFL/ESL) listening skill "is seen not only as something valuable for its own sake but as something that supports the growth of other aspects of language use, such as speaking and reading" (Richards, 2005, p. 85). Currently, there have been a positive change in the teaching of EFL listening comprehension, but learners still confront with such difficulties as insufficient classroom instructions and a lack of sustained listening practice to guide them to be successful listeners (Hwang et al., 2016; Vandergrift & Goh, 2012). In recent years, the popularity of mobile-assisted language learning (MALL) technologies has reshaped the traditional FL listening instructions, since "listening comprehension skill is an invisible mental process that is more difficult to develop" (Mendoza et al., 2020, p. 61), and compared to other language skills MALL is mostly used to autonomously and ubiquitously develop learners' listening skill with sufficient exposure of multimodal listening materials (Li, 2022a; Li, 2022b). Motivated by the possibility of overcoming the barriers of insufficient classroom instructions and the poverty of listening practice, pedagogical potentials of MALL for EFL learners' listening skill development have been well-recognized to create ubiquitous learning environments.

While researchers (Alabsi, 2020; Liu et al., 2018; Saeedakhtar et al., 2021; Tai & Chen, 2021) have claimed that MALL can effectively improve learners' listening skill, a quantitative meta-analysis of the overall effect size among these studies is still in paucity. Drawing on evidence-based applied linguistics (EBAL), the effects of a pedagogical intervention should be supported with sound evidence available in foreign language education. In other words, by combining the results detailed in multiple studies with larger sample sizes (Li, 2022a), a meta-analysis from the drive of evidenced-based practice will afford more accurate estimates of the effects of MALL for EFL learners' listening skill development, and offer a deeper understanding of related moderators that potentially affect the overall effects. More specifically, this study aims to (a) generalize empirical findings of MALL for listening skill development research while (b) tackling the variability of the aggregated effects with a meta-analytic approach.

# 2. Literature review

# 2.1. Review of MALL for EFL listening skill development

With the rapid development of mobile technologies, such as smartphones, tablet PCs and other portable devices, pedagogical affordances of MALL have received considerable attention in the existing studies (e.g., Holden & Sykes, 2011; Lai et al., 2022; Thorne et al., 2021; Zheng et al., 2017; Zheng et al., 2018). Researchers have

maintained that MALL supports seamless or ubiquitous learning (Burston & Giannakou, 2022; Chen et al., 2020), overcomes a lack of individualized learning (Sung et al., 2016), and affords instant and timely feedback (Sung et al., 2015), etc. For years, while MALL has been documented to facilitate EFL learners' four main language skills (Chen et al., 2020; Sung et al., 2015; Sung et al., 2016), researchers have gradually realized the importance of using MALL to develop EFL learners' listening skill. More specifically, since listening comprehension has been considered the most difficult of the four main language skills (Siegel, 2014), the ubiquitous feature of MALL could afford learners with increased practice and multimodal materials. Furthermore, the individualized and interactive features of MALL could encourage learners to overcome the limitation of traditional classroom instructions (Li, 2021a).

MALL for EFL listening skill development is defined as the use of MALL devices, e.g., PDAs (Chang et al., 2018), MP3 (Rahimi & Soleymani, 2015), iPods (de la Fuente, 2014), mobile phones (Alabsi, 2020; Al-Shamsi et al., 2020) and mobile virtual reality (VR) gear (Tai & Chen, 2021), among others, to facilitate EFL learners' listening skill development. Researchers have conducted extensive experimental or quasi-experimental studies to compare the experimental group using MALL for EFL listening, such as mobile captions and subtitles (Alabsi, 2020), micro dialogues (de la Fuente, 2014), mobile English listening system (Liu et al., 2018), mobile VR (Tai & Chen, 2021), and portable media players (Rashtchi & Mazraehno, 2019) and the control group using traditional approaches, e.g., conventional paper-and-pencil (Azar & Nasiri, 2014), DVDs and audio CDs from the book (Rashtchi & Mazraehno, 2019) and indoor computer classroom (Chang et al., 2018).

While an emerging array of empirical studies on MALL for EFL listening skill development, researchers have resulted in mixed findings. Some have found that MALL can effectively improve learners' listening comprehension skill (Alabsi, 2020; Liu et al., 2018; Saeedakhtar et al., 2021; Tai & Chen, 2021). Despite the promising results, not everyone could be convinced, as other researchers (e.g., Hwang et al., 2016; Rashtchi & Mazraehno, 2019; Seo & Choi, 2014) have obtained the limited effects of MALL for EFL listening skill development. Besides the aforementioned empirical studies, researchers have also conducted reviews in an attempt to gain a holistic understanding of the current state. To our knowledge, only Coşkun and Marlowe (2020) began to adopt a narrative review and investigate the use of MALL for EFL listening skill development. From their review, they summarized that most studies reported the facilitative effects.

Taken together, although the aforementioned studies are valuable to shed light on MALL for EFL listening research, several issues remain open for debate. On the one hand, given the growing number of empirical studies that have progressively accumulated sufficient data sources, the mixed results may justify the need to scrutinize the pedagogical effects of MALL on EFL learners' listening skill development and which moderators may underline such discrepancy. Compared to those empirical studies, meta-analytic results based on multiple studies and increased sample sizes are more reliable and generalizable (Li, 2022d; Li, in press). On the other hand, besides the empirical studies, while the qualitative review (Coşkun & Marlowe, 2020) might provide a comprehensive vision, little is still known about the effect size with a quantitative meta-analytic approach, and how the effect of MALL for listening skill development was affected by some possible moderators. Given that understanding the overall effect and moderator effect could not only index the effectiveness of an intervention, but also inform pedagogy, it is necessary to conduct a meta-analysis that examines the effectiveness and potential moderators on MALL for EFL learners' listening skill development.

#### 2.2. Moderators of MALL for EFL listening skill development

The meta-analysis regarding the effectiveness of MALL for listening skill development is not a simple blackand-white issue, rather rigorous analysis of potential factors that can moderate the effectiveness is also required. The selection of moderators was based on the existing literature and widely referred to previous studies (Chen et al., 2020; Sung et al., 2015; Sung et al., 2016), resulting in the following six moderators: educational levels, software types, control conditions, intervention settings, intervention durations and measured outcome types.

*Educational levels*. Participants of different educational levels tend to have different learning outcomes. For instance, Costabile and colleagues (2003) employed 9 to 10-year-old children as participants of primary educational level to evaluate the effects of a tutoring hypermedia and found children's affective perceptions predict their learning outcomes. By contrast, in a quasi-experiment, Li (2021a) recruited 17 to 21-year-old college EFL learners to understand the effects of game-based vocabulary learning on their vocabulary achievement, motivation and self-confidence. Results indicated that those participants of tertiary educational level might not care too much about the entertainment feature of games that tend to arouse their affective perceptions, implying learner educational level might be an important moderator. In this study, we took

educational levels (viz. primary, secondary and tertiary education) as a moderator to test whether there exists moderator effect of educational levels on the MALL for EFL learners' listening skill development.

*Software types*. Depending on the specific educational purposes of MALL technologies, the software could be further categorized into general purposes (technologies that were designed for non-educational purposes, e.g., MP3 and iPods) and educational purposes (technologies that were designed for educational purposes, e.g., listening management system). It is generally believed that domain-specific MALL technologies developed for educational purposes are better tailored to EFL learners' individualized needs compared to those of general purposes (Chen et al., 2020; Sung et al., 2016). However, it remains largely unclear whether the research findings of the domain-general MALL technologies could be generalizable to the domain-specific type (Li, 2022c), viz. MALL for listening skill development in particular.

*Control conditions*. While participants of the experimental group who adopted MALL technologies have been extensively described, those of the control group should be considered with caution as well. Thus, two main control conditions are classified: paper-and-pencil (participants of the control group who used traditional paper-and-pencil method to practice listening) and computer-based methods (participants of the control group who used traditional computer-based method, e.g., watching videos, DVDs and CDs, to practice listening). Importantly, as all effect sizes are potentially caused by a comparison with a control condition, it is very important to understand what the control conditions look like in the studies reviewed. As such, the moderator effect of control conditions deserves further scrutiny in this study.

*Intervention settings*. Since EFL learning or teaching activities might occur in the classroom or outside of the classroom, intervention settings of MALL for listening skill development are classified into two kinds: classroom and outdoor (Sung et al., 2015). The moderator effect of intervention settings receives considerable attention in the existing MALL studies (Chen et al., 2020; Sung et al., 2015; Sung et al., 2016). While these studies found the stronger effect of learning in outdoor settings than classroom settings, it is still unclear whether similar results could be applicable to the domain-specific EFL listening comprehension. In this study, we examine the moderator effect of intervention settings on the effectiveness of MALL for EFL listening skill development.

*Intervention durations*. The consensus regarding the moderator effect of intervention durations has not yet been reached thus far. For instance, in a systematic review of MALL research, Hwang and Fu (2019) argued that longer treatment durations can fully represent the real effects of MALL on learning outcomes. However, in a recent meta-analysis of MALL, Chen and colleagues (2020) claimed that shorter-term interventions yielded larger effect sizes than longer-term ones, because EFL learners' curiosity might not be sustain for a long period of time. As such, we examine the moderator effect of intervention durations in this study.

*Measured outcome types*. According to Xu and colleagues (2019), the heterogeneity of measured outcomes might impact the results of an intervention. In this study, we categorize two measured outcome types, viz. standardized vs. researcher-designed, to understand the moderator effect of measured outcome types.

#### 2.3. Research statements and questions

This study is designed to provide a brand-new quantitative perspective on MALL for EFL listening skill development, using a meta-analysis of rigorously peer-reviewed empirical research. The research purposes are twofold: First, drawing on the data collected from the primary studies, we conduct a meta-analysis to calculate the aggregated overall effect regarding MALL for EFL listening skill development. Second, according to similar meta-analyses (Chen et al., 2020; Sung et al., 2015; Sung et al., 2016), the moderator effects, such as educational levels, software types, control conditions, intervention settings, intervention durations and measured outcome types, are analyzed as well. Consequently, two research questions to be addressed are as follows.

*Research question 1*: What is the overall effect size of MALL vs. non-MALL for EFL learners' listening skill development?

**Research question 2**: How do moderators, such as educational levels, software types, control conditions, intervention settings, intervention durations and measured outcome types, affect the aggregated effect on listening skill development?

# 3. Research design

This study aimed to meta-analyze the experimental or quasi-experimental studies on MALL for EFL listening skill development. Specifically, it firstly dealt with the overall effect of MALL for EFL listening; and secondly it calculated the moderator effects of the related moderators, such as educational levels, software types, control conditions, intervention settings, intervention durations and measured outcome types, on the overall effect size. As a result, practical implications could be obtained from the meta-analytic findings.

## **3.1. Data collection source**

To exhaustively retrieve the related primary studies on MALL for EFL listening skill development, data collection procedures were observed as follows: First, data were searched from several electronic online databases (Web of Science, ProQuest, Scopus and ERIC), publishers (e.g., ScienceDirect, Springer, SAGE, Taylor & Francis and Wiley) and search engines (Google Scholar and Baidu Scholar) by using a combination of the following MALL-related and listening-related keywords integrated with Boolean operators, i.e., (mobileassisted language learning OR MALL OR mobile applications OR portable devices OR handheld devices OR mobile technologies OR mobile learning OR m-learning OR seamless learning OR ubiquitous learning OR ulearning OR mobile phone OR cell phone OR smartphone OR iPod OR iPhone OR tablets OR MP3 OR personal digital assistants OR PDAs OR podcasts) AND (listen OR listening comprehension OR listening skill OR listening ability OR listening performance OR listening score OR listening test). Then, to further avoid the insufficient search of a significant portion of the relevant literature in the first-round, a second-round backward and forward citation search was conducted with snowballing technique (Biernacki & Waldorf, 1981) by scanning references in some review articles (e.g., Chen et al., 2020; Coşkun & Marlowe, 2020; Jia & Hew, 2021). Third, each of the following EdTech journals (Computers & Education, Internet and Higher Education, British Journal of Educational Technology, Computers in Human Behavior, Interactive Learning Environments, Educational Technology Research and Development, Educational Technology & Society, Journal of Computing in Higher Education, Journal of Educational Computing Research, Journal of Computer Assisted Learning, Australian Journal of Educational Technology, and The Asia-Pacific Education Researcher, among others) and language education journals (Computer Assisted Language Learning, Language Learning & Technology, ReCALL, System and CALICO Journal) was manually searched to avoid the incomplete inclusion. The initial retrieval resulted in 74 primary studies after duplicates removed.

# **3.2. Inclusion and exclusion criteria**

To exclude irrelevant literature, two researchers independently and manually narrowed down the search to only cover the articles related to MALL for EFL listening skill development based on the inclusion and exclusion criteria:

(1) *Timespan constraints*. Given that studies of MALL for EFL listening skill development remained few in number before 2010 (Hwang & Fu, 2019), the studies published in English during 2010 to 2021 were considered.

(2) *Topic and publication type constraints*. The studies should adopt a form of MALL devices/tools (e.g., mobile phones, PDAs and iPods) on EFL/ESL listening skill development. Those studies that failed to use educational technologies, or used CALL technologies (e.g., computers and projectors), used MALL technologies on first language or other FL listening skill development were excluded. The studies should be rigorously peer-reviewed publications, including journal articles, book chapters and conference proceedings. Those non-peer-reviewed publications were excluded. Furthermore, other review publications (e.g., review articles, book reviews, and editorial materials) were excluded. As a result, those irrelevant (k = 31) were excluded, resulting in k = 43 left for further analysis.

(3) *Material constraints*. For the data transformation or effect size calculation, only the experimental or quasiexperimental studies that reported means, *SD*, and number of the participants were included. More specifically, the independent variables should include different interventional modes (e.g., traditional learning method as a control group vs. MALL method as a treatment group), and the dependent variable should include a measure of the researcher-designed or standardized pre- and post-tests on EFL listening comprehension performance between different modes. Among them, k = 25 studies were excluded for the following reasons: eight publications were about qualitative research on listening pedagogical or theoretical recommendations, five on EFL learners' attitudes or perceptions and twelve without sufficient information for effect size calculation. As a result, 20 eligible publications (18 articles and 2 conference proceedings) were finalized.

# 3.3. Study quality evaluation

Study quality is one of the most important issues to consider, as it "can affect study results, which can in turn affect the conclusions drawn. It is thus necessary to develop an explicit, transparent, and reproducible instrument of assessing study quality" (Valentine, 2019, p. 130). Informed by Valentine (2019), the commonly used study quality instrument, viz. the Medical Education Research Study Quality Instrument (MERSQI) (Reed et al., 2007), was adopted to evaluate the methodological quality of the selected studies. Although the MERSQI was originally developed to evaluate the study quality of medical education research, it has been increasingly proven to be discipline neutral and commonly applied in evaluating the study quality of educational research (e.g., Jensen & Konradsen, 2018). The MERSQI contains six domains, including study design, sampling, type of data, validity of evidence for evaluation instrument scores, data analysis and outcome. Each domain has a maximum score of 3, making a maximum total score of 18 and potential range 5 to 18. The average total score of 11 or so could be taken as the benchmark of satisfied study quality. For instance, Jensen and Konradsen (2018), in their meta-analysis, obtained the average total score of 10.9, with a range of 6 to 14.5. Likewise, Cook and Reed (2015) achieved the average score of 11.3, with a range of 8.9 to 15.1. Our meta-analysis showed that the average total score is 14.684 (SD = 0.682), with a range of 13.5 to 15, suggesting that the selected primary studies had the high quality for the coming meta-analysis.

## 3.4. Coding scheme

Drawing on the existing meta-analytic studies on MALL (Chen et al., 2020; Sung et al., 2015; Sung et al., 2016; Xu et al., 2019), all the 20 selected publications were coded with the following major categories (Table 1).

After the code scheme was developed, the following coding procedures were observed. First, given the recommended practice for data dependencies (Plonsky & Oswald, 2014), multiple effect sizes reported in a single publication involved different participants or different treatment interventions were coded separately to ensure the reliability of the analyses, resulting in 20 eligible primary studies that yielded 21 effect sizes as experimental studies. Second, two coders negotiated with each other to ensure the consistent understanding of each coding items. The discrepancies were resolved by consensus through discussions, along with any necessary reviews of the existing coding scheme.

Coding types	Subtypes	Operational definitions	References			
Educational levels	• primary education	Kindergarten or primary school students.	Chen et al. (2020)			
	<ul> <li>secondary education</li> </ul>	Junior middle school or senior high school students.				
	<ul> <li>tertiary education</li> </ul>	College students.				
Control conditions	• paper-and-pencil	Participants of the control group used traditional paper-and-pencil method to practice listening.	Researcher- designed			
	• computer-based	Participants of the control group used traditional computer-based method (e.g., watching videos, DVDs and CDs) to practice listening.				
Intervention settings	• classroom	Teaching/learning activities occurred in the classroom.	Chen et al. (2020)			
	• outdoor	Teaching/learning activities occurred outside of the classroom.				
Software types	• general purposes	Apps that were NOT designed for educational purposes, e.g., WhatsApp, captions and subtitles, podcasts and portable media players.	Chen et al. (2020)			
	<ul> <li>educational purposes</li> </ul>	al Apps that were designed for educational purposes e.g., mobile English listening system and mobile VR assisted listening system				
Measured outcome types	• standardized	Standardized TOEIC/IELTS/OPT listening comprehension scores.	Xu et al. (2019)			

<i>nume 1.</i> The descriptive mornation of coding scheme
---

	<ul> <li>researcher- designed</li> </ul>	Listening comprehension tests made by researchers.		
Intervention durations	<ul> <li>short (one session, ≤1 week)</li> </ul>	Durations fewer than one week or only one session.	Chen et (2020)	al.
	<ul> <li>intermediate (&gt; 1 week, ≤ 4 weeks)</li> </ul>	Durations over one week, but fewer than four weeks.		
	<ul> <li>long (&gt; 4 weeks, ≤ one semester)</li> </ul>	Durations over four weeks, but within one semester.		

*Note.* Apps = Applications; VR = virtual reality; TOEIC = Test of English for International Communication; IELTS = International English Language Testing System; OPT = Oxford Placement Test.

#### 3.5. Calculation and outlier diagnosis of the effect sizes

Hedges' g that could "provide a simple correct for the bias of small sample size" was taken as the effect size index (Lipsey & Wilson, 2001, p. 48). For most studies, Hedges' g was calculated based on means, sample sizes and standard deviations for the experimental and control group. When the means, sample sizes and standard deviations were not sufficiently reported, other statistical values, e.g., t-value, confidence interval, difference in means and sample sizes, to calculate effect sizes. For instance, we used means, sample sizes and p-values for both the experimental and control group (Hsu et al., 2013), and means, sample sizes and t-values (Hwang et al., 2014; Liu et al., 2018), to compute the effect sizes. Furthermore, a between-study Q test was executed to ensure whether the moderators played a role in the between-study heterogeneity (Li, 2022c).

Furthermore, according to Lipsey and Wilson (2001, p.108), potential outliers with the extreme effect sizes that were "more than 3 standard deviations from the mean of all the effect sizes" should be excluded. As such, one selected research (g = 3.926, Saeedakhtar et al., 2021) out of the 20 eligible studies that yielded extremely large effect sizes was excluded, resulting in 19 remaining primary studies that yielded 20 effect sizes for the final analysis.

#### 3.6. Publication bias analysis

The publication bias refers to the phenomenon that nonsignificant results are unlikely to be published, and the unpublished studies might differ from the published studies (Borenstein et al., 2005; Borenstein et al., 2009). According to some existing studies (Borenstein et al., 2005; Borenstein et al., 2009; Lipsey & Wilson, 2001), the inspection of publication bias was based on a fail-safe N method. In other words, by evaluating how many unpublished studies with nonsignificant results would change the meta-analytic results from significant to nonsignificant, Rosenthal (1991) proposed a classical fail-safe  $N(N_{fs})$  method to avoid the file-drawer problem. The calculation is as follows:

$$N_{fs} = (\frac{N_0}{Z_c^2})(N_0 \bar{Z}_0^2 - Z_c^2)$$

where  $N_0$  is the number of studies,  $Z_c$  is the critical value of Z, and  $\overline{Z_0}$  is the mean of Z obtained for the  $N_0$  studies. It indicated that there existed no publication bias, since the result of fail-safe N was 896, which was significantly higher than the respective observed number 20 (Z = 13.259, p < .001), that Rosenthal (1991) suggested for the file-drawer problem.

#### 4. Results

## 4.1. Overall analysis results

Twenty experimental studies involving 1218 participants were selected for the meta-analysis. The overall effect size was estimated using a *random effect model*, which "assumes that each observed effect size differs from the sampling error plus a value that represents other sources of variability" (Lipsey & Wilson, 2001, p. 119). As shown in Table 2, the overall effect size computed from 20 effect sizes is moderate-to-large, g = 0.792, 95% CI [0.536, 1.047] and significant, Z(19) = 6.072, p < .001, indicating the use of MALL for EFL learners' listening skill development is more effective than traditional methods.

Table 2. Overall effect size results of the included studies

		Heterog	geneity					
k	g	LL	UL	Z-value	Q-value	$I^2$		
20	0.792	0.536	1.047	6.072	93.318***	79.640		
Note, $k =$ number of effect sizes; $g =$ Hedges' g: LL = lower limit, UL = upper limit; 95 % CI = 95% confidence								

*Note.* k = number of effect sizes; g = Hedges' g; LL = lower limit, UL = upper limit; 95 % CI = 95% confidence interval; \*\*\*p < .001.

#### 4.2. Moderator analysis results

As apparent in Table 2, *Q*-value was 93.318 with p < 0.001, indicating that there were between-group differences among the effect sizes resulting from factors other than subject-level sampling error. The  $I^2$  for the overall model showed high heterogeneity ( $I^2 = 79.640$ ), indicating the need for moderator analysis (Borenstein et al., 2005; Borenstein et al., 2009). To achieve this aim, moderator analysis results were presented in Table 3.

Table 3. Moderator analysis results								
Moderators	k	g	95%	o CI	Z-value	Hete	rogeneity	у
			LL	UL	-	Q-value	df	р
Educational levels	20	0.590	0.418	0.762	6.717***	6.386*	2	0.041
1. primary education	4	0.485	0.247	0.724	3.990***	primary vs. se	condary,	p = 0.748
2. secondary education	3	0.415	0.059	0.772	$2.282^{*}$	secondary vs.	tertiary,	p = 0.028
3. tertiary education	13	0.974	0.628	1.320	5.512***	tertiary vs. p	rimary, p	p = 0.023
Control conditions	20	0.802	0.547	1.056	$6.184^{***}$	1.210	1	0.271
1. paper-and-pencil	10	0.645	0.269	1.022	3.357**			
2. computer-based	10	0.932	0.588	1.276	5.309***			
Intervention settings	20	0.793	0.532	1.054	5.961***	0.051	1	0.821
1. classroom	11	0.821	0.469	1.172	$4.576^{***}$			
2. outdoor	9	0.760	0.371	1.149	3.826***			
Software types	20	0.724	0.513	0.935	6.736***	0.599	1	0.439
1. general purposes	8	0.666	0.408	0.923	$5.072^{***}$			
2. educational purposes	12	0.843	0.476	1.210	$4.500^{***}$			
Measured outcome types	20	0.765	0.534	0.996	$6.484^{***}$	0.056	1	0.813
1.standardized	11	0.805	0.401	1.208	3.907***			
2.researcher-designed	9	0.745	0.463	1.027	$5.180^{***}$			
Intervention durations	20	0.771	0.517	1.025	$5.952^{***}$	0.341	2	0.843
1. short	2	1.017	-0.158	2.193	1.696			
2. intermediate	6	0.709	0.033	0.355	3.921***			
3. long	12	0.817	0.195	0.038	4.185***			

*Note.* k = number of effect sizes; g = Hedges' g; LL = lower limit, UL = upper limit; 95 % CI = 95% confidence interval; to further locate the sources of variation, post-hoc comparisons were reported for the significant heterogeneity; \*p < .05; \*\*p < .01; \*\*\*p < .001.

*Educational level.* As shown in Table 3, the effect size was largest for college students of tertiary education (g = 0.974, p < .001), followed by the moderate effect size of primary (g = 0.485, p < .001) and secondary (g = 0.415, p < .050) educational levels. Between-group comparisons reached a statistical significance (Q = 6.386, p = .041), indicating that the effect sizes of three categories differ significantly from each other. Post-hoc comparisons were further executed. College students of tertiary education who used MALL for EFL listening skill development were found to have higher beneficial effects than those of primary (Q = 5.188, p = .023) and secondary (Q = 6.386, p = .028) educational levels. However, no significant difference was obtained between those of the primary and secondary educational levels (Q = 0.103, p = .748).

*Control conditions.* Regarding the moderator effect of control conditions, both paper-and-pencil (g = 0.645, p < .001) and computer-based (g = 0.932, p < .001) conditions were found to be significant. The moderator effect of control conditions did not reach a significant level (Q = 1.210, p = .271).

*Intervention settings*. Intervention settings consist of instructional activities occurred in the classroom (classroom setting) and outside of the classroom (outdoor setting). The classroom setting (g = 0.821, p < .001) had a high effect size, while the outdoor setting (g = 0.760, p < .001) had a moderate effect size. The moderator effect of intervention settings did not reach a significant level (Q = 0.051, p = .821).

Software types. According to Rosell-Aguilar (2017), MALL software could be taxonomized into two categories: educational and general purposes. The MALL software for educational purposes achieved a significantly high effect size (g = 0.843, p < .001), and general purposes had a moderate effect size (g = 0.666, p < .001). No significant between-group difference was found between them (Q = 0.599, p = .439).

*Measured outcome types.* The standardized tests achieved a large effect size (g = 0.805, p < .001), and researcher-designed tests had a moderate-to-large effect size (g = 0.745, p < .001). The effect size did not vary based on whether the standardized or researcher-designed tests were used (Q = 0.056, p = .813), indicating the robustness of research outcomes between studies.

*Intervention durations*. When integrated with MALL for EFL listening skill development, long-term (> 4 weeks,  $\leq$  one semester) durations had a high effect size, and intermediate-term (> 1 week,  $\leq$  4 weeks) durations had a moderate effect size (Table 3). However, the significant effect size of short-term (one session,  $\leq$  1 week) durations was not found, Z = 1.696, 95% CI [-0.158, 2.193], p = .090. Between-group comparisons did not achieve a statistical significance level (Q = 0.341, p = .843), suggesting the effect size did not significantly differ among the three categories.

# 5. Discussion

This study represents a meta-analytic approach to the effectiveness of MALL for EFL listening skill development over the past decade. Through the meta-analysis, we found 20 experimental studies involving a total of 1218 EFL learners that met the proposed inclusion and exclusion criteria. Results of the meta-analysis showed that the overall effect size was moderate-to-large, suggesting that MALL for EFL listening skill development is more effective than traditional methods. In addition to the overall effect of MALL for EFL listening skill development, this study also reported the different moderator effects of educational levels, software types, control conditions, intervention settings, measured outcome types and intervention durations. Research findings were discussed to answer the two research questions in the remainder of this section accordingly.

Regarding research question 1, the result showed that EFL learners benefit more from MALL for their listening skill development than from traditional methods, and the effect size of 0.792 could be comparable to several recently published meta-analyses of MALL for language learning (g = 0.720, Burston & Giannakou, 2022; g =0.722, Chen et al., 2021). After a scrutiny of those studies included, three possible reasons could be tentatively concluded to explicate the pedagogical benefits of MALL for EFL listening skill development. First, the mobility feature of MALL in its own right supports ubiquitous and autonomous learning. In other words, EFL learners can use MALL to autonomously practice listening comprehension from anywhere and at any time (Li, 2022b), which will increase the input exposure rates of learning materials compared to the highly temporal and spatialconstrained traditional methods, e.g., indoor computer classroom (Chang et al., 2018) and conventional paper and pencil (Azar & Nasiri, 2014). The increased input exposure rates of MALL also lend support from the old saying-practice makes perfect. Second, the multimodal materials of MALL reduce working memory loads and facilitate listening comprehension process. According to the Dual Coding Theory (Paivio, 1991), there are verbal and non-verbal channels of working memory that process information independently from one another. When the two channels are interconnected with each other, working memory loads will be decreased and the learning outcomes will be improved (Li, 2021a). Compared to the unimodal presentation of the traditional methods, EFL learners who used MALL can make full use of multimodal materials to practice listening comprehension (Mayer, 2009). For instance, EFL learners' listening skill was greatly improved with a mobile VR due to the multimodal virtual presence and the high degree of immersion (Tai & Chen, 2021). Third, the interactivity of MALL affords enjoyable listening experiences, which increases EFL learners' flow experiences (Li et al., 2021), motivation and engagement (Li, 2021a; Li et al., 2019), and self-efficacy (Li, 2021b). For instance, within a mobile VR environment, the interaction between realistic environment and spatial audio allows learners to feel "being there" and "being participants", which triggers their flow experiences, motivation and engagement, and facilitates listening skill development in turn (Tai & Chen, 2021).

Research question 2 dealt with the moderator effects of educational levels, software types, control conditions, intervention settings, measured outcome types and intervention durations. Moderator analysis results are discussed as follows.

Educational levels. While pedagogical benefits of MALL for EFL listening skill development are obtained among EFL learners of all educational levels, larger effect regarding college students of tertiary education over

those of primary and second education may be attributed to the restricted use of mobile devices for children of primary and secondary education, rather than adults of tertiary education (Wiederhold, 2019). Another possible explanation might lie in researchers' inadequate attention to participants of primary and secondary education, warranting further research in this regard. A closer look at the unbalanced distribution of educational levels indicates that, among the 20 eligible experimental studies, only 35% of the participants (k = 7 out of 20) who adopted MALL for EFL listening skill development are children of primary and secondary education. By contrast, 65% of the participants (k = 13) are college students.

*Control conditions*. While both paper-and-pencil and computer-based conditions have significant effect sizes, the moderator effect of control conditions did not reach a significant level, suggesting that MALL for listening skill development is robustly effective, regardless of the difference in control conditions. To elucidate the reasons, it will come as no surprise to find that, although some researchers (e.g., Alabsi, 2020; Kargozari & Tafazoli, 2012; Shiri, 2015) designed the control group with the paper-and-pencil method, while others (e.g., Rashtchi & Mazraehno, 2019 using DVDs and audio CDs; Zhang, 2016 using traditional CD Room; Rahimi & Soleymani, 2015 using desktop computers) with the computer-based method, all the studies included are under strict (quasi-)experimental design, warranting the reliability and robustness of the results reported.

*Intervention settings*. While both the informal outdoor and the formal classroom settings are beneficial, the pedagogical benefit of classroom setting is higher than that of outdoor setting, suggesting that EFL learners who used MALL for formal classroom listening practice would outperform those did informally. This result is, however, inconsistent with domain-general meta-analyses on MALL (Chen et al., 2020; Sung et al., 2015; Sung et al., 2016), advocating larger effect size in informal outdoor setting than in classroom setting. A plausible explanation for the discrepancy might rest on the different domains of investigation: domain-general vs. domain-specific. In other words, different from those domain-general meta-analyses of MALL research, this study meta-analyzed the domain-specific MALL for EFL listening skill development that requires formal and intensive classroom instruction to warrant its pedagogical benefit (Sung et al., 2015).

*Software types.* The higher effect of MALL devices for educational purposes than for general purposes corroborates the result of Sung and colleagues (2016), who posited that compared to MALL devices for general purposes, MALL devices for educational purposes integrated functionalities well with the content of curriculums (Li, 2022b). Thus, MALL devices for educational purposes, e.g., mobile English listening systems (Liu et al., 2018) and subtitles for English listening (Alabsi, 2020), are better tailored to EFL learners' personalized needs than those of general purposes (Chen et al., 2020).

*Measured outcome types.* The effect size did not vary between the standardized or researcher-designed tests, which is consistent with previous finding that compared the moderator effect between standardized and researcher-designed tests in a meta-analysis of writing devices (Xu et al., 2019), suggesting both standardized tests and researcher-designed tests could warrant a robust reliability to measure EFL learners' listening comprehension performance.

Intervention durations. Although there is no significant between-group difference of short-term, intermediateterm and long-term durations, long-term durations achieved a high effect size of 0.817, showing a consistent result compared to researchers who advocating that "long-term teaching interventions are important for obtaining reliable results" (Sung et al., 2016, p. 264). Moreover, unlike other skills (e.g., vocabulary learning) that could be improved over a short period of time, EFL listening skill development requires long-term practice and training (Sung et al., 2015). It is promising to observe that 60% of the studies (k = 12 out of 20) involved listening tasks are carried out for long-term (> 4 weeks,  $\leq$  one semester) durations, while only 10% of the studies (k = 2) are for short-term (one session,  $\leq 1$  week) durations, suggesting that appropriate long-term intervention durations and opportunities for mobile listening practice are most favored. This may explain the reason why the beneficial effects of MALL for EFL listening skill development for short-term is not significant.

# 6. Implications

With regard to the aforementioned findings, the following practical implications could also be inferred for instructors, designers and researchers.

### **6.1. Implications for instructors**

First, since MALL for EFL listening skill development has been confirmed to be more effective than traditional methods, its pedagogical potentials should be encouraged to explore among learners of different educational levels, especially among children of primary and secondary educational levels who are often banned from mobile devices use (Wiederhold, 2019). Second, while robust effects of MALL for listening skill development could be obtained regardless of the difference in control conditions, an experimental-control group comparison regarding "MALL vs. computer-assisted" design achieves higher effect size ( $g_{computer-based}$  vs.  $g_{paper-and-pencil} = 0.932$  vs. 0.645) than that of "MALL vs. paper-and-pencil" design, suggesting that instructors should adopt the former design in their EFL listening instruction to achieve better pedagogical effectiveness. Third, given the higher effect of MALL for educational purposes than for general purposes, instructors should help EFL learners select and tailor appropriate MALL devices and prepare useful multimodal listening materials based on their personalized needs.

#### 6.2. Implications for designers

First, MALL technology designers should consider EFL learners' educational levels. That is to say, the difficulty of listening tasks could be dynamically tailored with reference to the ongoing detection of educational levels. Second, designers should collect instructors and learners' opinions regarding the use of MALL applications to adjust the different intervention settings. Third, to strengthen the educational purposes, designers should try to integrate curriculums into the MALL applications depending on the educational needs and purposes.

#### **6.3. Implications for researchers**

First, as scant attention has been paid on participants of primary and secondary educational levels, researchers should focus on how the use of MALL can facilitate the listening skill development among learners of primary and secondary education. Second, given that longer durations achieve higher effect sizes regarding the use of MALL for EFL listening skill development, researchers should design the research with long-term durations to increase the reliability of findings (Sung et al., 2016). Third and importantly, while this study sheds light on the effectiveness of MALL for EFL learners' listening skill development, there is a desperate need for researchers to focus on MALL for other foreign language listening comprehension, e.g., Chinese as a foreign language (CFL).

# 7. Conclusion

To gain a more comprehensive understanding into the effects of MALL for EFL learners' listening skill development, this study conducted a meta-analysis of 20 selected experimental studies involving 1218 EFL learners. Results in response to the two main research questions were concluded: First, the overall effect size was moderate-to-large, suggesting the use of MALL for EFL listening skill development is more effective than traditional methods. Second, moderator analysis results indicated that educational levels were found to be a significant moderator, while no significant moderator effects of software types, control conditions, intervention settings, measured outcome types and intervention durations were obtained.

While results of this study may contribute to advancing our understanding regarding the effects of MALL for EFL listening skill development and the potential moderators that may affect such effects, there are some limitations with the study: First, given the limited studies of MALL for other FL listening skill development, it is rather premature to consider other foreign languages in the meta-analysis this time. As such, we only considered EFL learners who used MALL for listening skill development. Future research should not only consider the moderator effect of different foreign languages, but also meta-analyze the effectiveness of MALL for a particular FL listening comprehension. Second, due to the limited keyword combinations and strict inclusion/exclusion criteria, only 20 experimental studies that met the criteria were analyzed. Future study may consider the inclusion of more publications on MALL for EFL listening skill development. Third, due to small number of experimental studies with affective outcomes that are premature for meta-analysis, this study only involved standardized or researcher-designed test scores for listening outcome. Future study should take EFL learners' affective outcomes into account when the number of primary studies is sufficient for analysis. Lastly, drawing on existing meta-analyses for MALL, we only reported the moderator effects of educational levels, software types, control conditions, intervention settings, measured outcome types and intervention durations for the overall effect size. Future attempt should consider other potential moderators with sufficient information for calculation.

# Acknowledgement

This research was supported by the Social Science Foundation of Hunan Province (grant number 20ZDB005). The author would like to thank the anonymous reviewers and Prof. Yu-Ju Lan for their constructive and insightful comments.

## References

Articles marked with an asterisk (\*) were analysed in the meta-analysis.

\*Alabsi, T. (2020). Effects of adding subtitles to video via Apps on developing EFL students' listening comprehension. *Theory and Practice in Language Studies*, *10*(10), 1191–1199. https://doi.org/10.17507/tpls.1010.02

\*Al-Shamsi, A., Al-Mekhlafi, A. M., Busaidi, S. A., & Hilal, M. M. (2020). The Effects of mobile learning on listening comprehension skills and attitudes of Omani EFL adult learners. *International Journal of Learning, Teaching and Educational Research*, *19*(8), 16–39. https://doi.org/10.26803/IJLTER.19.8.2

\*Andujar, A., & Hussein, S. A. (2019). Mobile-mediated communication and students' listening skills: A Case study. *International Journal of Mobile Learning and Organisation*, *13*(3), 309–332. https://doi.org/10.1504/IJMLO.2019.100432

<sup>\*</sup>Azar, A. S., & Nasiri, H. (2014). Learners' attitudes toward the effectiveness of mobile assisted language learning (MALL) in L2 listening comprehension. *Procedia - Social and Behavioral Sciences*, *98*, 1836–1843. https://doi.org/10.1016/j.sbspro.2014.03.613

Biernacki, P., & Waldorf, D. (1981). Snowball sampling: Problems and techniques of chain referral sampling. *Sociological Methods & Research*, *10*(2), 141–163. https://doi.org/10.1177/004912418101000205

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2005). *Comprehensive meta-analysis version 2.0* [Computer software]. Biostat.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). Introduction to meta-analysis. John Wiley & Sons, Ltd.

Burston, J., & Giannakou, K. (2022). MALL language learning outcomes: A Comprehensive meta-analysis 1994–2019. *ReCALL*, 34(2), 147 - 168. https://doi.org/10.1017/S0958344021000240

<sup>\*</sup>Chang, C., Warden, C. A., Liang, C., & Chou, P. (2018). Performance, cognitive load, and behaviour of technology-assisted English listening learning: From CALL to MALL. *Journal of Computer Assisted Learning*, *34*(2), 105–114. https://doi.org/10.1111/jcal.12218

Chen, Z., Chen, W., Jia, J., & An, H. (2020). The Effects of using mobile devices on language learning: A Meta-analysis. *Educational Technology Research and Development*, 68, 1769–1789. https://doi.org/10.1007/s11423-020-09801-5

Cook, D. A., & Reed, D. A. (2015). Appraising the quality of medical education research methods: The Medical education research study quality instrument and the Newcastle-Ottawa scale-education. *Academic Medicine*, 90(8), 1067–1076. https://doi.org/10.1097/ACM.00000000000786

Coşkun, A., & Marlowe, Z. (2020). The Place of technology-assisted language learning in EFL listening: A Review of literature and useful applications. In M. Durnali & İ. Limon (Eds.), *Enriching Teaching and Learning Environments with Contemporary Technologies* (pp. 102–116). IGI Global.

Costabile, M. F., De Angeli, A., Roselli, T., Lanzilotti, R. & Plantamura, P. (2003). Evaluating the educational impact of a tutoring hypermedia for children. *Information Technology in Childhood Education Annual*, *1*, 289–308. https://www.learntechlib.org/primary/p/18867/

<sup>\*</sup>de la Fuente, M. J. (2014). Learners' attention to input during focus on form listening tasks: The Role of mobile technology in the second language classroom. *Computer Assisted Language Learning*, 27(3), 261–276. https://doi.org/10.1080/09588221.2012.733710

Holden, C. L., & Sykes, J. M. (2011). Leveraging mobile games for place-based language learning. In P. Felicia (Ed.), *Developments in Current Game-Based Learning Design and Deployment* (pp. 27-45). IGI Global.

<sup>\*</sup>Hsu, C., Hwang, G., Chang, Y., & Chang, C. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society*, *16*(1), 403–414.

Hwang, G., & Fu, Q. (2019). Trends in the research design and application of mobile language learning: A Review of 2007-2016 publications in selected SSCI journals. *Interactive Learning Environments*, 27(4), 567–581. https://doi.org/10.1016/j.compedu.2018.01.004

<sup>\*</sup>Hwang, W., Huang, Y., Shadiev, R., Wu, S., & Chen, S. (2014). Effects of using mobile devices on English listening diversity and speaking for EFL elementary students. *Australasian Journal of Educational Technology*, *30*(5), 503–516. https://doi.org/10.14742/ajet.237

<sup>\*</sup>Hwang, W., Shih, T. K., Ma, Z., Shadiev, R., & Chen, S. (2016). Evaluating listening and speaking skills in a mobile gamebased learning environment with situational contexts. *Computer Assisted Language Learning*, 29(4), 639–657. https://doi.org/10.1080/09588221.2015.1016438

Jensen, L., & Konradsen, F. (2018). A Review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, 23(4), 1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Jia, C., & Hew, K. F. (2021). Toward a set of design principles for decoding training: A Systematic review of studies of English as a foreign/second language listening education. *Educational Research Review*, 33, 1–18. https://doi.org/10.1016/j.edurev.2021.100392

<sup>\*</sup>Kargozari, H. R., & Tafazoli, D. (2012). Podcasting: A Supporting tool for listening. In *Proceedings of International Technology, Education and Development (INTED2012) Conference* (pp. 3870–3873). International Association of Technology, Education and Development (IATED).

\*Kim, H. (2013). Emerging mobile apps to improve English listening skills. *Multimedia-Assisted Language Learning*, *16*(2), 11–30. https://doi.org/10.15702/mall.2013.16.2.11

Lai, Y., Saab, N., & Admiraal, W. (2022). Learning strategies in self-directed language learning using mobile technology in higher education: A Systematic scoping review. *Education and Information Technology*, 27, 7749–7780. https://doi.org/10.1007/s10639-022-10945-5

Li, R. (2021a). Does game-based vocabulary learning app influence Chinese EFL learners' vocabulary achievement, motivation and self-confidence? *SAGE Open*, *11*(1), 1–12. https://doi.org/10.1177/21582440211003092

Li, R. (2021b). Modeling the continuance intention to use automated writing evaluation. SAGE Open, 11(4), 1–13. https://doi.org/10.1177/21582440211060782

Li, R. (2022a). Effects of mobile assisted language learning on EFL/ESL reading comprehension. *Educational Technology & Society*, 25(3), 15–29.

Li, R. (2022b). Research trends of blended language learning: A Bibliometric synthesis of SSCI-indexed journal articles during 2000–2019. *ReCALL*, 34(3), 309–326. https://doi.org/10.1017/S0958344021000343

Li, R. (2022c). Foreign language reading anxiety and its correlates: A Meta-analysis. *Reading and Writing*, *35*(4), 995–1018. https://doi.org/10.1007/s11145-021-10213-x

Li, R. (2022d). Effects of blended language learning on EFL learners' language performance: An Activity theory approach. *Journal of Computer Assisted Learning*, *38*, 1273–1285. https://doi.org/10.1111/jcal.12697

Li, R. (in press). Effects of mobile-assisted language learning on foreign language learners' speaking skill development. Language Learning & Technology.

Li, R., Meng, Z., Tian, M., Zhang, Z., & Xiao, W. (2021). Modelling Chinese EFL learners' flow experiences in digital game-based vocabulary learning: The Roles of learner and contextual factors. *Computer Assisted Language Learning*, *34*(4), 483–505. https://doi.org/10.1080/09588221.2019.1619585

Li, R., Meng, Z., Tian, M., Zhang, Z., Ni, C., & Xiao, W. (2019). Examining EFL learners' individual antecedents on the adoption of automated writing evaluation in China. *Computer Assisted Language Learning*, 32(7), 784–804. https://doi.org/10.1080/09588221.2018.1540433

Lipsey, M. W., & Wilson, D. B. (2001). Practical meta-analysis. SAGE.

<sup>\*</sup>Liu, G., Chen, J., & Hwang, G. (2018). Mobile-based collaborative learning in the fitness center: A Case study on the development of English listening comprehension with a context-aware application. *British Journal of Educational Technology*, 49(2), 305–320. https://doi.org/10.1111/bjet.12581

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

Mendoza, V., la Torre, A. V., & Páez-Quinde, C. (2020). Mobile Apps (English listening and 6 Minutes English) and the listening skill. In M. E. Auer, H. Hortsch & P. Sethakul (Eds.), *The Impact of the 4th Industrial Revolution on Engineering Education* (pp. 60–66). Springer.

<sup>\*</sup>O, K.-M. (2015). The Effectiveness of mobile assisted language learning on L2 listening comprehension. *Multimedia-Assisted Language Learning*, *18*(2), 135–158. https://doi.org/10.15702/mall.2015.18.2.135

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Journal of Psychology*, 45(3), 255–287. https://doi.org/10.1037/h0084295

Plonsky, L., & Oswald, F. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64, 878–912. https://doi.org/10.1111/lang.12079

\*Rahimi, M., & Soleymani, E. (2015). The Impact of mobile learning on listening anxiety and listening comprehension. *English Language Teaching*, 8(10), 152–161. https://doi.org/10.5539/elt.v8n10p152

\*Rashtchi, M., & Mazraehno, M. R. T. (2019). Exploring Iranian EFL learners' listening skills via TED talks: Does medium make a difference? *Journal of Language and Education*, 5(4), 81–97. https://doi.org/10.17323/jle.2019.9691

Reed, D. A., Cook, D. A., Beckman, T. J., Levine, R. B., Kern, D. E., & Wright, S. M. (2007). Association between funding and quality of published medical education research. *JAMA*, 298(9), 1002–1009. https://doi.org/10.1001/jama.298.9.1002

Richards, J. C. (2005). Second thoughts on teaching listening. *RELC Journal*, 36(1), 85–92. https://doi.org/10.1001/jama.298.9.1002

Rosell-Aguilar, F. (2017). State of the App: A Taxonomy and framework for evaluating language learning mobile applications. *CALICO Journal*, *34*(2), 243–258. https://doi.org/10.1558/cj.27623

Rosenthal, R. (1991). Meta-analytic procedures for social research. Sage.

Saeedakhtar, A., Haqiu, R., & Rouhi, A. (2021). The Impact of collaborative listening to podcasts on high school learners' listening comprehension and vocabulary learning. *System*, *101*, 102588. https://doi.org/10.1016/j.system.2021.102588

\*Seo, W., & Choi, I. (2014). The Effect of using a smart-phone speaking application on Korean middle school students' English expression learning and satisfaction. *Multimedia-Assisted Language Learning*, *17*(1), 34–57. https://doi.org/10.15702/mall.2014.17.1.34

\*Shiri, S. (2015). The Application of podcasting as a motivational strategy to Iranian EFL learners of English: A View toward listening comprehension. *Advances in Language and Literary Studies*, *6*(3), 155–165. https://doi.org/10.7575/aiac.alls.v.6n.3p.155

Siegel, J. (2014). Exploring L2 listening instruction: Examinations of practice. *ELT Journal*, 68(1), 22–30. https://doi.org/10.1093/elt/cct058

Sung, Y., Chang, K., & Liu, T. (2016). The Effects of integrating mobile devices with teaching and learning on students' learning performance: A Meta-analysis and research synthesis. *Computers & Education*, *94*, 252–275. https://doi.org/10.1016/j.compedu.2015.11.008

Sung, Y., Chang, K., & Yang, J. (2015). How effective are mobile devices for language learning? A Meta-analysis. *Educational Research Review*, *16*, 68–84. https://doi.org/10.1016/j.edurev.2015.09.001

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A Second-order meta-analysis and validation study. *Review of Educational Research*, *81*, 4–28. https://doi.org/10.3102/0034654310393361

\*Tai, T., & Chen, H. H. (2021). The Impact of immersive virtual reality on EFL learners' listening comprehension. *Journal of Educational Computing Research*, 59(7), 1272–1293. https://doi.org/10.1177/0735633121994291

Thorne, S. L., Hellermann, J., & Jakonen, T. (2021). Rewilding language education: Emergent assemblages and entangled actions. *The Modern Language Journal*, 105(S1), 106–125. https://doi.org/10.1111/modl.12687

Valentine, J. C. (2019). Incorporating judgments about study quality into research syntheses. In H. Cooper, L. V., Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-analysis* (3rd ed., pp. 129–140). Russell Sage Foundation.

Vandergrift, L., & Goh, C. C. M. (2012). Teaching and learning second language listening: Metacognition in action. Routledge.

Wiederhold, B. K. (2019). Should smartphone use be banned for children? *Cyberpsychology, Behavior, and Social Networking*, 22(4), 235–236. https://doi.org/10.1089/cyber.2019.29146.bkw

Xu, Z., Banerjee, M., Ramirez, G., Zhu, G., & Wijekumar, K. (2019). The Effectiveness of educational technology applications on adult English language learners' writing quality: A Meta-analysis. *Computer Assisted Language Learning*, 32(1-2), 132–162. https://doi.org/10.1080/09588221.2018.1501069

<sup>\*</sup>Zhang, Y. (2016, December). *The Impact of mobile learning on ESL listening comprehension* [Paper presentation]. 3rd International Conference on Advanced Education and Management (ICAEM 2016). https://doi.org/10.12783/dtssehs/icaem2016/4290

Zheng, D., Dai, Y., & Liu, Y., (2017). Place, identity and becomings with mobile technologies: Feminist and Chinese philosophies. *International Communication of Chinese Culture*, 4(3), 319–337. https://doi.org/10.1007/s40636-017-0106-1

Zheng, D., Liu, Y., Lambert, A., Lu, A., Tomei, J., & Holden, D. (2018). An Ecological community becoming: Language learning as first-order experiencing with place and mobile technologies. *Linguistics and Education*, 44, 45–57. https://doi.org/10.1016/j.linged.2017.10.004
# The Interplay Between Cognitive Load and Self-Regulated Learning in a Technology-Rich Learning Environment

# Tingting Wang<sup>1\*</sup>, Shan Li<sup>2</sup> and Susanne Lajoie<sup>1</sup>

<sup>1</sup>Department of Educational and Counselling Psychology, McGill University, Canada // <sup>2</sup>Lehigh University, United States // tingting.wang4@mail.mcgill.ca // shla22@lehigh.edu // susanne.lajoie@mcgill.ca \*Corresponding author

(Submitted January 19, 2022; Revised July 9, 2022; Accepted August 12, 2022)

ABSTRACT: Cognitive load can be induced by both learning tasks and self-regulated learning (SRL) activities, which compete for limited working memory capacity. However, there is little research on the relationship between cognitive load and SRL. This study explored how cognitive load interplayed with SRL behaviors and their joint effects on task performance (i.e., diagnostic efficiency) in the context of clinical reasoning. Specifically, twenty-seven (N = 27) medical students diagnosed three virtual patient cases in BioWorld, a simulation-based learning environment to improve medical students' clinical reasoning skills. Students' SRL behaviors were automatically recorded in BioWorld log files as they accomplished the tasks. We employed text mining techniques to extract four linguistic features from students' concurrent think-aloud, i.e., cognitive discrepancy, insight, causation, and positive emotions, which were further used to represent students' cognitive load. The latent profile analysis was then performed to cluster students into high- and low-load group. We also conducted a path analysis to investigate the mediation roles of SRL behaviors in the relationship between cognitive load and diagnostic efficiency (task performance). The results revealed that cognitive load negatively affected diagnostic efficiency, mediated by the ratio of SRL behaviors in the self-reflection phase. This study provides theoretical and methodological insights regarding the measurement of cognitive load and its interplay with SRL. This study informs the design of effective interventions for managing cognitive load in SRL within intelligent tutoring systems.

Keywords: Cognitive load, Self-regulated learning, Technology-rich learning environment, Text mining

# 1. Introduction

Technology-rich learning environments (TREs), such as multimedia, simulation, virtual reality, and intelligent tutoring systems, have been increasingly employed to foster medical students' clinical reasoning skills (Azevedo & Gašević, 2019). Clinical reasoning refers to a complex reasoning and decision-making process whereby health professionals get familiar with patient information, collect evidence, propose hypotheses, evaluate gathered evidence, and make final diagnostic decisions (Kuiper, 2013; Simmons, 2010). Due to the crucial role of clinical reasoning on patients' health, medical students need to plan, monitor, and control their problem-solving processes to achieve an accurate diagnosis, which is also referred to as self-regulated learning (SRL) (Artino et al., 2014; Brydges & Bulter, 2012; Cleary et al., 2016). SRL is a recursive process by which learners monitor and control their motivational, behavioral, emotional, and cognitive aspects to realize pre-determined goals (Greene & Azevedo, 2007; Panadero, 2017; Pintrich, 2000; Winne & Perry, 2000; Zimmerman, 2000). Researchers have attempted to examine medical students' clinical reasoning process from the perspective of SRL and demonstrated that strategical planning and reflective journal writing significantly promoted their performance in clinical reasoning tasks (Artino et al., 2014; Kuiper et al., 2014; Kuiper et al., 2009).

In addition to SRL, cognitive load is also an explanatory theoretical lens to understand clinical reasoning outcomes (Solhjoo et al., 2019). Cognitive load refers to the amount of working memory capacity (WMC) occupied by solving a specific task (Paas et al., 2003). Decades of research have shown that mental overload induces more negative emotions and leads to poorer academic performance across disciplines (Leutner et al., 2009; Scheiter et al., 2020). Given the complexity of clinical reasoning, including the intricate patient information, uncertainty about the diagnostic decisions, and the detrimental consequences of medical errors, medical students are likely to experience a high cognitive load during the diagnostic process (Durning et al., 2011; Solhjoo et al., 2019). As for the effects of cognitive load on clinical reasoning performance, empirical studies exhibited mixed results. For instance, Solhjoo et al. (2019) demonstrated a negative association between self-reported cognitive load and diagnostic performance, whereas Durning et al. (2011) and Fraser et al. (2012) indicated a positive relationship.

However, the research on cognitive load and SRL is conducted separately, and few studies have investigated the interplay between cognitive load and SRL. Therefore, the current research examines the interplay between cognitive load and SRL and their joint roles in explaining clinical reasoning performance. Specifically, we situated this study in the BioWorld system (Lajoie, 2009), a technology-rich learning environment which simulates virtual patient cases for medical students to improve their clinical reasoning skills. BioWorld keeps track of students' operations in log files, which are necessary to analyze fine-granular SRL behaviors. The following section provides the theoretical foundation and research questions.

# 2. Theoretical framework

### 2.1. Cognitive load theory

Cognitive load theory (CLT) is based on a cognitive architecture consisting of working memory and long-term memory (Paas et al., 2003; Sweller, 2011). The long-term memory system provides an infinite capacity to store acquired knowledge in cognitive schemas, a complex unit of interrelated information elements (Bower et al., 1975). Working memory temporarily stores and manipulates novel information (Baddeley, 1992), working as a conduit between external environments and the long-term memory system (Kirschner, 2002). Compared with long-term memory, the working memory system is limited in capacity and duration when dealing with new information. The finite WMC is necessary for mental tasks such as language comprehension, problem-solving, and planning (Cowan, 2011; Wiley & Jarosz, 2012).

Cognitive load refers to the load that performing a specific task exerts on the working memory (Paas et al., 1994; Sweller, 2011). CLT distinguishes three types of cognitive load: intrinsic, extraneous, and germane. Intrinsic load reflects the "necessary load" determined by task complexity and expertise levels. Specifically, the intrinsic load increases with the element interactivity of tasks and decreases with learners' expertise levels (Paas et al., 2010; Park et al., 2015; Sweller, 1988; Sweller, 2011). Charlin et al. (2007) demonstrated that medical experts tended to experience a lower intrinsic load during diagnostic tasks than novices by applying knowledge constructed from prior experiences. However, the extraneous load is ineffective since it is triggered by inappropriate instructional designs and display modes (Paas et al., 1994). For instance, Reedy (2015) indicated that suboptimal designs of simulation-based learning environments, such as redundant information and inappropriate presentation format, were associated with increased extraneous load in clinical reasoning. Lastly, germane load is induced by schema construction and automation and represents a kind of effective load that directly contributes to learning (Sweller, 1988). Moreover, intrinsic, extraneous, and germane load are additive, and the sum is referred to as the overall load (Paas et al., 2003). This study is particularly interested in the overall cognitive load students experienced in the clinical reasoning process.

### 2.2. Linguistic features of cognitive load

There are a variety of methods to measure cognitive load in TREs. Subjective questionnaires have been intensively used to measure cognitive load (Leppink et al., 2013; Paas, 1992). Physiological techniques, including eye-tracking (Joseph & Murugesh, 2020), electroencephalograph (EEG) (Antonenko et al., 2010), and heart rate variability (HRV) (Solhjoo et al., 2019), can objectively trace cognitive load changes. For the performance-based measures, study time, accuracy, and error rate are frequently used to represent students' cognitive load (Brünken et al., 2003; Paas et al., 2003; Sweller, 1988).

Words and language use can also reflect individuals' psychological processes (Darabi et al., 2010), and the advancement of text mining techniques makes linguistic features a promising measurement of cognitive load. In particular, this study employed the text mining program Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015), an automatic tool to count word frequency to extract participants' linguistic features. Linguistic features are more reliable indicators of cognitive load than other techniques since language is the most common and direct way to reflect internal mental states (Pennebaker et al., 2003; Pennebaker et al., 2015). The commonly used linguistic features include speech pause (Khawaja et al., 2014; Müller et al., 2001), use of first/third-person plurals (Sexton & Helmreich, 2000), patterns of personal pronoun (Khawaja et al., 2012), the use of different word categories, repetitive words, and grammatical structures. Researchers have applied linguistic features to infer students' cognitive load in learning contexts. For instance, Konopasky et al. (2020) found that participants used fewer first-person pronouns in high-load contexts because high cognitive load would induce more cognitive processes and thus less attention on the self. Khawaja et al. (2014) revealed that students were more likely to use cognitive words (e.g., think, know, and consider) when they experienced a high

cognitive load. Positive emotions words were also confirmed as a negative index of experienced cognitive load (Fraser et al., 2012).

### 2.3. Self-regulated learning in the context of clinical reasoning

According to Zimmerman (2000), SRL refers to self-generated thoughts, feelings, and behaviors that are monitored and adjusted in three loosely sequential phases to attain learning goals. During the forethought phase, learners prepare their motivation and self-belief, conduct task interpretation, and set goals and strategic plans. In the performance phase, learners take appropriate learning strategies to execute the task and monitor task progress, and they evaluate and reflect on performance during the self-reflection phase.

As aforementioned, the clinical reasoning task requires students to collect patient information, order lab tests, integrate and evaluate evidence, propose and reflect on diagnostic hypotheses, and make final decisions, which process has substantial overlaps with SRL (Artino et al., 2014; Kuiper, 2009; Kuiper, 2013; Simmons, 2010). From the perspective of SRL, medical students review patient information to familiarize themselves with task conditions and plan strategies and efforts to solve the task during the forethought phase (Li et al., 2020). Throughout the performance phase, medical students conduct lab tests, search for external resources, and integrate all gathered evidence to propose diagnostic hypotheses (Artino et al., 2014). During the last self-reflection phase, medical students evaluate and reflect on their diagnostic judgments to see if any additional actions are required (Brydges & Butler, 2012). The emerging literature within medical education showed that medical students did not often exhibit strategic thinking and self-evaluative judgments during clinical reasoning (Cleary et al., 2016). Moreover, several empirical studies have demonstrated that insufficient SRL, such as poor planning, deficits in self-monitoring, and scant self-reflection, led to poor clinical reasoning performance (Artino et al., 2014; Brydges & Butler, 2012).

### 2.4. Self-regulated learning and cognitive load

SRL can also be regarded as a set of information processes from the lens of information processing (Panadero, 2017; Winne, 2001; Winne, 2018). That is, learners process task-specific information (e.g., patient symptoms) and psychological information (motivational and emotional states) in the forethought phase and apply tactics, strategies, and schemas to solve the problem during the performance phase (Winne, 2001). Finally, all the information collected in the previous two phases should be integrated and evaluated to reach a final decision in the self-reflection phase.

More importantly, information processes in different SRL phases also demand working memory resources and generate additional cognitive load beyond problem-solving itself (Seufert, 2018). In this regard, SRL behaviors and cognitive load are two contrary forces to competing for the limited WMC, and SRL can be regarded as a function of WMC and cognitive load (de Bruin et al., 2020; Seufert, 2018). As illustrated by Figure 1, adapted from Seufert (2018, 2020), students might be unable to self-regulate their learning when the cognitive load triggered by a task is high since little WMC remains for SRL. However, students can easily achieve learning goals when they experience a low-level cognitive load, even without efficient SRL processes. In other words, conducting SRL activities in less cognitive-demanding contexts is unnecessary. Only when the cognitive load is moderate will the resources sufficient for efficient SRL.

Despite the potential effects of cognitive load on SRL activities from the theoretical perspective (see Figure 1), few scholars provided empirical evidence to support these assumptions, especially in the context of clinical reasoning. The majority of studies explored the separate role of cognitive load and SRL in diagnostic performance (Artino et al., 2014; Durning et al., 2011; Kuiper, 2009; Solhjoo et al., 2019); however, the interactions between cognitive load and SRL and their joint effects on learning have been underexplored (de Bruin et al., 2020; Scheiter et al., 2020; Seufert, 2018; Seufert, 2020). Investigating the associations between cognitive load and SRL can facilitate medical students to achieve an accurate and efficient diagnosis, guaranteeing patients' safety. Therefore, this study addresses this issue by answering the following research questions: (1) Did medical students demonstrate different cognitive load patterns when diagnosing virtual patient cases in TREs? (2) Did cognitive load patterns affect SRL behaviors? and (3) How did cognitive load and SRL behaviors jointly predict diagnostic efficiency?



### 3. Method

### 3.1. Participants, procedures, and learning context

Twenty-seven (N = 27) participants were recruited from a large North American university. They consisted of 14 first- and 13 second-year medical students. Among them, seventeen (63%) were male students, and ten (37%) were female students, with a mean age of 23 (SD = 2.66). All students had completed a prerequisite course module on endocrinology, metabolism, and nutrition. Therefore, they mastered the necessary knowledge to complete the tasks.



Participants were first provided with a sample case to familiarize themselves with the BioWorld system (Lajoie, 2009), a simulation-based learning environment developed to help medical students practice clinical reasoning skills. During the formal experiments, each student was tasked to solve three virtual patient cases in BioWorld. The correct diagnoses for the three tasks are *Diabetes (Type1), Hyperthyroid (Grave's disease), and Pheochromocytoma*, respectively. Therefore, there were a total of 81 (27 x 3) different cases in this study. Participants were also instructed to concurrently speak out whatever comes to their minds without modifying their feelings and thoughts. The think-aloud protocols were audio-recorded and transcribed afterward for further

analysis. Each of the three cases lasted about 40-50 minutes, and the whole process took approximately 2-2.5 hours.

As shown in Figure 2, students initiated each diagnostic task by reading patient history and extracting critical symptoms. To collect additional information and evidence, students ordered appropriate lab tests and searched the online library. Students could propose one or more clinical hypotheses and manage hypotheses throughout the processes. They also linked gathered evidence with specific hypotheses and classified it into three categories, i.e., useful, neutral, and useless. In the end, students were also required to prioritize their hypotheses based on their subjective judgments. After submitting the final hypothesis, the BioWorld system provided individualized performance feedback for students.

### 3.2. Measures

### 3.2.1. Linguistic indicators of cognitive load

Cognitive load was inferred from transcribed think-aloud using the LIWC (Pennebaker et al., 2015). Based on an embedded dictionary that defines categories of word lists, LIWC automatically calculates the percentage of specific function words and provides a practical approach to detecting individuals' psychological processes from their linguistic patterns (Robinson et al., 2012). The current study extracted four markers, i.e., *positive emotion*, *cognitive discrepancy, insight*, and *causation*, from the LIWC output variables to represent cognitive load for each case. *Positive emotion* was a negative indicator of cognitive load and was calculated based on the percentage of words indicating pleasant perceptions. *Cognitive discrepancy* refers to the students' perceived inconsistency between prior knowledge and current task contexts. For instance, the words *should* and *would* were counted as cognitive discrepancies. *Cognitive insight* and *causation* indicated the intensity of cognitive efforts in solving the tasks. The three variables in the cognitive category were regarded as positive indicators of cognitive load indicators to infer students' cognitive load levels by the following latent profile analysis and calculated an individual's cognitive load value by the equation: *Cognitive load* = *Discrepancy* + *Insight* + *Causation* – *Positive Emotion*.

### 3.2.2. SRL behaviors

A total of 22 different operations were recorded in the BioWorld log files. Among them, ten operations were extracted as SRL behaviors (Table 1), and the remaining activities, such as *switch area* and *unlink evidence*, were excluded from our analysis. As shown in Table 1, this study conducted both a macro- and micro-level analysis of SRL. The forethought phase consisted of one SRL behavior, i.e., *Task Analysis*, whereby students review and interpret patient information. The performance phase of SRL included four behaviors: *Orientation*, *Execution*, *Help-Seeking*, and *Link Evidence*. Readers can find a detailed description of these behaviors in Table 1. In the self-reflection phase, students evaluated and reflected on the problem-solving process by the Evidence Evaluation and Hypothesis Evaluation behaviors. Moreover, we calculated the relative ratio of each SRL behavior to represent SRL behavior frequencies.

	<i>Table 1.</i> SRL behaviors extracted from log files						
Macro-level	Micro-level	Description	Sample				
Forethought	Task Analysis	Collect patient symptoms to construct an overall view of the problem	Add evidence				
Performance	Orientation	Propose or mange hypotheses based on the collected information and prior knowledge	Add hypothesis				
	Execution	Order lab tests to collect additional evidence	Add tests				
	Help-Seeking	Seek for external help from online library embedded in BioWorld	Search library Search library category				
	Link Evidence	Link evidence with specific hypotheses to claim progress	Link evidence				
Self-Reflection	Evidence Evaluation	Evaluate and classify the evidence into supportive group and against group	(Re)categorize				
	Hypothesis Evaluation	Evaluate possibilities of each hypothesis	(Re)prioritize				
37 . 751 11	1 1.1(						

Note. The coding scheme was adapted from Lajoie and Lu (2012).

### 3.2.3. Diagnostic efficiency

Diagnostic efficiency was automatically measured by the BioWorld system. Diagnostic efficiency refers to the matching degree between medical students' clinical reasoning process and experts' steps to reach the diagnosis. For instance, students would obtain a score of 50 if 50% of their clinical reasoning steps were matched with experts' solution steps (that were embedded in the BioWorld system). Diagnostic efficiency was designed to range from 0 to 100.

### 3.3. Data analysis

This study employed the latent profile analysis (LPA) to model different cognitive load patterns, *t*-tests to examine the effects of cognitive load patterns on SRL behaviors, and path analysis to explore the joint effects of cognitive load and SRL behaviors on diagnostic efficiency. The following session described how we used LPA and path analysis to address our research questions.

### 3.3.1. Latent profile analysis

Latent profile analysis, a person-centered mixture modelling method, can detect homogeneous clusters from observed variables through a probabilistic framework. In contrast to traditional cluster analytical techniques, LPA is model-based, whereas the hierarchical and *K*-means clustering methods are not (McLachlan et al., 2019; Pastor et al., 2007). Remarkably, this study had 81 different cases, and we treated each case as a sample due to the small sample size. To reach a stable solution, we constrained the variances of cluster indicators to be equal, but the means can vary across clusters (Scherer et al., 2017).

Using the "tidyLPA" packages in R (Rosenberg et al., 2019), we applied the maximum likelihood (ML) algorithm to estimate the model parameters and generate fit statistics for six candidate models with k values ranging from 1 to 6. Because the number of clusters k is unknown priori in LPA (Nylund et al., 2007), multiple alternative models with different k values should be compared. There are several well-established model fit indices to determine the goodness-of-fit of specific models. First, Akaike Information Criteria (*AIC*), Bayesian Information Criteria (*BIC*), and Sample-Size-Adjusted Bayesian Information Criteria (*SSA-BIC*) were utilized in LPA to decide the number of clusters. The lower the values of these three indices, the better the model fit (Schwarz, 1978). Second, the Bootstrapped Likelihood Ratio Test (*BLRT*) revealed whether a model with k clusters was significantly better than k-l clusters (Lo et al., 2001). The significant result (i.e., p < .05) of BLRT implied that adding a cluster increased the model fit. Third, the estimate of classification certainty was also essential in LPA, and Entropy values > .70 indicated an acceptable accuracy (Celeux & Soromenho, 1996). In addition, the appropriate size of each cluster should be no less than 5% of the sample, which guarantees latent profiles to be theoretically significant and generalized (Pastor et al., 2007).

### 3.3.2. Path analysis

A path analysis was performed to examine the mediating role of cognitive load in SRL behaviors and diagnostic efficiency, using PROCESS Macro in SPSS (Hayes, 2012). The PROCESS Macro can automatically execute computation, run the analysis, and generate meaningful mediation output (Uchechukwu Onu et al., 2020). Specifically, this study employed the "Model 4" embedded in PROCESS Macro, which allows researchers to examine the significance of parallel mediators. In this study, Cognitive load served as the independent variable, the Performance-phase behavior Ratio (PR) and the Self-reflection-phase behavior Ratio (SR) were entered as two parallel mediator variables, and diagnostic efficiency was the dependent variable. The Forethought-phase behavior Ratio (FR) was excluded to avoid multicollinearity. To increase inference accuracy, bootstrapping with 10000 bias-corrected bootstrap samples was conducted to depict the sampling distribution of direct and indirect effects.

### 4. Results

In this section, we first employed the LPA to classify medical students as high- and low-load clusters. Then we compared the differences in macro- and micro-level SRL behaviors between the high- and low-load cluster to

examine the effects of cognitive load on SRL. To further investigate the joint predictive functions of cognitive load and SRL, we performed the path analysis.

# 4.1. Did students demonstrate different cognitive load patterns when they diagnosed virtual patient cases in a technology-rich learning environment?

As aforementioned, we utilized the LPA to identify medical students' profiles of cognitive load based on four linguistic markers extracted by the LIWC, i.e., *positive emotion, cognitive discrepancy, insight,* and *causation.* The model fit indices for LPA with clusters ranging from 1 to 6 are shown in Table 2. The 2-cluster solution was deemed superior to the one-cluster solution due to the lower *AIC, BIC,* and *SSA-BIC* values and the significant result of *BLRT* (p = .01). However, the *BLRT* demonstrated that the 3-cluster solution did not have a significantly better fit than the 2-cluster solution (p = .70). The 4-cluster solution fitted better than a 3-cluster solution with decreased *AIC* and *SSA-BIC* values and significant *BLRT* revealed that they did not fit better than a 4-cluster solution (p = .62 and .31, respectively). As for the comparison between the 2-cluster solution and 4-cluster solution, we deemed the former was a better fit than the latter, considering the higher Entropy of a 2-cluster solution (*Entropy* = .81) than that of the 4-cluster solution (*Entropy* = .74). Overall, the 2-cluster solution is optimal for differentiating medical students' cognitive load profiles, with 21% and 79% of the students being labelled as high- and low-load cluster.

As shown in Figure 3, the two clusters represent distinct cognitive load patterns. Cluster 1 consisted of 17 (21%) cases in which students experienced more cognitive discrepancy (M = 3.25), insight (M = 3.27), and causation processes (M = 2.09), but less positive emotions (M = 3.10). In contrast, the 64 cases (79%) in Cluster 2 demonstrated less cognitive discrepancy (M = 1.68), insight (M = 2.93), and cognitive causations (M = 1.69) bur more positive emotions (M = 2.97). Therefore, we labelled Cluster 1 and Cluster 2 as high- and low-load group, respectively.

Table 2. Fit indices for models with number of clusters ranging from 1 to 6

Models	AIC	BIC	SSA-BIC	Entropy	BLRT_p	n_min
1 Cluster	857	876	851	1.00		1.00
2 Clusters	840	871	830	.81	.01	.21
3 Clusters	845	888	831	.63	.70	.11
4 Clusters	835	890	817	.74	.02	.14
5 Clusters	838	905	817	.74	.62	.14
6 Clusters	840	919	815	.76	.31	.04

*Note.* BLPT\_*p* refers to the *p* values for the Bootstrapped Likelihood Ratio Test, *n\_min* refers to the ratio of respondents in clusters with the smallest sample size.



#### 4.2. Did cognitive load patterns affect SRL behaviors?

A series of independent t-tests were conducted to examine how high- and low-load groups differed in macrolevel (i.e., forethought, performance, and self-reflection) and micro-level SRL behaviors. Table 3 showed that the ratio of SRL behaviors in the forethought phase (FR) did not significantly differ between high- and low-load cases. The ratio of SRL behaviors in the performance phase (PR) was significantly higher in high-load cases (M = 52.53, SD = 14.99) than that in the low-load cases (M = 41.23, SD = 12.92), t(79) = 3.10, p = .003. However, high-load cases (M = 32.06, SD = 15.14) led to a significantly lower ratio of SRL behaviors in the self-reflection phase than the low-load cases (M = 41.69, SD = 32.06), t(79) = -2.54, p = .013. In addition, the effect sizes of cognitive load levels on PR and SR were large (Cohen's d = .85 and -.69, respectively).

The results of the micro-level analysis are illustrated in Table 4. Students in the high-load case showed significant differences from those in the low-load cases in two micro-level SRL behaviors: Execution (performance phase) and Hypothesis Evaluation (self-reflection phase). The ratio of Execution behavior was higher for the high-load cluster (M = 14.41, SD = 7.13) compared to the low-load cluster (M = 10.48, SD = 5.83), t = 2.35, p = .21, Cohen's d = .59. In contrast, the ratio of Hypothesis Evaluation behavior was significantly higher in low-load cases (M = 24.80, SD = 14.14) than that in high-load cases (M = 16.53, SD = 12.95), t = -2.15, p = .035, Cohen's d = ..59.

*Table 3.* The predicative role of cognitive load level on macro-level SRL behavior ratios

	High-load cases $(n = 17)$		Low-load cases $(n = 64)$		t	Sig.	Cohen's d
	М	SD	М	SD			
FR	15.29	3.74	17.14	4.88	-1.45	.152	40
PR	52.53	14.99	41.23	12.92	3.10	.003**	.85
SR	32.06	15.14	41.69	13.56	-2.54	.013*	69

*Note.* FR = Forethought-phase behavior Ratio, PR = Performance-phase behavior Ratio, SR = Self-reflection-phase behavior Ratio. p < .05; p < .01, p < .01.

Table 4. The predicative role of cognitive load level on micro-level SRL behavior ratios

	High-load ca	High-load cases $(n = 17)$		Low-load cases $(n = 64)$		Sig.	Cohen's d
	М	SD	М	SD			
TAR	15.29	3.74	17.14	4.88	-1.45	.152	40
ORR	16.88	7.11	16.26	5.66	.37	.714	.10
EXR	14.41	7.13	10.48	5.83	2.35	.021*	.64
HSR	5.88	6.70	4.34	5.50	.98	.331	.23
LER	15.35	10.46	10.28	9.81	1.87	.065	.52
EER	15.47	3.81	16.91	5.46	-1.02	.312	28
HER	16.53	12.95	24.80	14.14	-2.15	.035*	59

*Note.* TAR = Task Analysis Ratio, ORR = Orientation Ratio, EXR = Execution Ratio, HSR = Help-Seeking Ratio, LER = Link Evidence Ratio, EER = Evidence Evaluation Ratio, HER = Hypothesis Evaluation Ratio. \*p < .05; \*p < .01, \*\*p < .001.

### 4.3. How did cognitive load and SRL behaviors jointly predict diagnostic efficiency?

As aforementioned, path analysis was conducted to test the hypothesis that SRL behavior ratios mediated the relationships between cognitive load and diagnostic efficiency. This study did not include the Forethought-phase behavior Ratio (FR) in the model to avoid the issue of collinearity. The results in Table 5 showed that the direct effect of cognitive load on diagnostic efficiency was not significant ( $R^2 = .25$ , F(3, 77) = 1.42, p > .05). Cognitive load positively predicted the Performance-phase behavior Ratio (PR) ( $\beta = .26$ , p < .05), but it was a negative predictor of the Self-reflection-phase behavior Ratio (SR) ( $\beta = -.24$ , p < .05). In addition, SRL behaviors in the self-reflection phase positively predicted diagnostic efficiency ( $\beta = .88$ , p < .01), whereas SRL behaviors in the performance phase had no significant predictive effect on diagnostic efficiency ( $\beta = .53$ , p = .11). Although the direct effect of cognitive load on diagnostic efficiency was not significant, the indirect effect through SRL was significant ( $\beta = -.21$ , 95% *CI* [-.46, -.01]). Overall, the effect of cognitive load on diagnostic efficiency are shown in Figure 4.

Figure 4. The mediating role of SRL behavior ratios between cognitive load and diagnostic efficiency



*Note.* PR = Performance-phase behavior Ratio, SR = Self-reflection-phase behavior Ratio. The dotted lines represent insignificant effects. \*p < .05; \*\*p < .01, \*\*\*p < .001.

Table 5. Mediation model						
Model	β	SE	t	р	CI (lower)	CI (upper)
$X \rightarrow M_1(a1)$	.26	.66	2.27	.026*	.18	2.80
$X \rightarrow M_2(a2)$	24	.68	-2.00	$.049^{*}$	-2.71	01
$M_1 \rightarrow Y(b1)$	.53	.53	1.62	.109	20	1.91
$M_2 \rightarrow Y(b2)$	.88	.51	2.73	$.008^{**}$	.38	2.43
$X \rightarrow Y(c)$	03	1.08	24	.808	-2.42	1.89
$X \rightarrow Y(c')$	.04	1.03	.35	.725	-1.68	2.41
$X \rightarrow M_1 \rightarrow Y(a1*b1)$	.14	.10			02	.34
$X \rightarrow M_2 \rightarrow Y (a2*b2)$	21	.12			46	01

*Note.*  $X = \text{cognitive load}, M_1 = \text{Performance-phase behavior Ratio}, M_2 = \text{Self-Reflection-phase behavior Ratio}, Y = diagnostic efficiency; a1 and a2 represent the direct effect of X on <math>M_1$  and  $M_2$ , respectively; c means the total effect of X on Y; c' refers to the direct effect of X on Y. \*p < .05; \*\*p < .01.

### 5. Discussion

We applied text mining techniques to extract cognitive load indicators from students' concurrent think-aloud protocols as they diagnosed virtual patient cases in BioWorld. Specifically, four indicators, including *cognitive discrepancy, insight, causation,* and *positive emotions,* were selected in this study. We then employed the LPA on cognitive load indicators to cluster medical students to see if any patterns of cognitive load emerged in solving the tasks. The results from LPA demonstrated that medical students could be identified as high- and low-load group when addressing the tasks. Notably, medical students with a high cognitive load experienced fewer positive emotions and exerted more cognitive effort (i.e., more cognitive discrepancies, insight, and causation activities) compared to those who experienced a low-level cognitive load.

Specifically, students may encounter large knowledge gaps in high cognitive load situations and find more inconsistencies between their acquired knowledge and the ongoing learning task (Reiser, 2004). Therefore, they tended to use words such as *should* and *would*, to express their *cognitive discrepancies*. As well, a high-level cognitive load required students to conduct more cognitive operations, such as thinking, evaluation, and analyses, to achieve predetermined learning goals (Baddeley, 1992; Khawaja et al., 2014; Sweller, 2011). In this study, medical students demonstrated more *insights* (words such as *think* and *know*) and performed more *causal inferences* (words such as *because* and *so*) in high-load contexts. Consistent with Fraser and McLaughlin (2019), we also found that the proportion of positive emotion words decreased with increased cognitive load. According to Pekrun's (2006) control-value theory, individuals' appraisal of perceived controllability over the diagnostic task would be weakened by the high cognitive load level; thus, students were expected to generate more negative emotions in cognitive-demanding contexts. Noticeably, linguistic features developed by the LIWC indicate the overall load instead of distinguishing between three types of cognitive load. For instance, linguistic features cannot differentiate the intrinsic load caused by task complexity from the extraneous load induced by the interface and presentations of the BioWorld. However, the interactions between multidimensional load and SRL matter in learning (Seufert, 2020) and are worthy of further investigation.

The interplay between cognitive load, SRL, and diagnostic efficiency is of primary interest to this study. We found that students with a higher cognitive load had a significantly higher ratio of SRL behaviors in the performance phase but a significantly lower ratio of SRL behaviors in the self-reflection phase. Cognitive load did not affect the ratio of SRL behaviors in the forethought phase. As aforementioned, the high-level cognitive load led students to experience more cognitive discrepancies and uncertainties; thus, it was essential for them to try more operations in the performance phase to collect additional evidence for diagnoses. In this regard, students' mental efforts in the performance phase would occupy a vast of limited working memory resources, suggesting that few cognitive capacities remained for self-reflection behaviors (Seufert, 2018; Sweller, 2011; Winne, 2001). Therefore, students in the high-load cluster had a lower ratio of SRL behaviors in the self-reflection phase compared to the low-load cluster.

Furthermore, this study found that the micro-level SRL behaviors in the performance and self-reflection phases were affected by cognitive load levels. Specifically, students conducted relatively more *Execution* behaviors and fewer *Hypothesis Evaluation* behavior when experiencing a higher cognitive load. Students with a high cognitive load might activate all relevant cognitive schemas stored in the long-term memory; thus, they were inclined to conduct more lab tests (Execution) to reduce the feeling of uncertainty and to ease the cognitive load. The SRL behavior of *Hypothesis Evaluation* required students to integrate all information obtained from the forethought and performance phases to make a judgment about the proposed hypotheses. This SRL behavior imposed

enormous cognitive burdens on the working memory system (de Bruin & van Merriënboer, 2017). However, a high cognitive load indicates limited working memory resources, which prevent students from performing many *Hypothesis Evaluation* behaviors.

As for the joint effects of SRL behaviors and cognitive load on diagnostic efficiency, the path analysis revealed that cognitive load negatively predicted diagnostic efficiency by influencing the ratio of SRL behaviors in the self-reflection phase. Specifically, cognitive load positively predicted students' efforts in performance-phase operations but negatively predicted the ratio of SRL behaviors in the self-reflection phase due to the limited cognitive capacity. However, the ratio of SRL behaviors in the self-reflection phase was a positive indicator of diagnostic efficiency. The self-reflection behaviors facilitated students to construct a more comprehensive and deeper understanding of the task and led them to elaborate on their problem-solving processes (Lew & Schmidt, 2011). Thus, SRL behaviors in the self-reflection phase were beneficial to students' diagnostic efficiency, which was in line with the findings of Zheng et al. (2020). Moreover, we found that SRL behaviors in the self-reflection phase results suggested that effective allocation strategies of working memory resources matter to diagnostic efficiency.

### 6. Conclusion

In conclusion, this paper is the first to explore how cognitive load interacts with SRL behaviors and their joint roles in predicting diagnostic efficiency in the context of clinical reasoning. Theoretically, findings from this study provide empirical evidence for integrating cognitive load and SRL frameworks. Furthermore, this study made a methodological contribution to the measurement of cognitive load by applying text mining techniques to extract cognitive load indicators from students' think-aloud protocols. Moreover, this study has educational implications in that it provides educators with insights regarding how to facilitate students' self-regulated learning and academic performance when they experience a high cognitive load. For example, educators should pay particular attention to students' self-reflection behaviors. When students experience a high-level cognitive load, they may conduct fewer self-reflection behaviors (a positive performance indicator). Under this condition, instructors can provide metacognitive scaffoldings to foster students' metacognitive awareness and self-reflective activities. Moreover, this study informs TREs developers and instructors to design optimal instructional activities to minimize the extraneous load. Otherwise, the high extraneous load would occupy limited cognitive resources and limit SRL behaviors.

While the present study has theoretical, methodological, and practical significance, it is not without limitations. First, this study has a small sample size. Therefore, additional research is needed to verify the findings of this study with a larger number of participants. Second, we did not differentiate between the cognitive load caused by problem-solving tasks and SRL activities due to the limitation of linguistic features. Given that individuals only have limited WMC, it is vital to balance the cognitive load caused by problem-solving and SRL activities to avoid mental overload. Further research should employ various techniques to differentiate these two cognitive load sources. Lastly, the measurement of cognitive load relies entirely on the linguistic features of students, and the validity and reliability of this approach have not yet been verified in the literature. Scholars should combine linguistic features and other cognitive load measures, such as self-ratings, to address this issue in future research.

There are several areas that we will pursue as future research directions. First, we will use multimodal data, such as physiological signals and self-rating, to measure students' cognitive load. A second research direction is to examine the temporal interplay between cognitive load and SRL during the dynamic learning process. Third, this study emphasized the overall load students experienced in a task, and it did not explore the cognitive load in more fine-grained SRL behaviors, such as during planning, monitoring, and evaluation behaviors. The investigation of these research topics will facilitate an integrative framework incorporating cognitive load and SRL theories and guide educators to design more effective instruction activities.

# Acknowledgement

This research was made possible by the Social Sciences and Humanities Research Council of Canada (SSHRC) and the China Scholarship Council (CSC). This study was approved by the Research Ethics Board of McGill University. The authors declare that they have no competing interests.

### References

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review*, 22(4), 425-438. https://doi.org/10.1007/s10648-010-9130-y

Artino Jr, A. R., Cleary, T. J., Dong, T., Hemmer, P. A., & Durning, S. J. (2014). Exploring clinical reasoning in novices: A Self-regulated learning microanalytic assessment approach. *Medical Education*, 48(3), 280-291. https://doi.org/10.1111/medu.12303

Azevedo, R., & Gašević, D. (2019). Analyzing multimodal multichannel data about self-regulated learning with advanced learning technologies: Issues and challenges. *Computers in Human Behavior*, 96, 207–210. https://doi.org/10.1016/j.chb.2019.03.025

Baddeley, A. (1992). Working memory. Science, 255(5044), 556–559. https://doi.org/10.1126/science.1736359

Bower, G. H., Karlin, M. B., & Dueck, A. (1975). Comprehension and memory for pictures. *Memory & Cognition*, 3(2), 216–220. https://doi.org/10.3758/BF03212900

Brünken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 53–61. https://doi.org/10.1207/S15326985EP3801\_7

Brydges, R., & Butler, D. (2012). A Reflective analysis of medical education research on self-regulation in learning and practice. *Medical Education*, 46(1), 71-79. https://doi.org/10.1111/j.1365-2923.2011.04100.x

Celeux, G., & Soromenho, G. (1996). An Entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2), 195–212. https://doi.org/10.1007/BF01246098

Charlin, B., Boshuizen, H. P., Custers, E. J., & Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education*, 41(12), 1178-1184. https://doi.org/10.1111/j.1365-2923.2007.02924.x

Cleary, T. J., Durning, S. J., & Artino, A. R. (2016). Microanalytic assessment of self-regulated learning during clinical reasoning tasks: Recent developments and next steps. *Academic Medicine*, *91*(11), 1516-1521.

Cowan, N. (2011). The Focus of attention as observed in visual working memory tasks: Making sense of competing claims. *Neuropsychologia*, 49(6), 1401–1406. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2011.01.035

Darabi, A., Arrastia, M. C., Nelson, D. W., Cornille, T., & Liang, X. (2010). Cognitive presence in asynchronous online learning: A Comparison of four discussion strategies. *Journal of Computer Assisted Learning*, 27(3), 216–227. https://doi.org/10.1111/j.1365-2729.2010.00392.x

de Bruin, A. B., Roelle, J., Carpenter, S. K., & Baars, M. (2020). Synthesizing cognitive load and self-regulation theory: A Theoretical framework and research agenda. *Educational Psychology Review*, 32(4), 903-915. https://doi.org/10.1007/s10648-020-09576-4

de Bruin, A. B., & van Merriënboer, J. J. (2017). Bridging cognitive load and self-regulated learning research: A Complementary approach to contemporary issues in educational research. *Learning and Instruction*, 51, 1–9.

Durning, S., Artino Jr, A. R., Pangaro, L., van der Vleuten, C. P., & Schuwirth, L. (2011). Context and clinical reasoning: understanding the perspective of the expert's voice. *Medical Education*, 45(9), 927-938. https://doi.org/10.1111/j.1365-2923.2011.04053.x

Fraser, K., Ma, I., Teteris, E., Baxter, H., Wright, B., & Mclaughlin, K. (2012). Emotion, cognitive load and learning outcomes during simulation training. *Medical Education*, 46(11), 1055–1062. https://doi.org/10.1111/j.1365-2923.2012.04355.x

Fraser, K., & McLaughlin, K. (2019). Temporal pattern of emotions and cognitive load during simulation training and debriefing. *Medical Teacher*, 41(2), 184-189. https://doi.org/10.1080/0142159X.2018.1459531

Greene, J. A., & Azevedo, R. (2007). A Theoretical review of Winne and Hadwin's model of self-regulated learning: New perspectives and directions. *Review of Educational Research*, 77(3), 334–372. https://doi.org/10.3102/003465430303953

Hayes, A. F. (2012). *PROCESS: A Versatile computational tool for observed variable moderation, mediation, and conditional process modeling* [White paper]. http://www.afhayes.com/public/process2012.pdf

Joseph, A. W., & Murugesh, R. (2020). Potential eye tracking metrics and indicators to measure cognitive load in humancomputer interaction research. *Journal of Scientific Research*, 64(01), 168–175. https://doi.org/10.37398/jsr.2020.640137

Khawaja, M. A., Chen, F., & Marcus, N. (2012). Analysis of collaborative communication for linguistic cues of cognitive load. *Human Factors*, *54*(4), 518–529. https://doi.org/10.1177/0018720811431258

Khawaja, M. A., Chen, F., & Marcus, N. (2014). Measuring cognitive load using linguistic features: Implications for usability evaluation and adaptive interaction design. *International Journal of Human-Computer Interaction*, 30(5), 343–368. https://doi.org/10.1080/10447318.2013.860579 Kirschner, P. A. (2002). Cognitive load theory: Implications of cognitive load theory on the design of learning. *Learning and instruction*, *12*(1), 1-10. https://doi.org/10.1016/S0959-4752(01)00014-7

Konopasky, A., Durning, S. J., Artino, A. R., Ramani, D., & Battista, A. (2020). The Linguistic effects of context specificity: exploring affect, cognitive processing, and agency in physicians' think-aloud reflections. *Diagnosis*, 7(3), 273–280. https://doi.org/10.1515/dx-2019-0103

Kuiper, R. A. (2013). Integration of innovative clinical reasoning pedagogies into a baccalaureate nursing curriculum. *Creative Nursing*, *19*(3), 128–139. https://doi.org/10.1891/1078-4535.19.3.128

Kuiper, R., Pesut, D., & Kautz, D. (2009). Promoting the self-regulation of clinical reasoning skills in nursing students. *The Open Nursing Journal*, *3*, 76. https://doi.org/10.2174/1874434600903010076

Lajoie, S. P. (2009). Developing professional expertise with a cognitive apprenticeship model: Examples from Avionics and Medicine. In K. A. Ericsson (Ed.), *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments* (pp. 61–83). Cambridge University Press.

Lajoie, S. P., & Lu, J. (2012). Supporting collaboration with technology: Does shared cognition lead to co-regulation in medicine? *Metacognition and Learning*, 7(1), 45-62.

Leppink, J., Paas, F., Van der Vleuten, C. P. M., Van Gog, T., & Van Merriënboer, J. J. G. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods*, 45(4), 1058–1072. https://doi.org/10.3758/s13428-013-0334-1

Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior*, 25(2), 284–289. https://doi.org/10.1016/j.chb.2008.12.010

Lew, M. D., & Schmidt, H. G. (2011). Self-reflection and academic performance: Is there a relationship? Advances in Health Sciences Education, 16(4), 529-545.

Li, S., Zheng, J., & Lajoie, S. P. (2020). Efficient clinical reasoning: Knowing when to start and when to stop. *Education in the Health Professions*, *3*(1), 1–7.

Lo, Y., Mendell, N. R., & Rubin D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88(3), 767–778. https://doi.org/10.1093/biomet/88.3.767

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. Annual Review of Statistics and Its Application, 6, 355–378. https://doi.org/10.1146/ANNUREV-STATISTICS-031017-100325

Müller, C., Großmann-Hutter, B., Jameson, A., Rummer, R., & Wittig, F. (2001, July). Recognizing time pressure and cognitive load on the basis of speech: An Experimental study. In *International Conference on User Modeling* (pp. 24-33). Springer.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 535-569. https://doi.org/10.1080/10705510701575396

Paas, F. (1992). Training strategies for attaining transfer of problem-solving skill in statistics: A Cognitive load approach. *Journal of Educational Psychology*, 84(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63–71. https://doi.org/10.1207/S15326985EP3801\_8

Paas, F., Van Gog, T., & Sweller, J. (2010). Cognitive load theory: New conceptualizations, specifications, and integrated research perspectives. *Educational Psychology Review*, 22(2), 115-121. https://doi.org/10.1007/s10648-010-9133-8

Paas, F., Van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual and Motor Skills*, 79(1 Pt 2), 419–430. https://doi.org/10.2466/pms.1994.79.1.419

Panadero, E. (2017). A Review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, *8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Park, B., Korbach, A., & Brünken, R. (2015). Do learner characteristics moderate the seductive-details-effect? A Cognitive-load-study using eye-tracking. *Educational Technology & Society*, 18(4), 24–36.

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A Latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, *32*(1), 8–47. https://doi.org/10.1016/j.cedpsych.2006.10.003

Pekrun, R. (2006). The Control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, 18(4), 315–341. https://doi.org/10.1007/s10648-006-9029-9

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The Development and psychometric properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.15781/T29G6Z

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54, 547–577. https://doi.org/10.1146/ANNUREV.PSYCH.54.101601.145041

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaert, P. R. Pintrich, & M. Zerdner (Eds.), *Handbook of self-regulation* (1st ed., pp. 451-502). Academic Press. https://doi.org/10.1016/b978-012109890-2/50043-3

Reedy, G. B. (2015). Using cognitive load theory to inform simulation design and practice. *Clinical Simulation in Nursing*, 11(8), 355-360. https://doi.org/10.1016/j.ecns.2015.05.004

Reiser, B. J. (2004). Scaffolding complex learning: The Mechanisms of structuring and problematizing student work. *Journal of the Learning Sciences*, 13(3), 273–304. https://doi.org/10.1207/s15327809jls1303\_2

Robinson, R. L., Navea, R., & Ickes, W. (2012). Predicting final course performance from students' written selfintroductions: A LIWC analysis. *Journal of Language and Social Psychology*, 32(4), 469–479. https://doi.org/10.1177/0261927X13476869

Rosenberg, J. M., Beymer, P. N., Anderson, D. J., Van Lissa, C. J., & Schmidt, J. A. (2019). tidyLPA: An R package to easily carry out Latent Profile Analysis (LPA) using open-source or commercial software. *Journal of Open Science Software*, *3*(30), 978. https://doi.org/10.21105/joss.00978

Scheiter, K., Ackerman, R., & Hoogerheide, V. (2020). Looking at mental effort appraisals through a metacognitive lens: Are they biased? *Educational Psychology Review*, *32*(4), 1003–1027. https://doi.org/10.1007/s10648-020-09555-9

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.

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.

Seufert, T. (2018). The Interplay between self-regulation in learning and cognitive load. *Educational Research Review*, 24, 116–129. https://doi.org/10.1016/j.edurev.2018.03.004

Seufert, T. (2020). Building bridges between self-regulation and cognitive load—An Invitation for a broad and differentiated attempt. *Educational Psychology Review*, *32*(4), 1151-1162. https://doi.org/10.1007/s10648-020-09574-6

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: The Links between language, performance, error, and workload. *Journal of Human Performance in Extreme Environments*, 5(1), 63–68.

Simmons, B. (2010). Clinical reasoning: Concept analysis. Journal of Advanced Nursing, 66(5), 1151–1158.

Solhjoo, S., Haigney, M. C., McBee, E., van Merrienboer, J. J., Schuwirth, L., Artino, A. R., Battista, A., Ratcliffe, T, A., Lee, H. D., & Durning, S. J. (2019). Heart rate and heart rate variability correlate with clinical reasoning performance and self-reported measures of cognitive load. *Scientific Reports*, 9(1), 1-9. https://doi.org/10.1038/s41598-019-50280-3

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2), 257-285. https://doi.org/10.1016/0364-0213(88)90023-7

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of Learning and Motivation* (pp. 37-76). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Uchechukwu Onu, D., Iorfa, S. K., & Ugwu, D. I. (2020). Negative centralisation of HIV/AIDS trauma and health-related quality of life: do post-traumatic stress symptoms explain the link? *African Journal of AIDS Research*, *19*(3), 206-213. https://doi.org/10.2989/16085906.2020.1797842

Wiley, J., & Jarosz, A. F. (2012). Working memory capacity, attentional focus, and problem solving. *Current Directions in Psychology Science*, 21(4), 258-262. https://doi.org/10.1177/0963721412447622

Winne, P. H. (2001). Self-regulated learning viewed from models of information Self-regulated learning viewed from models of information processing. In B. J. Zimmerman & D. H. Schunk (Eds.), *Self-Regulated Learning and Academic Achievement: Theoretical Perspectives* (pp. 145–178). Routledge.

Winne, P. H. (2018). Theorizing and researching levels of processing in self-regulated learning. *British Journal of Educational Psychology*, 88(1), 9-20. https://doi.org/10.1111/bjep.12173

Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaert, P. R. Pintrich, & M. Zerdner (Eds.), *Handbook of self-regulation* (1st ed., pp. 531-566). Academic Press. https://doi.org/10.1016/B978-012109890-2/50045-7

Zheng, J., Li, S., & Lajoie, S. P. (2020). The Role of achievement goals and self-regulated learning behaviors in clinical reasoning. *Technology, Knowledge and Learning*, 25(3), 541-556.

Zimmerman, B. J. (2000). Attaining self-regulation: A Social cognitive perspective. In M. Boekaert, P. R. Pintrich, & M. Zerdner (Eds.), *Handbook of self-regulation* (1st ed., pp. 13-39). Academic Press. https://doi.org/10.1016/b978-012109890-2/50031-7

# Lag Sequential Analysis for Identifying Blended Learners' Sequential Patterns of e-Book Note-taking for Self-Regulated Learning

# Christopher C.Y. Yang<sup>1\*</sup> and Hiroaki Ogata<sup>2</sup>

<sup>1</sup>Graduate School of Informatics, Kyoto University, Japan // <sup>2</sup>Academic Center for Computing and Media Studies, Kyoto University, Japan // yang.yuan.57e@st.kyoto-u.ac.jp // ogata.hiroaki.3e@kyoto-u.ac.jp \*Corresponding author

(Submitted April 18, 2022; Revised August 23, 2022; Accepted September 17, 2022)

ABSTRACT: Blended learning (BL) is regarded as an effective strategy for combining traditional face-to-face classroom activities with various types of online learning tools (e.g., e-books). An effective feature of e-books is the ability to use digital notes. When e-books are used in BL, the strategic adoption of note-taking provides benefits that influence the learners' progress for self-regulated learning (SRL) and course achievements. However, learners tend to be unsure about how note-taking is performed using online learning materials and lack knowledge of effective strategies for SRL. Furthermore, few studies have investigated blended learners' sequential patterns of e-book note-taking for SRL. Thus, in this paper, an exploratory study was conducted in an undergraduate course that implemented the BL design. The learning task for the blended learners in the present study was to study the learning material using BookRoll, an e-book system, during in-class and out-of-class learning sessions. Lag sequential analysis of the e-book learning behavior data was conducted to identify the blended learners' sequential behaviors of e-book note-taking for the cognitive strategy use of SRL. Moreover, the difference between higher- and lower-achievement blended learners in terms of their sequential behaviors of e-book note-taking for SRL was revealed. This study can help educators provide evidence-based educational feedback to learners regarding the identified sequential patterns of e-book note-taking that can be applied as effective strategies for promoting the cognitive strategy use of SRL and improvement of course achievement in BL.

Keywords: Lag sequential analysis, Sequential pattern, Note-taking, Blended learning, Self-regulated learning

# **1. Introduction**

Blended learning (BL) is regarded as an effective combination of face-to-face and online learning experiences, and this new education domain emphasizes the need to reflect on traditional learning experiences to redesign learning and teaching strategies (Garrison & Vaughan, 2008). BL aims to combine traditional face-to-face classroom activities with various types of online learning resources, and it enables learners to achieve improved learning outcomes through a well-defined interactive strategy.

In the early stages of BL development, time- and cost-related factors were major challenges (Míguez-Álvarez et al., 2020). However, studies have demonstrated that the effective use of online learning technologies in BL has a positive impact on improving learner learning engagement (Castro, 2019), learning performance (Yang et al., 2021), motivation (Álvarez et al., 2013), and self-efficacy (Moon & Hyun, 2019). In the online learning activities of a BL course, learners' interaction with various online educational platforms provides a massive amount of learning interaction data that can be captured and analyzed by educational technologies. These advanced educational technologies have been employed to automate the processes for information delivery by offering a personalized learning experience for the individual learner to enhance their engagement in learning (Castro, 2019).

BL describes a learner-centered, self-paced, and flexible digital environment in which traditional face-to-face classroom activities are supported by offline or online activities via educational technologies (Tang & Chaw, 2016; Anthonysamy et al., 2020). Consequently, the promotion of self-regulated learning (SRL) in BL contexts is essential since SRL generally refers to "self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals" (Song et al., 2021; Zimmerman, 2000, p. 14). Studies have examined instructional materials that foster learners' specific strategies for SRL including self-monitoring (e.g., Kauffman et al., 2008) and note-taking (e.g., Igo & Kiewra, 2007; Igo et al., 2005).

Note-taking behavior can be regarded as a reflection of the progress of learner learning. Thus, the analysis of note-taking plays a role in tracking and monitoring the learning process of learners who participate in BL or fully online courses (Nakayama et al., 2021), which may also promote the cognitive strategy use of SRL. The taking

of digital notes on online materials has been indicated as an essential strategy for learners studying online materials using educational tools (e.g., e-books) (van de Sande et al., 2017). Observations of learners in classrooms have revealed various note-taking strategies, some of which involve learners meeting their personal learning needs by modifying the materials given to them by faculty members. The note-taking behavior of learners is a topic that warrants further exploration (Stacy & Cain, 2015). Furthermore, note-taking behavior is correlated with achievement (Luo et al., 2018; van de Sande et al., 2017) because it improves retention and recall (Fisher & Harris, 1973), increases attention to material (Kane et al., 2017), and provides several memory benefits (i.e., storage and encoding; Peverly & Wolf, 2019). When learners do not apply effective note-taking techniques during lectures, they may overlook key concepts and content (Boyle, 2010). Note-taking is an essential skill that all learners must have to achieve success in a classroom. Learners should use the note-taking medium that maximizes their willingness and ability to achieve a delicate balance between practicality, ease of implementation, and efficacy concerning note-taking strategies (Dror, 2008).

In the present exploratory study, to understand blended learners' interactions with the note-taking systems (e.g., e-books) for the cognitive strategy use of SRL, lag sequential analysis (LSA) was applied to analyze learnergenerated e-book learning behavioral data collected in a BL environment since LSA was proposed by Sackett (1978) as an effective method that has been used to conduct detailed investigations of the sequential behaviors of learners in the educational domain (Yang et al., 2018; Zarzour et al., 2020; Zhang et al., 2021). Moreover, the difference between higher- and lower-achievement blended learners in terms of their sequential behaviors of e-book note-taking for SRL was revealed and discussed. The analytical results are expected to offer opportunities for educators to effectively understand learners' interactions with e-books and provide learners with evidence-based educational feedback regarding note-taking strategies for the promotion of SRL and improvement of course achievement in BL. The results are also expected to be considered as a basis by teachers at every education level and learners for adjusting their teaching and learning strategies in BL, respectively. In the present study, the following two research questions are addressed:

- To what extent can the blended learners' sequential patterns of e-book note-taking for SRL be identified by using LSA?
- What are the differences between higher- and lower-achievement blended learners in terms of their sequential patterns of e-book note-taking for SRL?

### 2. Literature review

### 2.1. Note-taking strategies for SRL

In higher education, note-taking is regarded as an effective strategy for learners to enhance their learning (Wu, 2020). The appropriate self-regulatory strategies regarding the certain actions and processes that individuals adopt to succeed is a key element for being self-regulated (Zimmerman, 1989). According to Pressley and Woloshyn (1995), the cognitive strategy use of SRL involves cognitive operations for the process of carrying out a task. A strategy that fosters students' abilities to efficiently locate and organize knowledge from the learning materials is important (Kauffman et al., 2011). Therefore, note-taking has been recognized as a key part of the organization aspect of a cognitive SRL strategy that aims at retaining information from the learning materials and monitoring the learning process of learners (Cengiz-Istanbullu & Sakiz, 2022; Pintrich et al., 1991).

Trevors et al. (2014) implied that learners with different self-regulatory skills may exhibit different behaviors of note-taking and note-reviewing. Hence, different patterns in the content of notes recorded were shown. Learners' patterns of note-taking may differ according to their level of prior knowledge, metacognitive awareness, capabilities of adopting effective self-regulatory strategies, and the instructional support available to them (Moos & Azevedo, 2008). Therefore, their results suggested that note-taking is an essential and challenging skill for learners to master and for scholars to uncover in the context of the learning process and achievement.

Given the importance of considering note-taking as an SRL strategy and the challenges of implementing these strategies, Alvi et al. (2016) conducted a qualitative study on 37 university learners in a two years Master's degree program to uncover the SRL strategies used by the learners. Their results suggested that learners tend to use a variety of SRL techniques ranging from the shallow strategy (i.e., repetition for memorization) to the cognitively deep processing strategy (i.e., note-taking and consulting notes). Particularly, high-achieving learners exhibit superior meta-cognitive awareness of taking and consulting notes. Therefore, they indicated that there is a need to guide and assist learners in moving beyond the traditional practice of note-taking to promote SRL.

The above studies demonstrated the role played by note-taking strategies for promoting the cognitive strategy use of SRL. However, SRL should be viewed as an ongoing process that is developed by learners over time (Azevedo et al., 2010). Analyzing sequence data can reveal the transitional relationships between the different categories of learning behaviors and the temporality perspective of the learning process of learners rather than using count-based measures to quantify learner behavior in specific contexts (Chen et al., 2017) such as SRL. In this sense, few studies have investigated the sequential patterns of note-taking of learners for SRL. Moreover, since the development of SRL strategies is essential for learners undergoing BL, there is a need to particularly uncover the sequential patterns of note-taking of blended learners.

### 2.2. Identification of patterns of e-book note-taking

With the increasing use of technology in education, e-books are gradually replacing traditional textbooks and changing the way learners learn, think, and interact with learning materials (Casselden & Pears, 2020; Sung & Wu, 2017; Wright et al., 2013). For e-book-based learning, note-taking strategies are useful for helping learners to understand online learning materials. Numerous studies have demonstrated that enriching learning activities with various advanced educational technologies lead to enhanced reading ability (Wu, 2016) and improved comprehension outcomes for learners (Huang & Liang, 2015). E-book-based learning systems have positive effects on aspects such as learning motivation, perceived usefulness and ease of use, rapid knowledge construction, and level of comfort during particular course activities, all of which can increase the engagement of learners in a learning process (Lin et al., 2018).

With the increasing and widespread use of e-books, learners can now take notes digitally through various electronic devices (e.g., laptops, tablets, or mobile phones) instead of using pen and paper (Chiu et al., 2013). Note-taking can enhance the learning activities of learners during a course by directing their attention and building both internal and external connections (Du, 2004). Studies have demonstrated a positive correlation between the frequency of annotation use by learners and their academic performance during e-book learning (Yang et al., 2021). To achieve success in e-book-based BL, learners must strategically record their notes in their online learning materials. However, for learners who study in traditional face-to-face teaching sessions, note-taking is still a challenging task (Hanafin et al., 2007). For learning to occur, learners must actively listen to their teachers, memorize relevant information, and connect and relate this new information to the ideas that they learned in their classes (O'Hara, 2005).

Although strategic digital note-taking is generally recommended for online learning materials, Dunn (2015) reported that learners were unsure about how they can take notes and lacked knowledge of the effectiveness of their strategies for learning. Analyzing learners' sequential patterns of note-taking may offer opportunities for educators to effectively understand learners' interactions with note-taking systems (e.g., e-books) and provide learners with corresponding interventions regarding note-taking strategies for learning. However, few studies have empirically investigated the sequential patterns of the note-taking performed by learners when they are using an e-book in a BL environment.

### 3. Method

### **3.1.** Participants and context of the exploratory study

An exploratory study was conducted in an undergraduate course called Accounting Information Systems. This course implemented the BL design with a total of 88 undergraduate learners participating. These participants were from the Department of Accounting. They had an average age of 21 years, and 30 (34.1%) and 58 (65.9%) of them were male and female, respectively. No participant dropped out of the study. The learning task designed for the participants in the present study was to study the learning material using BookRoll, an e-book system, which was developed by the Ogata et al. (2015). Figure 1 shows an example of the user interface of BookRoll. In addition to traditional face-to-face learning activities, the participants who enrolled in this course studied the learning material uploaded by the instructor before their classes; they achieved this by using various electronic devices (e.g., desktops, laptops, and mobile phones) to access the BookRoll system during in-class and out-of-class learning sessions. The system had several functions such as page-turning, marker drawing, memo taking, and page jumping. Data on the learning behaviors of the participants when they were using BookRoll were stored in its database. The functions of BookRoll are discussed in detail in a previous study by the Ogata et al. (2015).



### 3.2. Procedure

The present study was conducted following the research procedure presented in Figure 2. The duration of the learning task designed for the present exploratory study was 6 weeks. First, the course instructor introduced the course syllabus and the use of BookRoll. Second, the instructor uploaded the course learning materials to BookRoll and assigned several learning tasks to the participants taking the blended course. Third, the participants studied the learning materials by using the BookRoll system during both in-class face-to-face learning sessions and out-of-class self-learning sessions, and they accessed the system through their electronic devices (e.g., desktops, laptops, and mobile phones). In this stage, the participants were highly encouraged to take notes using the memo function of the BookRoll system to enhance their understanding of the knowledge contained in the learning materials. Fourth, after the participants completed their 6-week learning task, data on their learning behaviors while using BookRoll were collected from BookRoll's database and preprocessed for follow-up data analysis. Last, the collected learning behavior data of the participants were coded, such that an LSA could be performed to extract sequential patterns of note-taking behaviors. The extracted sequential patterns of e-book note-taking behaviors of higher- and lower-achievement participants were compared based on their learning achievements during the course.



### 3.3. Data collection, preprocessing, and analysis

In the present study, 82,443 data of the learning behaviors of learners while using the BookRoll system were collected from the system's database. To reduce redundant information during data analysis, each type of learning behavior was only counted once when it was observed to have occurred more than once over a continuous period. For example, if the learning behavior "ADD MEMO" consecutively occurred three times during a single learning session, it was still only counted as a single instance of the behavior. Furthermore, to improve the homogeneity of the collected learner data, data preprocessing was performed to remove outlier data relating to learning sessions (i.e., multiple successive learning actions performed in BookRoll). Specifically, data related to overly short sessions (i.e., those that involved only one type of learning behavior) and overly long sessions (i.e., those in which the number of learning behaviors observed was greater than those observed in 95% of all examined sessions) were removed (Jovanović et al., 2017). Table 1 provides an example of the collected BookRoll learning behavior data. Each piece of behavioral data of BookRoll interactions included user ID, content ID, operation name, and operation date information. The learning behavior data of the learners, which were collected using BookRoll, were coded to enable the subsequent sequential pattern mining of note-taking behaviors. The coded BookRoll learning behavior data and their corresponding descriptions are presented in Table 2. Notably, in the present study, only behavioral data on page-turning and the use of memos, markers, and bookmarks, were collected for the pattern analysis of e-book note-taking. In addition to the collection of BookRoll learning behavior data, a final examination was conducted to measure the learners' learning achievements for the course. The participants' scores for the final examination were compiled by the course instructor at the end of the learning task. The examination comprised 40 multiple-choice items, and a maximum score of 100 could be obtained. For each correctly answered item, 2.5 points were awarded; no points were awarded for incorrectly answered items. The final examination had a Kuder-Richardson Formula 20 value of 0.59, indicating that it had acceptable internal consistency (Cortina, 1993).

To analyze the sequential e-book note-taking behaviors of the blended learners, LSA was performed using the Generalized Sequential Querier (GSEQ) software (Bakeman & Quera, 1995). To further explore the differences between higher- and lower-achievement blended learners in terms of their sequential patterns of e-book note-taking, all the participants were classified into a higher-achievement group and a lower-achievement group by applying the percentile rank transformation method to classify their learning achievements (i.e., final examination scores). For example, learners A and B received final examination scores of 40 and 80, respectively, and they were ranked in the 40th and 80th percentiles, respectively, of the scores of all the learners in the course; thus, they were classified into the higher- and lower-achievement groups, respectively. Next, descriptive statistics of the BookRoll behaviors for higher- and lower-achievement groups were analyzed. Finally, LSA was performed to reveal the adjusted residuals of BookRoll sequential behaviors for the two groups. The analysis results are discussed in the next section.

<i>Table 1</i> . Examp	les of collected	BookRoll learnin	g behavior data

User_ID	Content_ID	Operation_Name	Operation_Date
15920	ec645f3851e	OPEN	2021/5/10 10:03:52
15920	ec645f3851e	ADD MEMO	2021/5/10 10:04:32
15920	ec645f3851e	CHANGE MEMO	2021/5/10 10:07:03
15929	ec645f3851e	OPEN	2021/5/10 11:27:14
15929	ec645f3851e	NEXT	2021/5/10 11:27:20

Table 2. Coding scheme of the BookRoll behavioral d	ata
---	-----

Code	BookRoll behavioral data	Description
NE	NEXT	A learner advances to the next page of the e-book learning material.
PR	PREV	A learner returns to the previous page of the e-book learning material.
AM	ADD MEMO	A learner adds a memo to the e-book learning material.
DM	DELETE MEMO	A learner deletes a memo in the e-book learning material.
CM	CHANGE MEMO	A learner modifies a memo in the e-book learning material.
AH	ADD MARKER	A learner adds a marker to the e-book learning material.
DH	DELETE MARKER	A learner deletes a marker in the e-book learning material.
AB	ADD BOOKMARK	A learner adds a bookmark to the e-book learning material.
DB	DELETE BOOKMARK	A learner deletes a bookmark in the e-book learning material.

### 4. Results

### 4.1. Sequential patterns of e-book note-taking

In the present study, LSA was applied to uncover the sequential behaviors of blended learners concerning e-book note-taking. Table 3 presents the collected frequency and percentage data of the BookRoll behaviors of all the blended learners. Among the 2,740 BookRoll-related behavioral data pieces that were collected, there were 882 occurrences of "NEXT," 336 occurrences of "PREV," 832 occurrences of "ADD MEMO," 109 occurrences of "DELETE MEMO", 107 occurrences of "CHANGE MEMO", 428 occurrences of "ADD MARKER", 28 occurrences of "DELETE MARKER", 18 occurrences of "ADD BOOKMARK", and 0 occurrence of "DELETE BOOKMARK". The top 3 behaviors that occurred most frequently were "NEXT" (32.19%), "ADD MEMO" (30.36%), and "ADD MARKER" (15.62%). These findings indicate that "ADD MEMO" and "NEXT" occurred with similar frequencies because the learners in this BL course were highly encouraged to take notes in the learning materials when using BookRoll for both in-class face-to-face and out-of-class self-learning sessions.

Table 3. Frequency and percentage data of the BookRoll behaviors for all the blended learners

Category	Frequency	Percentage (%)
NEXT (NE)	882	32.19
PREV (PR)	336	12.26
ADD MEMO (AM)	832	30.36
DELETE MEMO (DM)	109	3.98
CHANGE MEMO (CM)	107	3.91
ADD MARKER (AH)	428	15.62
DELETE MARKER (DH)	28	1.02
ADD BOOKMARK (AB)	18	0.66
DELETE BOOKMARK (DB)	0	0





Table 4 is an adjusted residual table of the BookRoll sequential behaviors for all the blended learners, and Figure 3 depicts their behavioral transition. A z-score of more than 1.96 indicates the presence of a significant sequential relationship between two analyzed items (Bakeman & Gottman, 1997), which is represented with an arrow icon. In the present study, 15 sequential behaviors were revealed to be significant based on their z-score values. The significant sequential behaviors that started with "NEXT" were NE  $\rightarrow$  PR (z-score = 22.09) and NE  $\rightarrow$  AM (z-score = 27.42). The significant sequential behavior that started with "PREV" was PR  $\rightarrow$  NE (z-score = 14.8). The significant sequential behaviors that started with "ADD MEMO" were AM  $\rightarrow$  DM (z-score = 5.5), AM  $\rightarrow$  CM (z-score = 10.23), and AM  $\rightarrow$  AH (z-score = 32.22). The significant sequential behaviors that started with "DELETE MEMO" were DM  $\rightarrow$  AM (z-score = 3.25) and DM  $\rightarrow$  DH (z-score = 25.48). The significant sequential behavior that started with "CHANGE MEMO" was CM  $\rightarrow$  AM (z-score = 23.48), AH  $\rightarrow$  DM (z-score = 3.4), and AH  $\rightarrow$  AB (z-score = 5.37). The significant sequential behavior that started with "DELETE

MARKER" was DH  $\rightarrow$  AM (z-score = 3.89). The significant sequential behaviors that started with "ADD BOOKMARK" were AB  $\rightarrow$  NE (z-score = 2.1) and AB  $\rightarrow$  AH (z-score = 2.66).

		J						
Code	NE	PR	AM	DM	CM	AH	DH	AB
NE	-24	<b>22.09</b> *	$27.42^{*}$	-5.29	-6.11	-15.1	-3.62	-2.9
PR	<b>14.8</b> *	-6.75	-1.54	-2.07	-1.62	-8.21	-1.97	-1.58
AM	-1.78	-10.62	-22.54	5.5*	10.23*	32.22*	-3.57	0.72
DM	-2.74	-1.95	3.25*	-2.21	-0.11	-4.63	25.48*	-0.89
CM	-4.29	-3.03	11.55*	-0.6	-2.07	-4.47	-1.07	-0.86
AH	23.48*	-5.57	-13.42	3.4*	-2.42	-9.49	-2.28	5.37*
DH	-1.22	-0.47	<b>3.89</b> *	-1.02	0.07	-2.15	-0.52	-0.41
AB	2.1*	-1.53	-2.81	0.3	-0.86	2.66*	-0.45	-0.36

Table 4. Adjusted residuals of BookRoll sequential behaviors for all the blended learners

*Note.* \**p* < .05.

# 4.2. Difference of the sequential patterns of e-book note-taking for SRL between blended learners in higher- and lower-achievement groups

Table 5 presents the collected frequency and percentage data of the BookRoll behaviors for higher- and lowerachievement groups. The top 3 behaviors that occurred most frequently for the higher-achievement group were "ADD MEMO" (31.79%), "NEXT" (30.93%), and "ADD MARKER" (16.61%). The top 3 behaviors that occurred most frequently for the lower-achievement group were "NEXT" (33.51%), "ADD MEMO" (28.87%), and "ADD MARKER" (14.58%). These descriptive statistics results reveal that the percentage of BookRoll behavior of page-turning (i.e., "NEXT" and "PREV") for the lower-achievement group (33.51% and 14.21%) is higher than that for the higher-achievement group (30.93% and 10.41%). Moreover, the percentage of BookRoll behavior of taking and reviewing notes (i.e., "ADD MEMO", "ADD MARKER", "ADD BOOKMARK", CHANGE MEMO, and DELETE MARKER) for the higher-achievement group (31.79%, 16.61%, 0.71%, 4.63%, and 1.07%) is higher than that for the lower-achievement group (28.87%, 14.58%, 0.6%, 3.14%, and 0.97%).

Table 5. Frequency and percentage data of the BookRoll behaviors for higher- and lower-achievement groups						
Group	Category	Frequency	Percentage (%)			
Higher-achievement group	NEXT (NE)	434	30.93			
(n = 44)	PREV (PR)	146	10.41			
	ADD MEMO (AM)	446	31.79			
	DELETE MEMO (DM)	54	3.85			
	CHANGE MEMO (CM)	65	4.63			
	ADD MARKER (AH)	233	16.61			
	DELETE MARKER (DH)	15	1.07			
	ADD BOOKMARK (AB)	10	0.71			
	DELETE BOOKMARK (DB)	0	0			
Lower-achievement group	NEXT (NE)	448	33.51			
(n = 44)	PREV (PR)	190	14.21			
	ADD MEMO (AM)	386	28.87			
	DELETE MEMO (DM)	55	4.11			
	CHANGE MEMO (CM)	42	3.14			
	ADD MARKER (AH)	195	14.58			
	DELETE MARKER (DH)	13	0.97			
	ADD BOOKMARK (AB)	8	0.6			
	DELETE BOOKMARK (DB)	0	0			

Table 6 and Table 7 are the adjusted residual tables of the BookRoll sequential behaviors for the higherachievement group and lower-achievement group, respectively. Figure 4 depicts their behavioral transition. For the higher-achievement group, 14 sequential behaviors were revealed to be significant based on their z-score values. The significant sequential behaviors that started with "NEXT" were NE  $\rightarrow$  PR (z-score = 14.83) and NE  $\rightarrow$  AM (z-score = 20.4). The significant sequential behavior that started with "PREV" was PR  $\rightarrow$  NE (z-score = 9.67). The significant sequential behaviors that started with "ADD MEMO" were AM  $\rightarrow$  DM (z-score = 3.36), AM  $\rightarrow$  CM (z-score = 8.85), and AM  $\rightarrow$  AH (z-score = 23.04). The significant sequential behavior that started with "DELETE MEMO" was DM  $\rightarrow$  DH (z-score = 19.1). The significant sequential behavior that started with "CHANGE MEMO" was CM  $\rightarrow$  AM (z-score = 10.16). The significant sequential behaviors that started with "ADD MARKER" were AH  $\rightarrow$  NE (z-score = 17.27), AH  $\rightarrow$  DM (z-score = 3.27), and AH  $\rightarrow$  AB (z-score = 5.48). The significant sequential behavior that started with "DELETE MARKER" was DH  $\rightarrow$  AM (z-score = 3.57). The significant sequential behaviors that started with "ADD BOOKMARK" were AB  $\rightarrow$  NE (z-score = 1.97) and AB  $\rightarrow$  AH (z-score = 1.97).

	<i>Table</i> 0. Adjusted residuals of bookkon sequential behaviors for higher-achievement group								
Code	NE	PR	AM	DM	CM	AH	DH	AB	
NE	-16.37	<b>14.83</b> *	$20.4^{*}$	-3.38	-5.38	-10.96	-2.6	-2.12	
PR	<b>9.67</b> *	-4.11	-0.38	-2	-1.84	-5.5	-1.31	-1.06	
AM	-1.7	-6.95	-17.04	3.36*	8.85*	23.04*	-2.7	0.85	
DM	-0.77	-1.57	0.73	-1.5	-0.35	-3.35	<b>19.1</b> *	-0.65	
CM	-3.97	-2.7	10.16*	-0.94	-1.77	-3.6	-0.86	-0.7	
AH	$17.27^{*}$	-3.68	-10.19	3.27*	-1.79	-7.24	-1.72	<b>5.48</b> *	
DH	-1.24	-0.3	3.57*	-0.73	-0.8	-1.63	-0.39	-0.32	
AB	1.97*	-1.07	-2.14	-0.64	-0.7	<b>1.97</b> *	-0.34	-0.28	

Table 6. Adjusted residuals of BookRoll sequential behaviors for higher-achievement group

*Note.* \**p* < .05.

Table 7. Adjusted residuals of BookRoll sequential behaviors for lower-achievement group

Code	NE	PR	AM	DM	CM	AH	DH	AB
NE	-17.62	<b>16.33</b> *	<b>18.4</b> *	-4.08	-3.02	-10.37	-2.52	-1.97
PR	11.09*	-5.48	-1.66	-1.1	-0.24	-6.06	-1.47	-1.15
AM	-0.74	-8	-14.84	4.45*	5.16*	$22.5^{*}$	-2.35	$2.05^{*}$
DM	-3.09	-1.23	3.88*	-1.62	-0.56	-3.19	<b>16.95</b> *	-0.61
CM	-1.87	-1.48	<b>5.73</b> *	0.2	-1.15	-2.7	-0.66	-0.51
AH	<b>16</b> *	-4.14	-8.78	1.54	-1.71	-6.17	-1.5	1.84
DH	-0.46	-0.36	1.84	-0.71	1.15	-1.4	-0.34	-0.27
AB	0.97	-1.08	-1.83	1.13	-0.51	1.79	-0.29	-0.23
4								

*Note.* \**p* < .05.

Figure 4. Behavioral transition diagrams for (a) higher-achievement group and (b) lower-achievement group



For the lower-achievement group, 11 sequential behaviors were revealed to be significant based on their z-score values. The significant sequential behaviors that started with "NEXT" were NE  $\rightarrow$  PR (z-score = 16.33) and NE  $\rightarrow$  AM (z-score = 18.4). The significant sequential behavior that started with "PREV" was PR  $\rightarrow$  NE (z-score = 11.09). The significant sequential behaviors that started with "ADD MEMO" were AM  $\rightarrow$  DM (z-score = 4.45), AM  $\rightarrow$  CM (z-score = 5.16), AM  $\rightarrow$  AH (z-score = 22.5), and AM  $\rightarrow$  AB (z-score = 2.05). The significant sequential behavior that started with "CHANGE MEMO" was CM  $\rightarrow$  AM (z-score = 16.95). The significant sequential behavior that started with "CHANGE MEMO" was CM  $\rightarrow$  AM (z-score = 5.73). The significant sequential behavior that started with "ADD MARKER" was AH  $\rightarrow$  NE (z-score = 16).

These results reveal that the BookRoll sequential behaviors indicating the consecutively and combined use of taking and reviewing notes such as  $AH \rightarrow DM$ ,  $DH \rightarrow AM$ , and  $AH \rightarrow AB$  occurred significantly only for the higher-achievement group. Moreover, the BookRoll sequential behaviors indicating the follow-up action after

the use of bookmarks such as AB  $\rightarrow$  AH and AB  $\rightarrow$  NE occurred significantly only for the higher-achievement group.

# 5. Discussion and conclusions

### 5.1. Blended learners' sequential patterns of e-book note-taking for SRL

To address the first research question, the present study used LSA to identify the learners' sequential behaviors of e-book note-taking for SRL. The analytical results of the present study are generally consistent with those of other studies (e.g., Yang et al., 2018; Zarzour et al., 2020; Zhang et al., 2021), suggesting that the use of LSA to analyze learner behavior-related data is effective in revealing, mapping, and monitoring the online learning processes of learners. Moreover, the results of the present study echo those reported by Chen et al. (2017); in contrast to the count-based measures used by other studies to quantify learner behavior in specific learning contexts, an LSA can reveal the transitional relationships between different categories of learning behaviors and, sometimes, the temporality perspective of the learning process of learners. When the learners were reading the learning materials through the e-book format, they tended to repeatedly click the NEXT button to go to the next page and the PREV button to return to the previous page. This finding is consistent with those reported in previous studies, that is, learners review previous pages frequently when they are reading e-book learning materials in sequence (Yang et al., 2021; Yin et al., 2019). Moreover, the learners exhibited a variety of the combination of note-taking and note-reviewing strategies (e.g., changing memos after adding memos, adding memos after deleting markers, deleting memos after adding markers, etc.) for the promotion of SRL. This finding is consistent with those reported by Alvi et al. (2016), that is, learners tend to use a variety of SRL techniques ranging from the shallow approach (repetition for memorization) to the cognitively deep processing approach (note-taking and consulting notes).

# 5.2. Comparisons of the sequential patterns of e-book note-taking for SRL between blended learners in higher- and lower-achievement groups

To address the second research question, the present study used LSA to uncover the differences between the blended learners in higher- and lower-achievement groups in terms of their sequential behaviors of e-book note-taking for SRL. The analytical results reveal that the percentage of BookRoll behavior of page-turning for the lower-achievement group is higher than that for the higher-achievement group. The percentage of BookRoll behavior of the taking and reviewing of notes for the higher-achievement group is higher than that for the behavior of note-taking, the analytical results reveal that the BookRoll sequential behaviors indicating the consecutively and combined use of taking and reviewing notes (e.g., adding memos after deleting markers, deleting memos after adding markers, etc) occurred more for the higher-achievement group. Moreover, the BookRoll sequential behaviors indicating the follow-up action after the use of bookmarks (e.g., adding markers after adding bookmarks, turning to the next page after adding bookmarks, etc) occurred more for the higher-achievement group compared with the lower-achievement group.

The aforementioned findings are consistent with those reported by Alvi et al. (2016) and Effeney et al. (2013), that is, high-achievement learners tend to exhibit greater engagement in using wider cognitive strategies (e.g., note-taking and note-reviewing) than low-achievement learners. The findings are also consistent with those of other empirical studies, suggesting that learners who engage in the use of note-taking functions outperform those who do not (Kiewra et al., 1989; Kiewra et al., 1991) in terms of their learning achievements. In a specific context such as BL, the findings are consistent with those reported by Yang et al. (2021), suggesting that the taking and reviewing of notes have a considerable effect on the learning achievements of learners, and the act of browsing without taking notes is associated with poor learning achievements.

### 5.3. Theoretical and practical implications

The emergence of advanced educational technologies for classroom environments is changing the way learners take digital notes and process the knowledge that they acquire during class. The accountability of learners to the management of their learning processes may increase, which changes their study methods (Stacy & Cain, 2015). The appropriate adoption of note-taking strategies improves learners' capabilities of memorizing information and helps them to perform better on tests (Peverly et al., 2003).

The analytical results of the present study have several implications. First, the findings suggest that, in addition to the traditional navigation functions that allow learners to browse e-book learning materials, note-taking features (including the creation, deletion, and revision of memos, markers, and bookmarks) were identified as key behaviors for cognitive strategy use of SRL in BL contexts. Second, the present study demonstrated that LSA can be used to analyze sequential behavioral patterns to generate findings that enable e-book developers and instructional designers to better understand the actual cognitive operations and behavioral patterns of learners when using e-books. Moreover, LSA enables instructors and researchers to explore the hidden behaviors of learners and develop an effective instructional mechanism for the self-regulatory use of e-books in BL contexts through a visualized transition diagram.

For course scenarios, the findings suggest that teachers can guide and encourage learners with low levels of engagement in note-taking to apply strategies for combined use of taking and reviewing notes that improve their engagement level and learning achievements in BL. The findings also suggest that learners can increase their interaction with e-books by using more note-taking and note-review features to enhance their retention of the information in their learning materials. From this perspective, educational tools that allow learners to monitor and diagnose their learning process and receive personalized feedback on how they can improve their cognitive and metacognitive strategies for self-regulation abilities can be helpful; this issue was also highlighted in Yang and Ogata (2022).

In summary, the present study conducted an exploratory study in an undergraduate course that implemented the BL design. The learning task for the blended learners in the present study was to study the learning material using an e-book system during in-class and out-of-class learning sessions. The present study employed an LSA to investigate the blended learners' sequential patterns of e-book note-taking for the cognitive strategy use of SRL. Moreover, the present study revealed the difference between higher- and lower-achievement blended learners in terms of their sequential behaviors of e-book note-taking. The major contribution of the present study is to offer opportunities for educators to effectively understand learners' interactions with note-taking and note-reviewing systems (e.g., e-books) and provide learners with evidence-based educational feedback and corresponding interventions regarding the combined use of note-taking strategies for the promotion of SRL and improvement of course achievement in BL. Teachers at every education level can use the findings of the present study as a basis for adjusting their teaching strategies or materials to achieve personalized learning for their courses. The findings can be applied to help learners to adjust their adopted learning strategies, such that they can better adapt to changing learning environments and learning goals when they are receiving information from educators or digital learning platforms in the context of BL; this issue was also highlighted in the literature (Kundu et al., 2021; Luan & Tsai, 2021; Yang et al., 2021).

### 5.4. Limitations

The present study has several limitations. First, a sample size of only 88 participants was used in the present exploratory study. Therefore, the results, although significant, cannot be generalized to larger populations. A general analytical model is required to examine a larger sample size of learners through the application of similar analytics methods. Second, since the present study focused exclusively on identifying the sequential patterns of blended learners by examining e-book learning logs relating to note-taking, the number of types of learner learning behaviors and the awareness of SRL that could be identified were relatively limited. Therefore, future studies that apply similar analytic methods should integrate a greater variety of digital learning platforms and questionnaires to obtain a greater range of learner data relating to e-book learning and the awareness of SRL. Third, in the present study, GSEQ-based LSA was applied individually to identify the blended learners' sequential patterns of e-book note-taking. Future studies should incorporate other techniques (e.g., clustering and process mining) to enrich their analytical process and obtain further insights into the sequential patterns of e-book note-taking of learners undergoing BL. Finally, we did not take into account the influences of the blended learners' learning styles or personality traits on their behavioral engagement of e-book note-taking or learning achievement before the present exploratory study, which could cause some bias in the analytical results. When similar analyses are conducted in future studies, these potential influences should also be taken into account.

### Acknowledgement

This work was partially supported by JSPS Grant-in-Aid for Scientific Research (S)16H06304 and NEDO Special Innovation Program on AI and Big Data 18102059-0 and Explainable AI P20006.

### References

Álvarez, A., Martín, M., Fernández-Castro, I., & Urretavizcaya, M. (2013). Blending traditional teaching methods with learning environments: Experience, cyclical evaluation process and impact with MAgAdI. *Computers and Education, 68*, 129–140. https://doi.org/10.1016/j.compedu.2013.05.006

Alvi, E., Iqbal, Z., Masood, F., & Batool, T. (2016). A Qualitative account of the nature and use of self-regulated learning (SRL) strategies employed by university students. *Australian Journal of Teacher Education (Online)*, 41(8), 40-59.

Anthonysamy, L., Koo, A-C., Hew, S.-H. (2020). Self-regulated learning strategies and non-academic outcomes in higher education blended learning environments: A One decade review. *Education and Information Technologies*, 1–28. https://doi.org/10.1007/s10639-020-10134-2.

Azevedo, R., Moos, D. C., Johnson, A. M., & Chauncey, A. D. (2010). Measuring cognitive and metacognitive regulatory processes during hypermedia learning: Issues and challenges. *Educational Psychologist*, 45(4), 210-223. https://doi.org/10.1080/00461520.2010.515934

Bakeman, R., & Gottman, J. M. (1997). Observing interaction: An Introduction to sequential analysis. Cambridge university press.

Bakeman, R., & Quera, V. (1995). Analyzing interaction: Sequential analysis with SDIS and GSEQ. Cambridge University Press.

Boyle, J. R. (2010). Note-taking skills of middle school students with and without learning disabilities. *Journal of Learning Disabilities*, 43(6), 530-540. https://doi.org/10.1177/0022219410371679

Casselden, B., & Pears, R. (2020). Higher education student pathways to ebook usage and engagement, and understanding: highways and cul de sacs. *Journal of Librarianship and Information Science*, 52(2), 601-619.

Castro, R. (2019). Blended learning in higher education: Trends and capabilities. *Education and Information Technologies*, 24(4), 2523-2546.

Cengiz-Istanbullu, B., & Sakiz, G. (2022). Self-regulated learning strategies impact fourth-grade students' positive outcomes in science class. *Journal of Baltic Science Education*, *21*(2), 192-206. https://doi.org/10.33225/jbse/22.21.192

Chen, B., Resendes, M., Chai, C. S., & Hong, H. Y. (2017). Two tales of time: Uncovering the significance of sequential patterns among contribution types in knowledge-building discourse. *Interactive Learning Environments*, 25(2), 162-175.

Chiu, C. H., Wu, C. Y., & Cheng, H. W. (2013). Integrating reviewing strategies into shared electronic note-taking: Questioning, summarizing and note reading. *Computers & Education*, 67, 229-238.

Cortina, J. M. (1993). What is coefficient alpha? An Examination of theory and applications. *Journal of applied psychology*, 78(1), 98.

Dror, I. E. (2008). Technology enhanced learning: The Good, the bad, and the ugly. Pragmatics & Cognition, 16(2), 215-223.

Du, M. C. (2004). *Personalized annotation management for web based learning service* (Unpublished master thesis). National Central University, Taiwan.

Dunn, K. (2015). The Challenges of launching a MOOC and reusing that material in a blended campus class. In *Proceeding* of 2015 Canadian Engineering Education Association (CEEA15) Conference (pp. 1–8). McMaster University.

Effeney, G., Carroll, A., & Bahr, N. (2013). Self-regulated learning: Key strategies and their sources in a sample of adolescent males. *Australian Journal of Educational & Developmental Psychology*, 13, 58-74.

Fisher, J. L., & Harris, M. B. (1973). Effect of note taking and review on recall. *Journal of Educational Psychology*, 65(3), 321–325. https://doi.org/10.1037/h0035640

Garrison, D. R., & Vaughan, N. D. (2008). Blended learning in higher education: Framework, principles, and guidelines. John Wiley & Sons.

Hanafin, J., Shevlin, M., Kenny, M., & Neela, E. M. (2007). Including young people with disabilities: Assessment challenges in higher education. *Higher education*, *54*(3), 435-448.

Huang, Y. M., & Liang, T. H. (2015). A Technique for tracking the reading rate to identify the e-book reading behaviors and comprehension outcomes of elementary school students. *British Journal of Educational Technology*, 46(4), 864–876.

Igo, L. B., & Kiewra, K. A. (2007). How do high-achieving students approach web-based, copy and paste note taking? Selective pasting and related learning outcomes. *Journal of Advanced Academics*, *18*(4), 512–529.

Igo, L. B., Bruning, R., & McCrudden, M. T. (2005). Exploring differences in students' copy-and-paste decision making and processing: A Mixed-methods study. *Journal of Educational Psychology*, 97(1), 103–116.

Jovanović, J., Gašević, D., Dawson, S., Pardo, A., & Mirriahi, N. (2017). Learning analytics to unveil learning strategies in a flipped classroom. *The Internet and Higher Education*, 33(4), 74-85.

Kane, M. J., Smeekens, B. A., von Bastian, C. C., Lurquin, J. H., Carruth, N. P., & Miyake, A. (2017). A Combined experimental and individual-differences investigation into mind wandering during a video lecture. *Journal of Experimental Psychology: General*, 146(11), 1649–1674.

Kauffman, D. F., Ge, X., Xie, K., & Chen, H. (2008). Prompting in web-based environments: Supporting self-monitoring and problem solving skills in college students. *Journal of Educational Computing Research*, 38(2), 115–137.

Kauffman, D. F., Zhao, R., & Yang, Y. S. (2011). Effects of online note taking formats and self-monitoring prompts on learning from online text: Using technology to enhance self-regulated learning. *Contemporary Educational Psychology*, *36*(4), 313-322.

Kiewra, K. A., DuBois, N. F., Christian, D., McShane, A., Meyerhoffer, M., & Roskelley, D. (1991). Note-taking functions and techniques. *Journal of Educational Psychology*, 83(2), 240–245. http://dx.doi.org/10.1037/0022-0663.83.2.240

Kiewra, K. A., Dubois, N., Christensen, M., Kim, S.-I., & Lindberg, N. (1989). A More equitable account of the note-taking functions in learning from lecture and from text. *Instructional Science*, *18*(3), 217–232. http://dx.doi.org/10.1007/bf00053360

Kundu, A., Bej, T., & Rice, M. (2021). Time to engage: Implementing math and literacy blended learning routines in an Indian elementary classroom. *Education and Information Technologies*, 26(1), 1201-1220.

Lin, P. H., Huang, Y. M., & Chen, C. C. (2018). Exploring imaginative capability and learning motivation difference through picture E-book. *IEEE Access*, *6*, 63416–63425.

Luan, H., & Tsai, C. C. (2021). A Review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1), 250-266.

Luo, L., Kiewra, K. A., Flanigan, A. E., & Peteranetz, M. S. (2018). Laptop versus longhand note taking: Effects on lecture notes and achievement. *Instructional Science*, 46(6), 947–971. https://doi.org/10.1007/s11251-018-9458-0

Míguez-Álvarez, C., Crespo, B., Arce, E., Cuevas, M., & Regueiro, A. (2020). Blending learning as an approach in teaching sustainability. *Interactive Learning Environments*, 1-16. https://doi.org/10.1080/10494820.2020.1734623

Moon, H., & Hyun, H. S. (2019). Nursing students' knowledge, attitude, self-efficacy in blended learning of cardiopulmonary resuscitation: A Randomized controlled trial. *BMC Medical Education*, 19(1), 1–8. https://doi.org/10.1186/s12909-019-1848-8

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The Role of prior domain knowledge. *Contemporary Educational Psychology*, 33(2), 270-298.

Nakayama, M., Mutsuura, K., & Yamamoto, H. (2021). Impact of learner's characteristics and learning behaviour on learning performance during a fully online course. In *Note Taking Activities in E-Learning Environments* (pp. 15-36). Springer.

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). Asia-Pacific Society for Computers in Education.

O'Hara. (2005). Taking notes. In Improving your study skills: Study smart, study less (pp. 57 -70). Wiley.

Peverly, S. T., & Wolf, A. D. (2019). Note-taking. In J. Dunlosky & K. A. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 320–355). Cambridge University Press.

Peverly, S. T., Brobst, K. E., Graham, M., & Shaw, R. (2003). College adults are not good at self-regulation: A Study on the relationship of self-regulation, note taking, and test taking. *Journal of Educational Psychology*, 95(2), 335-346. http://dx.doi.org/10.1037/0022-0663.95.2.335

Pintrich, P.R., Smith, D.F., Garcia, T., & McKeachie, W. (1991). A Manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ). http://files.eric.ed.gov/fulltext/ED338122.pdf

Pressley, M., & Woloshyn, V. (1995). Cognitive strategy instruction that really improves children's academic performance (2nd ed.). Brookline.

Sackett, G. P. (1978). Observing behaviour: Theory and applications in mental retardation (Vol. 1). University Park Press.

Song, D., Hong H., & Oh, E. Y. (2021). Applying computational analysis of novice learners' computer programming patterns to reveal self-regulated learning, computational thinking, and learning performance. *Computers in Human Behavior, 120*, 106746. https://doi.org/10.1016/j.chb.2021.106746

Stacy, E. M., & Cain, J. (2015). Note-taking and handouts in the digital age. American Journal of Pharmaceutical Education, 79(7), 107. https://doi.org/10.5688/ajpe797107

Sung, T. W., & Wu, T. T. (2017). Dynamic e-book guidance system for English reading with learning portfolio analysis. *The Electronic Library*, 35(2), 358–373.

Tang, C. M., & Chaw, L. Y. (2016). Digital literacy: A Prerequisite for effective learning in a blended learning environment? *The Electronic Journal of e-Learning*, 14(1), 54–65.

Trevors, G., Duffy, M., & Azevedo, R. (2014). Note-taking within MetaTutor: Interactions between an intelligent tutoring system and prior knowledge on note-taking and learning. *Educational Technology Research and Development*, 62(5), 507-528.

van de Sande, C., Abramson, J., & Judson-Garcia, J. (2017). An Exploration of note-taking in an online calculus course. *Journal of Computers in Mathematics and Science Teaching*, 36(1), 75-99.

Wright, S., Fugett, A., & Caputa, F. (2013). Using e-readers and internet resources to support comprehension. *Educational Technology & Society*, *16*(1), 367-379.

Wu, J. Y. (2020). The Predictive validities of individual working-memory capacity profiles and note-taking strategies on online search performance. *Journal of Computer Assisted Learning*, *36*(6), 876–889. https://doi.org/10.1111/jcal.12441

Wu, T. T. (2016). A Learning log analysis of an English-reading e-book system combined with a guidance mechanism. *Interactive Learning Environments*, 24(8), 1938–1956.

Yang, C. C., & Ogata, H. (2022). Personalized learning analytics intervention approach for enhancing student learning achievement and behavioral engagement in blended learning. *Education and Information Technologies*, 1-20. https://doi.org/10.1007/s10639-022-11291-2

Yang, C. C., Chen, I. Y., & Ogata, H. (2021). Toward precision education: Educational data mining and learning analytics for identifying students' learning patterns with ebook systems. *Educational Technology & Society*, 24(1), 152-163.

Yang, X., Li, J., & Xing, B. (2018). Behavioral patterns of knowledge construction in online cooperative translation activities. *The Internet and Higher Education*, *36*, 13-21.

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.

Zarzour, H., Bendjaballah, S., & Harirche, H. (2020). Exploring the behavioral patterns of students learning with a Facebookbased e-book approach. *Computers & Education*, *156*, 103957. https://doi.org/10.1016/j.compedu.2020.103957

Zhang, J., Gao, M., Holmes, W., Mavrikis, M., & Ma, N. (2021). Interaction patterns in exploratory learning environments for mathematics: A Sequential analysis of feedback and external representations in Chinese schools. *Interactive Learning Environments*, 29(7), 1211-1228.

Zimmerman, B. J. (1989). A Social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81(3), 329-339. https://doi.org/10.1037/0022-0663.81.3.329

Zimmerman, B. J. (2000). Attaining self-regulation: A Social cognitive perspective. In *Handbook of self-regulation* (pp. 13-39). Academic Press.

# Effects of Undergraduate Student Reviewers' Ability on Comments Provided, Reviewing Behavior, and Performance in an Online Video Peer Assessment Activity

# Liang-Yi Li<sup>1\*</sup> and Wen-Lung Huang<sup>2</sup>

<sup>1</sup>Program of Learning Sciences, Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taiwan // <sup>2</sup>Department of Communication, Fo Guang University, Yilan, Taiwan // lihenry12345@ntnu.edu.tw // wlhuang@mail.fgu.edu.tw

\*Corresponding author

(Submitted January 28, 2022; Revised August 16, 2022; Accepted Aug 29, 2022)

**ABSTRACT:** With the increasing bandwidth, videos have been gradually used as submissions for online peer assessment activities. However, their transient nature imposes a high cognitive load on students, particularly low-ability students. Therefore, reviewers' ability is a key factor that may affect the reviewing process and performance in an online video peer assessment activity. This study examined how reviewers' ability affected the comments they provided and their reviewing behaviors and performance. Thirty-eight first-year undergraduate students participated in an online video peer assessment activity for 3 weeks. This study analyzed data collected from the teacher's and peer reviewers' ratings, comments provided by peer reviewers, and system logs. Several findings are significant. First, low-ability reviewers preferred to rate higher scores than high-ability reviewers did. Second, low-ability reviewers had higher review errors than high-ability reviewers. Third, high-ability reviewers provided more high-level comments, while low-ability reviewers provided more low-level comments. Finally, low- and high-ability reviewers showed different behavior patterns when reviewing peers' videos. In particular, low-ability reviewers invested more time and effort in understanding video content, while high-ability reviewers invested more time and effort in understanding video content, while high-ability reviewers invested more time and effort in understanding video content, while high-ability reviewers invested more time and effort in diagnosing problems. These findings are discussed, and several suggestions for improving the instructional and system design of online video peer assessment activities are provided.

Keywords: Video peer assessment, Learning analytics, Comments provided, Behavior pattern

# 1. Introduction

Peer assessment (PA) is a process whereby students assign grades to peers' submissions and provide comments for peers to improve their work (Tenorio et al., 2016). PA can reduce the teacher's workload and improve students' attitudes, critical thinking, and judgement skills (Tenorio et al., 2016), and has already been applied in many disciplines such as science, language, and programming.

With the rapid development of information and communication technologies (ICT), web-based peer assessment approaches have been widely used (Formanek et al., 2017; Hsia et al., 2016a; Hsu et al., 2018). They can help teachers to share the assessment tasks and results and monitor students' progress (Lin et al., 2001). Students can also conduct peer assessment activities on the Web without the limitations of time and space.

Generally, written text is the primary target assessed in online peer assessment activities (Tenorio et al., 2016). However, with the increasing bandwidth of the internet, videos have been gradually used as submissions. In contrast to static text and images, videos provide several advantages for peer assessment activities. First, videos, which present dynamic visual and verbal information, are especially useful for evaluating actions and voices (Hsia et al., 2016b; Lai et al., 2020). Second, video playing interfaces provide multiple operations (e.g., resume, pause, fast forward, and back). Students' operations can be recorded in system logs when reviewing peers' videos. These logs can then be analyzed in order to understand how students review peers' videos (Li, 2019).

Although online video peer assessment provides these advantages, not all students can benefit from it because reviewing peers' assignments on the Web is a self-regulated process in which learners freely control their reviewing path and pace. Students with different individual characteristics may have different behavior patterns when reviewing peers' assignments. These different behavior patterns may also result in different outcomes (Shirvani Boroujeni & Dillenbourg, 2019). To support students with different individual characteristics, teachers should understand how students perform during the Web learning process and what the relationships are between students' individual characteristics and learning behaviors and performance.

Reviewers' ability, which is defined as task-related knowledge and skills, is one individual difference that is often examined in peer assessment activities. Previous studies have shown that reviewers' ability affects the quantity and quality of comments they provide for peers' submissions (Huisman et al., 2018; Patchan et al., 2013). However, these studies primarily examined written text (e.g., writing compositions). Because videos are transient media in which the information presented changes dynamically, students easily experience the problems of cognitive overload and disorientation when viewing peers' videos. Therefore, whether the same effects also happen when the submissions are videos is unknown. In addition, previous studies have found that individual differences such as prior knowledge and cognitive style can affect learners' video watching behaviors and performance when viewing instructional videos (de Boer et al., 2016; de Boer et al., 2011; Li, 2019). Whether students' ability also affects their reviewing behaviors and performance is also unknown. Therefore, the purpose of this study was to examine how reviewers' abilities affect the comments they provide and their reviewing behaviors (e.g., review error and the quantity and quality of the comments provided).

# 2. Related works

This section first reviews previous studies related to the use of videos in online peer-assessment activities. The review focuses on what has been done about online video peer-assessment activities. Next, this section presents studies that examined the effects of reviewers' ability on the quantity and quality of comments and review errors when the submissions were static documents. These are the variables examined in this study. Because we used several learning analytics techniques to explore reviewers' behavioral patterns, we then introduced how the learning analytics community has studied peer assessment. Finally, a theoretical framework was proposed to present the relationships between the variables examined in this study. Based on the review, we then proposed the research questions of this study.

### 2.1. Use of videos in online peer-assessment activities

Videos have been used as submissions in PA activities for more than 30 years. Because of the limitation of internet bandwidth, the delivery of videos in the early period was via videotape, CD, or USB. Teachers needed to make additional efforts to collect and share students' videos. With the increasing bandwidth in the recent decade, however, videos have been gradually used as submissions in online peer assessment activities. They can present dynamic visual and verbal information. Therefore, they are especially useful for evaluating actions and voices, and have been used in sport (Hsia et al., 2016b), communication skills (Lai, 2016; Lai et al., 2020), and presentations (Wu & Kao, 2008).

Studies have examined the effects of online video peer assessment activities and found that such activities can improve students' learning performance and satisfaction (Hsia et al., 2016a; Lai, 2016; Wu & Kao, 2008). For example, Hsia et al. (2016a) examined the effects of the web-based peer assessment approach on students' learning performance, self-efficacy, and satisfaction in a junior high school performing arts course. They found that, in comparison with the web-based streaming video-supported learning approach, the web-based peer assessment approach could significantly improve the students' performance and learning satisfaction. Lai (2016) implemented an online video peer assessment system for scaffolding students' communication skills. They found that students' communication performance was significantly improved. The students were satisfied with the online peer assessment learning activities.

In addition to examining the effects of the online video assessment approach, studies have developed systems and instructional approaches for supporting online video peer assessment activities (Lai et al., 2020; Lin et al., 2021; Wu & Kao, 2008). For example, Lai et al. (2020) developed a video annotation system that helped students comment on any video position. They examined the effects of the system on students' communication skills and professional attitudes during an online peer assessment activity. They found that the video annotation system was helpful for promoting students' development of communication skills, but not their professional attitudes. The students using the video-annotation tool provided more suggestion comments than those who did not use it. They concluded that the video system with the annotation function was better than the video system without the function. Lin et al. (2021) proposed an online interactive peer assessment approach with an online video peer assessment to compare the approach with a one-way peer assessment approach and found that the proposed approach demonstrated significantly better learning achievement.

In sum, previous studies primarily examined the effects of online video peer assessment on performance, students' attitudes, and motivation. They found that online video peer assessment activities can improve learning performance and satisfaction. They also developed systems and instructional approaches for improving video peer assessment, and examined their effects. These developed systems and instructional approaches can provide ideas for system and instructional designers to improve online video peer assessment activities.

### 2.2. Peer assessment and reviewers' ability

Reviewing peers' assignments is a complex process. It consists of two intertwined tasks, providing feedback and rating. Regarding providing feedback, it involves the steps of reading, problem detection, and problem diagnosis (Patchan & Schunn, 2016). Regarding rating, it involves the steps of reading and understanding with concurrent evaluation, articulating scoring decisions, and making scoring decisions (Crisp, 2010; Cumming et al., 2002). Reviewers' ability is a key factor that may affect the reviewing process and performance.

Reviewers' ability was defined as task-related knowledge and skills (e.g., essay writing) in previous studies (Huisman et al., 2017; Xiong & Schunn, 2021). Generally, they determined reviewers' ability by a test that measured task-related knowledge and skills (Huisman et al., 2018; Patchan et al., 2013; Patchan & Schunn, 2015; Patchan & Schunn, 2016) or the quality of students' submissions (Huisman et al., 2017; Wang et al., 2016; Xiong & Schunn, 2021). For example, Patchan et al. (2013) determined the ability level of the students based on self-reported SAT verbal scores. Xiong and Schunn (2021) measured reviewers' ability by the writing quality of the submitted documents, which were evaluated by two experts. In this study, we measured reviewers' ability by the quality of their submissions, because it is most relevant to the current reviewing task (Xiong & Schunn, 2021).

When reviewing a peer's video, reviewers first watch the video. Because video is a transient medium in which the content is dynamically changing, it imposes a high cognitive load on learners (Li, 2019; Mayer, 2002). This high cognitive load may be more suitable for high-ability reviewers, because they have more prior knowledge and reserve more cognitive resources to handle the cognitive load (Moos & Azevedo, 2008; Song et al., 2016). While watching, reviewers have to concurrently detect problems. Reviewers compare their prior knowledge and the watched content to detect the problems. Because high-ability reviewers have richer knowledge of each type of problem, they should be able to easily detect problems and provide more comments (Patchan & Schunn, 2016). After detecting a problem, reviewers must provide enough information for authors to revise their submissions. A diagnosis can vary in its degree of explicitness. Providing suggestions can be seen as a more explicit diagnosis than identifying problems (Wu & Schunn, 2020). High-ability reviewers who have more knowledge of the subject and problems should be able to provide more elaborate diagnoses (Patchan & Schunn, 2015). Finally, reviewers have multiple considerations for making a final decision. The detected problems are the primary source. High-ability reviewers can effectively detect and diagnose problems. Therefore, they should be able to make more correct decisions (Xiong & Schunn, 2021).

Studies examining the effects of reviewers' ability on peer assessment activities are rare. They primarily examined how reviewers' ability affected the quantity and quality of comments that peer reviewers provide for their peers' submissions (Huisman et al., 2018; Huisman et al., 2017; Patchan et al., 2013; Patchan & Schunn, 2015; Patchan & Schunn, 2016; Xiong & Schunn, 2021). However, their results were mixed (Huisman et al., 2017; Patchan & Schunn, 2015; Patchan & Schunn, 2016). Several studies have found that reviewers' ability did not affect the quantity of comments provided. For example, Patchan and Schunn (2016) found that the number of comments of high- and low-ability reviewers was not significantly different. However, high-ability reviewers provided more high-level comments than low-ability reviewers provided. Huisman et al. (2017) also found that reviewer ability did not affect the provided feedback quantity. However, higher ability reviewers provided more suggestions and explanatory feedback than low-ability reviewers.

Patchan et al. (2013) found that high-ability reviewers provided more comments than low-ability reviewers. They examined how ability pairing (e.g., a high-ability reviewer with a high-ability author) affected the quantity and quality of comments. They found that high-ability reviewers provided more feedback, and their feedback was more likely to be implemented than that of low-ability reviewers. In particular, high-ability reviewers provided more problems, low prose issues, and substance issues for low-ability writers than low-ability reviewers on high-ability reviewers provided more positively emotional comments than high-ability reviewers on high-ability submissions. Although the experimental results obtained in these studies differed slightly, they reported one consistent result, namely, that high-ability reviewers provided more high-level feedback than low-ability reviewers.

In addition to the comments provided, we found only one study that examined the relationship between reviewers' ability and review error. Xiong and Schunn (2021) examined the relationships between the factors related to reviewer, essay, and reviewing process and whether the factors could predict two types of review errors: severity and leniency. They defined review error as the discrepancy between peer reviews and expert reviews. Review errors were calculated using the difference between students' ratings and expert ratings on a given essay. Review errors were further categorized as severe and lenient. Their study found that reviewers' ability could predict severe errors, but could not predict lenient errors. In particular, reviewers' ability was found to be negatively related to severe errors, and lower ability reviewers were more likely to produce severe ratings. These results indicated that reviewers' ability can significantly affect review error. In our study, the definition and measurement of reviewer error is the same as the definition and measurement used in Xiong and Schunn (2021).

### 2.3. Peer assessment in learning analytics

Learning Analytics (LA) is a field that offers tools and techniques to analyze educational data in order to understand the process of learning and improve the education environment. Previous studies in applying learning analytics to support peer assessment activities have focused on several areas, such as learning analytics dashboards (Er et al., 2021), automatic feedback (Cavalcanti et al., 2021; Shibani et al., 2019; Shibani et al., 2022), automatically classifying reviewers' comments (Dood et al., 2022), and predicting review errors (Xiong & Schunn, 2021). For example, Er et al. (2021) proposed a theoretical framework of collaborative peer feedback and designed a learning analytics dashboard based on the framework. The dashboard, which provides an overview of participation in assessments, class-wide statistics about feedback, and an overview of several engagement indicators, aims to support instructor actions for pedagogical decisions in a peer assessment activity. Shibani et al. (2022) introduced a writing analytics tool which used natural language processing to automatically identify rhetorically salient structures in writing. The tool can then provide contextualized automated writing feedback for students' assignments. Students revised their assignments based on both automated and peer feedback.

In addition to supporting peer assessment activities, studies have applied LA techniques and tools to explore learners' behavior patterns (Er et al., 2021; Hsu et al., 2018) for peer assessment activities. Clustering analysis and sequential behavior analysis are frequently used techniques. Clustering analysis was commonly used for exploring unanticipated trends or patterns (Cerezo et al., 2016; Li & Tsai, 2017). For example, Mirriahi et al. (2016) used clustering analysis on the behavioral variables (e.g., number of annotations, video watching time, and number of pauses) of a video annotation tool used for a video peer assessment activity. They found that students' viewing behaviors showed great variety and were clustered into four behavior patterns: minimalists, task-oriented, disenchanted, and intensive users. They then found that these behavior patterns were affected by external factors (e.g., grading). Sequential behavior analysis was used for exploring the behavior transitions (Li et al., 2022; Zarzour et al., 2020). For example, Chen et al. (2020) used sequential analysis to explore students' behavioral sequences in three online video peer assessment activities: comment only, scoring only, and comment with scoring. They then compared the differences in students' behavioral patterns among the three activities. They found that the students in the comment with scoring group had better musical theater performance, provided more critical feedback, and performed more behaviors of reading the rubrics, watching example videos, watching peers' work, and reading peers' feedback.

### 2.4. Theoretical framework

According to the above discussions, this study aimed to examine the effects of reviewers' ability on reviewing process and performance. The Presage-Process-Product (3P) model (Biggs, 1987) was applied as a theoretical framework. This model identifies three sections: presage, process, and product. The presage section considers pre-existing individual characteristics (e.g., gender, ability, and prior knowledge) and contextual issues (e.g., learning activities, instructor effects, and learning systems). The process variables are the ways in which learners handle their learning tasks. They are the results of the interaction between individual characteristics and contextual factors. Because learners with different individual characteristics have different perceptions of their contexts, these perceptions affect their choices regarding learning behaviors and strategies. Finally, the product section includes the learning outcomes of each learner (Cybinski & Selvanathan, 2005). In this study, the presage factor is reviewers' ability; the product factors are the quantity and quality of comments provided and the reviewing error. The process factor is the behavioral pattern acquired by analyzing the system logs. The behavioral analysis may clarify the role of reviewers' ability in peer assessment activities (Chen et al., 2020; Topping, 1998). Based on this framework, this study aimed to answer the following four research questions.

- Did the reviewers' ability affect the scores they gave?
- Did the reviewers' ability affect their review error?
- Did the reviewers' ability affect the quantity and quality of the comments they provided?
- Did the reviewers' ability affect their behavioral patterns?

This study focuses on the effects of reviewers' ability on rating error, comments provided, and behavior patterns in a video peer assessment activity. There are three reasons for this focus. First, videos have been gradually used as submissions in an online peer assessment activity. However, their transient nature imposes a high cognitive load on students, particularly for low-ability students. Reviewers' ability is a key factor that may affect the reviewing process and performance. However, we have not found any study that has examined the effects of reviewers' ability in video peer assessment activities. Second, previous studies that examined the effects of reviewers' ability on the quantity and quality of comments provided revealed mixed results. In addition, we found only one study that examined the relationship between reviewers' ability level and review error. Therefore, more research should be conducted to provide more empirical findings. Third, because reviewers can freely control their pace and path when reviewing peers' videos in online peer assessment systems, their reviewing outcomes, such as comments provided and scores rated, are influenced by a range of factors. Understanding how reviewers' ability affects their reviewing process and performance can help instructional and system designers to improve the system and instructional design and to design personalized supports for reviewers with different ability levels (Li & Tsai, 2020; Wang et al., 2016).

# 3. Method

### **3.1.** Participants and course

This study was conducted by a quasi-experimental design. A total of 38 first-year undergraduate students (20 males and 18 females) participated in this study. They were film design majors enrolled in a one-semester course called digital editing at a university of northern Taiwan. They attended face-to-face classes, where the course teacher introduced storytelling, digital editing skills, and film editing software for 2 hours each week in a computerized classroom, in which each student used one computer with internet access. In addition to lectures in the classroom, the course teacher published peer assessment assignments on a video peer assessment system. To meet the ethical requirements, before conducting the peer assessment activity, the students were informed of the purposes of the study and read the consent letter to confirm their rights in this study. The students who had signed the consent letter were involved in the study.

### **3.2. Video peer assessment system**

The video peer assessment system is a subsystem of a learning management system (LMS). It consists of three components: submitting, reviewing, and sharing. Teachers can create a video assignment using the submitting component. Each video assignment is presented on a submitting page where the students can upload their videos.

Teachers can use the reviewing component to create a reviewing assignment. When creating a reviewing assignment, teachers have to select a video assignment and an evaluation rubric. The rubric, which the teacher previously created using the LMS, was used by the reviewers for evaluating the assigned videos. This provides flexibility that allows teachers to assign different rubrics for different video assignments. The system then randomly assigns two peers as reviewers for each submitted video and automatically creates a reviewing page for each reviewer to review peers' videos.

The review was anonymous. When reviewing the assigned videos, a reviewer can link to the reviewing page. Figure 1 is the reviewing page. The page presents a rubric link, which is associated with a rubric page, and a video link, which is associated with a video page, for each assigned video. On the rubric page, the evaluation rubric, which was selected by the teacher when created a reviewing assignment, is presented. A reviewer can evaluate the assigned video by the rubric. On the video page, a video annotation interface is presented (Figure 2), where the reviewer can view the assigned video and comment on any position of the video timeline.

This video annotation interface allows reviewers to add a comment at any position of the video timeline. To add a comment at a specific position, a reviewer first drags the timeline to the position. Next, he/she clicks the right mouse button and then a menu with an "Insert a comment" button is displayed. The reviewer clicks the button

and then a dialog is immediately presented. The reviewer can type his/her comments into the dialog and click the submit button; then a comment tag (red rectangle) is immediately added at that position.





Figure 2. The video annotation interface



One feature of the video annotation interface is in-context comments. The interface associated the comments and timeline, so users can easily identify how many comments have been created and where they are, and quickly view the comments and the associated video content.

Finally, teachers can share the results of a reviewing assignment by the sharing component. When a reviewing assignment is shared, a sharing page is immediately generated. The page lists all students' videos with reviewing results, including links to the corresponding rubric pages and video pages. Authors can view the reviewing results of their videos to revise their submitted videos accordingly and learn from peers' videos and reviewing results.

### **3.3. Procedure**

In the first 5 weeks of the course, teachers not only introduced the concepts, skills, and software of film editing, but also used one or two examples, which were the videos submitted by the students of the previous year, to teach the students how to rate the videos, how to provide comment for the videos, and how to use the video peer assessment system in each week. A peer assessment activity was implemented during week 6 to week 8 of the course. The 1st (week 6), 2nd (week 7), and 3rd (week 8) weeks were for submitting, reviewing, and revising, respectively. At the beginning of the 1st week, the course teacher published a video assignment requiring students to edit a video and submit it to the system within 1 week.

At the beginning of the 2nd week, the teacher published a reviewing assignment where each student was assigned two peers' videos for reviewing. The students were required to finish their reviews within 1 week. At the beginning of the 3rd week, the teacher shared the reviewing results. He also published a video assignment which required the students to submit their revised video and a document on which the students responded to the peers' comments before the end of the 3rd week. A student was rewarded with a 16%, 8%, and 8% portion of the final grade for the quality of the submitted video, the quality of the comments provided for peers' videos, and the quality of the revised video, respectively.

### **3.4.** The evaluation rubric

The rubric used for evaluating the students' videos was designed by the course teacher and a film editing expert who had taught film editing for 3 years. The course teacher collected three evaluation rubrics used in video editing competitions and discussed them with the expert to determine the dimensions, detailed descriptions of the dimensions, and the rating scheme. There are three dimensions, namely rhythm, creativeness, and technical skill. The raters gave a score of 1 to 5 to every submitted video on each dimension. A higher score represents higher video quality. The detailed descriptions of the rating scheme are listed in Table 1.

<i>Table 1</i> . The detailed descriptions of the evaluation rubric							
Dimensions	Excellent(5)	Good(4)	Average(3)	Partial(2)	Unsatisfactory(1)		
Rhythm	The rhythm of the film is comprehensive , accurate, and persuasive.	The rhythm of the film is good and persuasive.	The rhythm of the film is not comprehensive and /or persuasive.	The rhythm of the film is incomplete.	The film is not presented with rhythm at all.		
Creativeness	The film shows excellent ideas that can be understood by the audience.	The film shows good ideas that can be understood by the audience.	The film does not show good ideas, but the audience can understand the content.	The film does not show good ideas, and the audience can only partially understand the content.	The film does not show good ideas, and the audience cannot understand the content.		
Technical skills	The film was edited with excellent quality editing skills and fully presented proper video effects and volume	The film was edited with good quality editing skills and partially presented proper video effects and volume	The film was edited with general quality editing skills and presented no proper video effects and volume.	The film was edited with poor quality editing skills and presented a few improper video effects and volume	The film was edited with very poor quality editing skills and presented improper video effects and volume		
	volume.	volume.		and volume.	volume.		

### 3.5. Data collection and analysis

In order to answer the four research questions, there were three kinds of data collected in this study, consisting of (1) the results of the teacher's ratings and peer reviewers' ratings, (2) peer reviewers' comments, and (3) system logs.

### 3.5.1. Reviewers' ability

A student's ability was determined by his/her submitted video. Each submitted video was rated by the course teacher and the film editing expert. The two raters independently rated 30% of the submitted videos based on the evaluation rubric. Cohen's Kappa analysis was performed to assess the inter-rater reliability of the two raters. The coefficients were 0.71, 0.68, and 0.72 for rhythm, creativeness, and technical skill respectively, showing that there was a high degree of consistency between the two raters. Finally, the course teacher evaluated the rest of the submitted videos. A median split was used to determine which students had higher ability and which had lower ability. Because one student did not review peers' videos, we excluded him from our analysis. Finally, the low-ability and high-ability groups comprised 19 and 18 students, respectively. The high-ability reviewers gained significantly higher scores in the rhythm dimension (U = 336.000, z = 5.206, p = .000, r = 0.856), creativeness dimension (U = 334.500, z = 5.107, p = .000, r = 0.840), and skill dimension (U = 309.000, z = 4.376, p = .000, r = 0.720) of the evaluation rubric and the sum of the three dimensions (U = 342.000, z = 5.233, p = .000, r = 0.861) than low-ability reviewers did (see Table 2).

Indicators	Low-a	Low-ability reviewers		ability reviewers	Mann-	
	(n = 19)			( <i>n</i> = 18)	Whitney test	
	Median 25/75 percentiles		Median	25/75 percentiles	p	
Rhythm	2.00	1.00/2.00	4.00	3.00/4.00	.000	
Creativeness	2.00	1.00/2.00	4.00	3.00/4.00	.000	
Skill	2.00	1.00/3.00	4.00	3.00/4.00	.000	
Sum of the three subscales	6.00	5.00/7.00	11.50	10.00/12.25	.000	

Table 2. The scores of the high-ability group and low-ability group's video products

### 3.5.2. Reviewing score

In addition to the teacher's ratings, each reviewer had to rate two peers' videos. The reviewing score of a peer reviewer is the average score of the two videos rated by the reviewer. There were four indicators generated from the reviewing scores. Three indicators, RhythmMeanScore, CreativenessMeanScore, and SkillMeanScore, are the reviewing scores of the three subscales respectively; and one indicator (TotalMeanScore) is the sum of the reviewing scores of the three subscales. These indicators were used for answering research question one.

### 3.5.3. Review error

The review error of a peer reviewer's rating for a video is the discrepancy between the scores of the course teacher and the peer reviewers. A lower review error represents higher review accuracy (Xiong & Schunn, 2021). There were four indicators generated from the review errors. Three indicators, RhythmError, CreativenessError, and SkillError, are the review errors of the three subscales respectively; one indicator (TotalError) is the sum of the review errors of the three subscales. These indicators were used for answering research question two.

### 3.5.4. Peer review comment coding

The reviewers' comments were qualitatively analyzed. The course teacher and the first author collaboratively developed a coding scheme (see Table 3) based on the previous studies (Cheng et al., 2015; Lu & Law, 2012). They then independently evaluated 20% of the comments based on the coding scheme. The inter-rater agreement between the two raters was calculated using Cohen's Kappa analysis, indicating a good reliability of 0.85, which is considered high agreement. Finally, the first author analyzed the rest of the comments. There were seven indicators generated from the reviewers' comments, consisting of the number of negative comments (NumNegative), number of positive comments (NumPositive), number of affective comments (NumAffective), number of comments identifying problems (NumIdentifyingProblems), number of suggestion comments (NumSuggestion), number of cognitive comments (NumCognitive), and number of all comments (NumComment). These indicators were used for answering research question three.

Table 3. Coding scheme for reviewers' comments							
Categories	Definition	Example					
Affective							
Negative	Giving criticism	The quality is bad.					
Positive	Praising the work	Very good					
Cognitive							
Identifying problems	Proposing specific problems	The video effect is not naturally presented.					
Suggestions	Providing suggestions for dealing with a problem	The beginning of this video can be cut by one second to make the actors' action look smoother.					

3.5.5. Behavior pattern analysis

The reviewers' operations in using the system were recorded in system logs. Generally, each recorded operation comprised four attributes: userId (who raised the operation), videoId (which video was viewed), operationName (the name of the operation, such as opening a Web page, closing a Web page, pausing, playing, adding a comment, mouse focusing on a Web page, and mouse focusing out of a Web page), and dateTime (the date and time of the operation performed). In this study, each reviewer needed to watch two videos on two video pages and rate two rubrics on two rubric pages. A reviewer performed different behavior patterns while watching videos and accessing the four pages. This study used k-means clustering analysis to explore the students' behavior patterns of watching the videos, and used lag sequential analysis to explore the behavior patterns of accessing the four pages (Hsu et al., 2018; Sun et al., 2018). These analyses intended to answer research question four.

A reviewer may open a video page several times and perform different behavior patterns on each opened page. In this study, the reviewers opened 203 video pages during the reviewing process. In order to understand the reviewers' viewing patterns on these opened video pages, this study established five variables for each opened video page and performed k-means clustering analysis on the five variables. The variables consisted of the time that the video was played (PlayTime), the time that the video was paused (PauseTime), the number of forward operations (NumForward), the number of backward operations (NumBackward), and the number of comment operations (NumCommentOperation). The five indicators were created because they are the most representative factors for actively viewing videos. It should be noted that the system cannot detect whether a student is actually on task. Students' inactivity (breaks, distractions etc.) could occupy a significant amount of time. Therefore, this study used time-oriented heuristics to place a threshold (4 min) (Kovanovic et al., 2015; Li & Tsai, 2017). The reason that we placed the threshold at 4 minutes is that the longest of the student's videos was 4 minutes. If a video was paused for a period of time longer than the threshold, the measured time was replaced with the threshold value.

K-means cluster analysis was performed on the five variables. Before doing the analysis, the five variables were transformed in order to reduce the bias in the cluster analysis (Li, 2019; Lust et al., 2011). The 0~20%, 21~40%, 41~60%, 61~80%, and 81~100% time durations or numbers were allocated a value of 1, 2, 3, 4, and 5, respectively, indicating very low, low, moderate, high, and very high. Two clusters were identified. The reviewers who spent more PlayTime and PauseTime and performed more NumForward, NumBackward, and NumCommentOperation in the opened video pages were classified into cluster 2, while those who spent less time were classified into cluster 1 (see Table 4). Therefore, Cluster 1 was labeled as "low active session" and Cluster 2 was labeled as "high active session."

Table 1	Cluster	analysis	of the	opened	video	nages
<i>1uvie</i> <del>4</del> .	Clusici	anary 515	or the	openeu	viuco	pages

	· · · · · ·	10
	Low Active session	High Active session
	Cluster1 ( $n = 107$ )	Cluster2 ( $n = 96$ )
PlayTime	2.084	3.990
PauseTime	2.206	3.938
NumForward	1.963	3.844
NumBackward	1.523	4.073
NumCommenOperation	1.252	3.594

To explore the reviewers' behavior patterns of accessing the four pages, we created nine codes and used lag sequential analysis to examine the patterns of accessing the four pages. The coding scheme is listed in Table 5.

	Table 5. The coding scheme of the reviewers' reviewing behaviors
Code	Description
Start	Starting the peer assessment activity
End	Finishing the peer assessment activity
Break	More than one hour break between two operations.
LAW1	Performing low active session on first video page
HAW1	Performing high active session on first video page
Rubric1	Viewing first rubric page
LAW2	Performing low active session on second video page
HAW2	Performing high active session on second video page
Rubric2	Viewing second rubric page

### 3.5.6. Statistical analyses

This study focused on between-group (high- vs. low-ability reviewers) differences in these indicators. Therefore, group comparison methods had to be conducted. SPSS software was used for analyzing the data. Because all of the indicators violated the assumption of normality, as assessed by the Shapiro–Wilk test (p < .05), Mann-

Whitney nonparametric tests were used for the indicators. The effect size was estimated by Cohen's r ( $r = z/\sqrt{n}$ ), with 0.1, 0.3, and 0.5 corresponding to small, medium, and large effect sizes (Fritz et al., 2012).

### 4. Results

#### 4.1. The reviewing scores

Four Mann-Whitney tests were conducted to compare the reviewing scores of the low- and high-ability reviewers. The results revealed that the low-ability reviewers rated RhythmMeanScore (U = 81.000, z = -2.790, p = .006, r = 0.459), CreativenessMeanScore (U = 106.000, z = -2.023, p = .049, r = 0.333), SkillMeanScore (U = 88.500, z = -2.584, p = .011, r = 0.425), and TotalMeanScore (U = 80.000, z = -2.776, p = .005, r = 0.456) significantly higher than the high-ability reviewers did (see Table 6).

These results may be caused by the difference in the quality of the videos reviewed by high-ability and lowability reviewers. To rule out the possibility, we compared the scores of the course teacher's ratings to the videos that were assigned for low-ability reviewers and high-ability reviewers. The results did not demonstrate any significant difference on the three subscales and the sum of the three subscales. These results may represent that the quality of the videos reviewed by low- and high-ability reviewers was similar.

Table 6. Reviewing scores of the low- and high-ability reviewers								
Indicators	Low-ability reviewers		High-a	ability reviewers	Mann-			
	(n = 19)			( <i>n</i> = 18)	Whitney test			
	Median	25/75 percentiles	Median	25/75 percentiles	р			
RhythmMeanScore	7.00	6.00/9.00	6.00	5.00/7.00	.006			
CreativenessMeanScore	7.00	6.00/8.00	6.00	5.00/6.25	.049			
SkillMeanScore	7.00	6.00/9.00	6.00	5.00/6.00	.011			
TotalMeanScore	22.00	18.00/24.00	18.00	15.75/19.00	.005			

### 4.2. The review errors

Four Mann-Whitney tests were conducted to compare the review errors of the low- and high-ability reviewers. The results revealed that the low-ability reviewers had significantly higher SkillError (U = 92.000, z = -2.471, p = .016, r = 0.406) than the high-ability reviewers did. However, the other indicators did not demonstrate any significant differences (see Table 7).

Table 7. Review errors of the low- and high-ability reviewers							
Indicators	Low-ability reviewers		High-a	ability reviewers	Mann-		
	(n = 19)			( <i>n</i> = 18)	Whitney test		
	Median	25/75 percentiles	Median	25/75 percentiles	р		
RhythmError	2.00	1.00/4.00	2.00	1.00/3.00	.518		
CreativenessError	3.00	2.00/3.00	2.00	1.75/3.25	1.000		
SkillError	2.00	1.00/4.00	1.00	0.75/2.25	.016		
TotalError	7.00	4.00/7.00	5.50	4.00/7.25	.199		

# 4.3. The numbers of comments provided

Seven Mann-Whitney tests were conducted to compare the numbers of different types of comments provided by the low- and high-ability reviewers. The results revealed that the high-ability reviewers provided significantly more NumIdentifyProblem (U = 235.500, z = 1.990, p = .049, r = 0.327), NumSuggestion (U = 240.500, z = 2.148, p = .032, r = 0.353), NumComment (U = 257.000, z = 2.614, p = .08, r = 0.430) and NumCognitive (U = 281.000, z = 3.345, p = .01, r = 0.550) than the low-ability reviewers did. However, low-ability reviewers provided marginally significantly more NumPositive (U = 240.500, z = 2.148, p = .032, r = 0.353) than high-ability reviewers (see Table 8).
Variable	Low-ability reviewers		High-ability reviewers		Mann-		
	(n = 19)			( <i>n</i> = 18)	Whitney test		
	Median	25/75 percentiles	Median	25/75 percentiles	р		
NumNegative	0.00	0.00/0.00	0.00	0.00/0.00	.775		
NumPositive	0.00	0.00/2.00	0.00	0.00/0.00	.081		
NumAffective	0.00	0.00/2.00	0.00	1.00/0.25	.169		
NumIdentifyProblem	2.00	0.00/3.00	3.00	2.00/5.00	.049		
NumSuggestion	2.00	0.00/2.00	4.00	1.00/6.00	.034		
NumCognitive	4.00	1.00/6.00	8.00	4.75/9.25	.002		
NumComment	4.00	3.00/6.00	8.50	4.75/9.25	.026		

Table 8. The number of comments provided by low- and high-ability reviewers

# 4.4. The behavior patterns

Three Mann-Whitney tests were conducted to compare the viewing patterns of the low- and high-ability reviewers. The results revealed that the low- and high-ability reviewers opened the same numbers of video pages. However, the high-ability reviewers demonstrated significantly more high active sessions (U = 225.500, z = 1.788, p = .098, r = 0.294) and fewer low active sessions (U = 112.000, z = -1.832, p = .057, r = 0.301) than the low-ability reviewers did (see Table 9).

Table 9. Quantitative reviewing behaviors of the low- and high-ability reviewers

Variable		Low-ability reviewers		ers	High-ability reviewers			Mann-	
		_	( <i>n</i> = 19)		(n = 18)		W	hitney test	
			Median	25/75 percentiles		Median	25/75 percentiles		p
Low Active	sessions		2.00	1.00/4.00		1.00	0.00/2.00		0.057
High active	sessions		2.00	0.00/3.	.00	2.00	2.00/4.00		0.049
Total Video	sessions		4.00	3.00/6.	.00	4.00	2.00/6.00 0.641		0.641
		Table 10	0. The adjus	ted residual	table of th	ne low-abili	ty reviewers		
	Start	LAW1	HAW1	Rubric1	Break	LAW2	HAW2	Rubric2	End
Start	0	$4.132^{*}$	$2.823^{*}$	-0.223	-1.144	-1.826	-1.467	-1.863	-1.467
LAW1	0	1.657	-0.481	-0.265	0.343	$2.543^{*}$	-1.184	-1.530	-1.775
HAW1	0	-1.691	-1.382	$5.066^{*}$	-1.111	0.380	0.275	-1.102	-1.424
Rubric1	0	-0.715	-0.762	-1.615	0.689	0.031	$2.937^{*}$	0.990	-0.855
Break	0	1.802	$4.135^{*}$	-0.872	-0.893	-1.424	-1.144	-0.601	-1.144
LAW2	0	-0.450	-1.772	-0.502	1.170	-1.090	-0.426	1.759	1.675
HAW2	0	-2.366	0.275	-0.223	-0.120	-1.826	0.191	$4.345^{*}$	0.191
Rubric2	0	-2.022	-1.102	-0.588	0.251	1.759	0.206	-1.793	$4.345^{*}$
End	0	0	0	0	0	0	0	0	0
<i>Note.</i> $*p < .05$ .									

Table 11. The adjusted residual table of the high-ability reviewers

	Start	LAW1	HAW1	Rubric1	Break	LAW2	HAW2	Rubric2	End
Start	0	1.712	3.683*	0.942	-1.478	-0.323	-1.826	-1.826	-1.395
LAW1	0	-0.900	1.207	0.540	$2.242^{*}$	-1.375	-0.020	-1.332	-0.711
HAW1	0	-1.381	-1.932	3.593*	-1.313	0.573	$3.195^{*}$	-2.438	-1.169
Rubric1	0	0.540	-0.411	-2.060	0.667	0.740	1.712	0.189	-0.932
Break	0	0.701	3.316*	-0.525	-1.567	1.352	-1.935	0.077	-1.478
LAW2	0	1.257	-0.933	-1.297	-0.442	0.899	-0.889	1.402	0.617
HAW2	0	-0.676	-1.875	-1.841	0.077	-0.889	-0.676	$6.179^{*}$	-0.420
Rubric2	0	-0.676	-1.875	0.697	1.418	-0.889	-0.676	-2.390	$5.204^{*}$
End	0	0	0	0	0	0	0	0	0

*Note.* \**p* < .05.

To explore the differences in the sequential patterns of the low- and high-ability reviewers, two lag sequential analyses were performed. The adjusted residual tables of the low- and high-ability reviewers are shown in Table 10 and Table 11 respectively, where the row presents the starting behavior and the column presents the following behavior. The value in each cell of the tables is the Z-score. The significant relationship is marked with a "\*" when the Z-score is greater than 1.96. Figure 3 and Figure 4 further present the behavioral transition diagrams of

low-ability and high-ability reviewers, respectively. They shows significant behavioral patterns. Some similar behavioral patterns were exhibited by the low- and high-ability reviewers. These are Start->HAW1, HAW1->Rubric1, Break->HAW1, HAW2->Rubric2, and Rubric2->End. In addition, several patterns differed between the low- and high-ability reviewers. The low-ability reviewers frequently performed the following patterns: Start->LAW1, LAW1->LAW2, and Rubric1->HAW2, while the high-ability reviewers frequently performed the following patterns: LAW1->Break and HAW1->HAW2.

*Figure 3.* Behavioral transition diagram of low-ability reviewers *Note.* The arrows refer to the direction of the sequential transfer. The value on an arrow is the z-score of the sequential transfer.



*Figure 4.* Behavioral transition diagram of high-ability reviewers

*Note.* The arrows refer to the direction of the sequential transfer. The value on an arrow is the z-score of the sequential transfer.



#### **5.** Discussion

There were four research questions in this study. For Question 1, "Did the reviewers' ability affect the scores they gave?", this study found that low-ability reviewers preferred to rate higher scores for their peers' videos than high-ability reviewers did. This result may be explained by Dunning-Kruger effect (Biango-Daniels & Sarvary, 2021; Kruger & Dunning, 1999). Lower-ability reviewers may overestimate their own knowledge or competence in film editing. When a low-ability reviewer assessed a peer's video that has the same quality as his/her video. He/She may rate a high score. The result is similar to previous studies that found peer-assessment was overestimated compared to instructors' assessment (Biango-Daniels & Sarvary, 2021; Lynch & Schmid, 2017). However, this result is inconsistent with Xiong and Schunn's (2021) study. They found that low-ability reviewers tended to be more severe. The inconsistent result may be that the dependent variables and the statistical methods were different. In our study, the dependent variable was reviewing score, a continuous data type, and was tested by Mann-Whitney nonparametric test. However, in Xiong and Schunn's (2021) study, the dependent variable was review errors using the difference between peer reviewers' and course teacher's ratings. They further categorized the review

errors as Severe, Lenient, and Accurate. Review errors below -1 were categorized as Severe; errors above 1 were categorized as Lenient; and everything between (including) -1 and 1 was categorized into the Accurate category. They then used logistic regression to examine the relation between reviewers' ability and each review error type. Therefore, further research should be conducted to ensure the effect of reviewers' ability on reviewers' rating preferences.

For question 2, "Did the reviewers' ability affect their review error?", we found that the review errors of the subscale "technical skill" of the low- and high-ability reviewers demonstrated a significant difference, but the review errors of the subscales "rhythm" and "creativeness" did not. Because high-ability reviewers had more knowledge of the problems related to technical skills, they could more accurately identify the problems related to technical skills and had lower review errors concerning technical skills than low-ability reviewers did. However, it is not clear why the subscales of rhythm and creativeness did not demonstrate significant differences. The reason may be that rhythm and creativeness knowledge is tacit and subjective. Tacit knowledge is accumulated from immense histories of life and work experience (Tee & Karney, 2010). Students have less experience of film editing and therefore may have difficulty evaluating rhythm and creativeness knowledge.

For question 3, "Did the reviewers' ability affect the quantity and quality of the comments they provided?", highability reviewers provided more cognitive comments and total comments on their peers' video assignments than low-ability reviewers did. This result is consistent with previous studies (Patchan et al., 2013). However, lowability reviewers provided more positive comments than high-ability reviewers did. Because high-ability reviewers had more knowledge of the subject and problems, they could more easily detect problems and provide more elaborate diagnoses (Patchan & Schunn, 2016). Therefore, they provided more cognitive comments. Reviewing peers' videos was a learning activity. Although the low-ability reviewers may have had difficulty identifying problems and providing suggestions, they still had to provide comments on their peers' videos. Emotional comments are more easily created than cognitive comments. Therefore, they posted more positive comments than high-ability reviewers did. These results are consistent with previous studies (Alqassab et al., 2018; Patchan & Schunn, 2015) which found that low-ability reviewers preferred to give emotional comments, while high-ability reviewers preferred to give cognitive comments.

For questions 4, "Did the reviewers' ability affect their behavioral patterns?", we found that the low- and highability reviewers showed different patterns of viewing peers' videos and navigating the four pages. In terms of the patterns of viewing videos, low-ability reviewers performed significantly more low active sessions than highability reviewers did. We have examined what reviewers did in the low active sessions and found two primary behavior patterns: long playing and short playing with a few forward and backward operations. The two viewing patterns were also observed by previous studies (de Boer et al., 2016). When the reviewers performed the pattern of long playing, they played the whole video or most of the video without any other operations or just a few other operations. This pattern may represent that they watched the video to understand the video content. On the other hand, students who performed short playing with a few forward and backward operations may have constructed an overview of the video content or wanted to find specific content. Therefore, this result that the low-ability reviewers performed significantly more low active sessions may imply that the low-ability reviewers invested more effort in understanding the video content than high-ability reviewers. However, the high-ability reviewers performed significantly more high active sessions than the low-ability reviewers did. In a highly active session, the reviewers watched the video for a long time and performed complex operations (e.g., adding and editing comments, moving forwards and backwards, and playing and pausing). This is similar to the strategic viewing behavior mentioned in de Boer et al. (2016). The reviewers performed these behaviors not only to understand the video content but also to detect and diagnose the problems. Therefore, this result that the highability reviewers performed significantly more high active sessions may imply that the high-ability reviewers invested more time and effort in detecting and diagnosing problems than the low-ability reviewers did. These results are similar to Li's (2019) study which found that low prior knowledge students spent most of the time viewing the videos for acquiring information, while the high prior knowledge students spent a considerable amount of time performing the viewing strategies for eliminating the discrepancies between their current knowledge state and the information presented in the videos.

In terms of navigational patterns, we found that the low- and high-ability reviewers performed some of the same sequential patterns and some different patterns for navigating the four pages. The low- and high-ability reviewers both performed the following patterns: start->HAW1->rubric1 and HAW2->rubric2->end. These patterns represent that they reviewed the first video at the beginning and the second video at the end of the whole reviewing process. These may imply that reviewers firstly reviewed the first assignment in the reviewing page and then reviewed the second one. This behavior is similar to the behavior of depth first processing of search result lists (Klöckner et al., 2004). Although low- and high-ability reviewers performed these similar behavior patterns, they also performed several different behavioral patterns. The low-ability reviewers performed the

pattern Rubric1->HAW2. This pattern may represent that they reviewed the second video after reviewing the first video. Additionally, the low-ability reviewers also performed the pattern Start->LAW1->LAW2. Because the reviewers performed the low active sessions to understand the video content, the sequential patterns may represent that low-ability reviewers spent more time and effort understanding the video content at the beginning of their reviewing. We also observed that low-ability reviewers performed the pattern HAW1->Rubric1->HAW2->Rubric2. They provided detailed feedback and rated the videos one by one. This pattern may imply that the low-ability reviewers assessed the two videos separately. While high-ability reviewers performed the scores. This pattern may imply that high-ability reviewers treated the two videos as a whole. They assessed the two videos in a summative way (Hsia et al., 2016b).

In sum, the low-ability reviewers provided fewer comments, demonstrated more low active sessions, and assessed the two videos separately. While the high-ability reviewers provided more comments, demonstrated more high active sessions, and assessed the two videos in a summative way. Two reasons may explain the different behaviors. First, low-ability reviewers may be less self-regulated learners. In this study, a student's ability was determined by his/her submitted video. The submitted video was an outcome of the peer assessment activity. It is closely linked to the three components of self-regulation: motivation, cognition, and metacognition (Trautwein & Koller, 2003). Second, low-ability reviewers were imposed a high cognitive load. The high cognitive load can significantly lower their self-regulated effort, the degree to which students can maintain motivation and persist with learning tasks (Hughes et al., 2018). Therefore, the low-ability reviewers may have a lower motivation to review peers' works, especially the submissions are videos. Compared with static documents (e.g., composition), the videos' navigational operations and transient nature can significantly increase students' cognitive load. On the one hand, the navigation operations in videos are more complex than in static documents (Leahy & Sweller, 2011). Learners move their eye focus to find a specific content in a static document. However, they drag the video timeline to a specific video frame and then move their eye focus to find a specific content in the video frame. The timeline does not provide any information cue for learners to locate a specific video frame. Therefore, learners may experience higher cognitive load and disorientation when navigating in a video than when navigating in a static document. On the other hand, the video is transient media, in which the content is dynamically changed. Learners must keep previously viewed content in working memory for comparing and integrating contents among different frames. It imposes a high cognitive load, especially the videos are long and complex (Leahy & Sweller, 2011; Leahy & Sweller, 2016). Because low-ability reviewers may be less self-regulated learners and may not persist with their review tasks, they provided fewer comments, performed simple operations (i.e., long playing and short playing with a few forward), and reviewed the two videos separately.

# 6. Conclusion

Reviewing peers' assignments is a complex process. It involves understanding the content, detecting and diagnosing the problems, and giving scores. Previous studies on written text have shown that reviewers' ability can significantly affect the comments and ratings they provide and their reviewing performance. Because video is a transient medium, watching videos imposes higher cognitive load than reading written text, and so reviewers' ability should have stronger effects on reviewing outcomes, behaviors, and performance. Therefore, this study examined how reviewers' ability affected their comments and ratings and their reviewing behaviors and performance. We found that low-ability reviewers tended to rate higher scores for peers' videos and demonstrated higher review errors than high-ability reviewers. In addition, low- and high-ability reviewers obviously performed different behavior patterns. In particular, the low-ability reviewers invested more time and effort in understanding the video content, while the high-ability reviewers also provided more comments for peers' videos, especially cognitive comments.

Although this study made a number of significant findings, several limitations should be mentioned. First, the sample size was small, which limits the extent of generalizability of the findings. In the future, we can involve more participants to examine the effects of reviewers' ability. Second, reviewers' ability measured by their submissions is an indirect measure. It may bias the research findings (Xiong & Schunn, 2021). In the future, reviewers' knowledge and skills relevant to detecting problems and providing feedback should be investigated. Third, the students were trained on how to provide comments and how to rate the videos in the first 5 weeks of the course and were provided with a detailed rubric. However, previous studies used different approaches to support participants' reviews. For example, the participants in Huisman et al.'s (2018) study were not trained. Patchan and Schunn (2016) provided their participants with a detailed rubric, including commonly-used general

reviewing suggestions and specific guidelines. How the students were trained and supported may also influence the research findings (Liu & Li, 2014). Therefore, future works can examine the effects of different training approaches and reviewing scaffolding for the reviewing process and performance. Fourth, this study examined the effects of reviewers' ability on the reviewing process and performance. Other individual characteristics, such as previous experience of peer review and online learning, might also affect the process and performance of online video peer assessment (Sahan & Razi, 2020; Zou et al., 2018). However, we did not control the variables. They may bias the research findings. Future work can control these variables or investigate the main and interactive effect of these individual characteristics.

Despite these limitations, this study contributes to our knowledge regarding online video peer assessment. The study provides a first insight into the relationships between reviewers' ability and the reviewing process and performance for online video peer assessment practices. Because design is a progressive and repeated process, the findings of this study can provide useful information for improving our instruction and system for video peer assessment activities. Three practical implications can be derived from the findings. First, because low-ability reviewers prefer to give higher scores for peers' videos, it is suggested that teachers should consider students' ability to assign the same number of low- and high-ability reviewers for a video assignment in order to ensure fairness. Second, low-ability reviewers provided more emotional comments and fewer cognitive comments. Providing feedback has been found to lead to greater improvements from pre-test to post-test than receiving feedback (Patchan & Schunn, 2015). In particular, providing cognitive comments has stronger effects on reviewers' learning than providing emotional comments, because students can practice detection and diagnosis skills rather than just detection skills (Patchan & Schunn, 2015). Therefore, teachers and system designers should help low-ability reviewers post cognitive comments. Several suggestions may help them. First, teachers can train the students and provide clear guidelines for providing cognitive comments. Second, systems can detect the type of comment. If the comment is emotional comment, the system can ask the reviewer to elaborate on the comment. Machine learning can be used to identify the type of comment and to provide instant recommendations (Dood et al., 2022). Finally, low-ability reviewers performed more low active sessions to understand the video content. In order to help them understand the content, system developers can provide tools to help students understand the video content. For example, the video playing interface can show reviewers' viewing history, so reviewers can understand what they have done before and what video content they have watched. This may decrease the time of viewing the videos and allow more time for detecting and diagnosing problems. In addition, the video playing interface can also provide a noting function. The notes can remind reviewers what they have done and what they have thought before about the videos.

## Acknowledgement

This work was supported by the Ministry of Science and Technology of Taiwan under contract numbers MOST108-2511-H-003-039-MY2 and MOST110-2511-H-003-031-MY2 and the "Institute for Research Excellence in Learning Sciences" of National Taiwan Normal University (NTNU) from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

## References

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education*, 33(1), 11-30. https://doi.org/10.1007/s10212-017-0342-0

Biango-Daniels, M., & Sarvary, M. (2021). A Challenge in teaching scientific communication: academic experience does not improve undergraduates' ability to assess their or their peers' writing. *Assessment & Evaluation in Higher Education*, 46(5), 809-820. https://doi.org/10.1080/02602938.2020.1812512

Biggs, J. B. (1987). *Student approaches to learning and studying research monograph*. Australian Council for Educational Research Ltd.

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A Systematic literature review. *Computers and Education: Artificial Intelligence*, *2*, 100027. https://doi.org/10.1016/j.caeai.2021.100027

Cerezo, R., Sanchez-Santillan, M., Paule-Ruiz, M. P., & Nunez, J. C. (2016). Students' LMS interaction patterns and their relationship with achievement: A Case study in higher education. *Computers & Education*, *96*, 42-54. https://doi.org/10.1016/j.compedu.2016.02.006

Chen, I. C., Hwang, G. J., Lai, C. L., & Wang, W. C. (2020). From design to reflection: Effects of peer-scoring and comments on students' behavioral patterns and learning outcomes in musical theater performance. *Computers & Education*, 150, 103856. https://doi.org/10.1016/j.compedu.2020.103856

Cheng, K. H., Liang, J. C., & Tsai, C. C. (2015). Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *Internet and Higher Education*, 25, 78-84. https://doi.org/10.1016/j.iheduc.2015.02.001

Crisp, V. (2010). Towards a model of the judgement processes involved in examination marking. Oxford Review of Education, 36(1), 1-21. https://doi.org/10.1080/03054980903454181

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A Descriptive framework. *Modern Language Journal*, 86(1), 67-96. https://doi.org/10.1111/1540-4781.00137

Cybinski, P., & Selvanathan, S. (2005). Learning experience and learning effectiveness in undergraduate statistics: Modeling performance in traditional and flexible learning environments. *Decision Sciences Journal of Innovative Education*, 3(2), 251–271.

de Boer, J., Kommers, P. A., de Brock, B., & Tolboom, J. (2016). The Influence of prior knowledge and viewing repertoire on learning from video. *Education and information technologies*, 21(5), 1135–1151.

de Boer, J., Kommers, P. A. M., & de Brock, B. (2011). Using learning styles and viewing styles in streaming video. *Computers & Education*, 56(3), 727-735. https://doi.org/10.1016/j.compedu.2010.10.015

Dood, A., Winograd, B., Finkenstaedt-Quinn, S., Gere, A., & Shultz, G. (2022). PeerBERT: Automated characterization of peer review comments across courses. *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 492–499). https://doi.org/10.1145/3506860.3506892

Er, E., Dimitriadis, Y., & Gasevic, D. (2021). Collaborative peer feedback and learning analytics: theory-oriented design for supporting class-wide interventions. *Assessment & Evaluation in Higher Education*, 46(2), 169-190. https://doi.org/10.1080/02602938.2020.1764490

Er, E., Villa-Torrano, C., Dimitriadis, Y., Gasevic, D., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., Eduardo Gómez-SánchezGómez-Sánchez, E., & Monés, A. M. (2021). Theory-based learning analytics to explore student engagement patterns in a peer review activity. *LAK21: 11th International Learning Analytics and Knowledge Conference* (pp. 196–206). https://doi.org/10.1145/3448139.3448158

Formanek, M., Wenger, M. C., Buxner, S. R., Impey, C. D., & Sonam, T. (2017). Insights about large-scale online peer assessment from an analysis of an astronomy MOOC. *Computers & Education*, *113*, 243-262. https://doi.org/10.1016/j.compedu.2017.05.019

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology-General*, 141(1), 2-18. https://doi.org/10.1037/a0024338

Hsia, L. H., Huang, I., & Hwang, G. J. (2016a). A Web-based peer-assessment approach to improving junior high school students' performance, self-efficacy and motivation in performing arts courses. *British Journal of Educational Technology*, 47(4), 618-632. https://doi.org/10.1111/bjet.12248

Hsia, L. H., Huang, I., & Hwang, G. J. (2016b). Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course. *Computers & Education*, *96*, 55-71. https://doi.org/10.1016/j.compedu.2016.02.004

Hsu, T. C., Chang, S. C., & Liu, N. C. (2018). Peer assessment of webpage design: Behavioral sequential analysis based on eye tracking evidence. *Educational Technology & Society*, 21(2), 305-321.

Hughes, C., Costley, J., & Lange, C. (2018). The Effects of self-regulated learning and cognitive load on beginning to watch and completing video lectures at a cyber-university. *Interactive Technology and Smart Education*, *15*(3), 220-237.

Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The Relationship between peer reviewers' ability and authors' essay performance. *British Journal of Educational Technology*, 49(1), 101-110. https://doi.org/10.1111/bjet.12520

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2017). Peer feedback on college students' writing: Exploring the relation between students' ability match, feedback quality and essay performance. *Higher Education Research & Development*, *36*(7), 1433-1447. https://doi.org/10.1080/07294360.2017.1325854

Klöckner, K., Wirschum, N., & Jameson, A. (2004). Depth- and breadth-first processing of search result lists. *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (pp. 1539-1539). https://doi.org/10.1145/985921.986115

Kovanovic, V., Gasevic, D., Joksimovic, S., Hatala, M., & Adesope, O. (2015). Analytics of communities of inquiry: Effects of learning technology use on cognitive presence in asynchronous online discussions. *Internet and Higher Education*, 27, 74-89. https://doi.org/10.1016/j.iheduc.2015.06.002

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6), 1121-1134. https://doi.org/10.1037/0022-3514.77.6.1121

Lai, C. Y. (2016). Training nursing students' communication skills with online video peer assessment. *Computers & Education*, 97, 21-30. https://doi.org/10.1016/j.compedu.2016.02.017

Lai, C. Y., Chen, L. J., Yen, Y. C., & Lin, K. Y. (2020). Impact of video annotation on undergraduate nursing students' communication performance and commenting behaviour during an online peer-assessment activity. *Australasian Journal of Educational Technology*, *36*(2), 71-88. https://doi.org/10.14742/ajet.4341

Leahy, W., & Sweller, J. (2011). Cognitive load theory, modality of presentation and the transient information effect. *Applied Cognitive Psychology*, 25(6), 943-951. https://doi.org/10.1002/acp.1787

Leahy, W., & Sweller, J. (2016). Cognitive load theory and the effects of transient information on the modality effect. *Instructional Science*, 44(1), 107-123. https://doi.org/10.1007/s11251-015-9362-9

Li, L. Y. (2019). Effect of prior knowledge on attitudes, behavior, and learning performance in video lecture viewing. *International Journal of Human-Computer Interaction*, 35(4-5), 415-426. https://doi.org/10.1080/10447318.2018.1543086

Li, L. Y., & Tsai, C. C. (2017). Accessing online learning material: Quantitative behavior patterns and their effects on motivation and learning performance. *Computers & Education*, 114, 286-297. https://doi.org/10.1016/j.compedu.2017.07.007

Li, L. Y., & Tsai, C. C. (2020). Students' patterns of accessing time in a text structure learning system: Relationship to individual characteristics and learning performance. *Etr&D-Educational Technology Research and Development*, 68(5), 2569-2594. https://doi.org/10.1007/s11423-020-09780-7

Li, S., Zheng, J., & Lajoie, S. P. (2022). Temporal structures and sequential patterns of self-regulated learning behaviors in problem solving with an intelligent tutoring system. *Educational Technology & Society*, 25(4), 1-14.

Lin, H.-C., Hwang, G.-J., Chang, S.-C., & Hsu, Y.-D. (2021). Facilitating critical thinking in decision making-based professional training: An Online interactive peer-review approach in a flipped learning context. *Computers & Education, 173*, 104266. https://doi.org/10.1016/j.compedu.2021.104266

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinkingstyles. *Journal of Computer Assisted Learning*, 17(4), 420-432.

Liu, X. Y., & Li, L. (2014). Assessment training effects on student assessment skills and task performance in a technology-facilitated peer assessment. *Assessment & Evaluation in Higher Education*, 39(3), 275-292. https://doi.org/10.1080/02602938.2013.823540

Lu, J. Y., & Law, N. (2012). Online peer assessment: Effects of cognitive and affective feedback. *Instructional Science*, 40(2), 257-275. https://doi.org/10.1007/s11251-011-9177-2

Lust, G., Vandewaetere, M., Ceulemans, E., Elen, J., & Clarebout, G. (2011). Tool-use in a blended undergraduate course: In Search of user profiles. *Computers & Education*, 57(3), 2135-2144.

Lynch, D., & Schmid, B. (2017). Peer evaluation: Enhancing learning opportunities and reducing marking effort. *Proceedings of the Canadian Engineering Education Association (CEEA)*. https://doi.org/10.24908/pceea.v0i0.10368

Mayer, R. E. (2002). Multimedia learning. Psychology of Learning and Motivation, 41, 85-139.

Mirriahi, N., Liaqat, D., Dawson, S., & Gasevic, D. (2016). Uncovering student learning profiles with a video annotation tool: Reflective learning with and without instructional norms. *Etr&D-Educational Technology Research and Development*, *64*(6), 1083-1106. https://doi.org/10.1007/s11423-016-9449-2

Moos, D. C., & Azevedo, R. (2008). Self-regulated learning with hypermedia: The Role of prior domain knowledge. *Contemporary Educational Psychology*, 33(2), 270-298. https://doi.org/10.1016/j.cedpsych.2007.03.001

Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The Effects of skill diversity on commenting and revisions. *Instructional Science*, 41(2), 381-405. https://doi.org/10.1007/s11251-012-9236-3

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science*, 43(5), 591-614. https://doi.org/10.1007/s11251-015-9353-x

Patchan, M. M., & Schunn, C. D. (2016). Understanding the effects of receiving peer feedback for text revision: Relations between author and reviewer ability. *Journal of Writing Research*, 8(2), 227-265. https://doi.org/10.17239/jowr-2016.08.02.03

Sahan, O., & Razi, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing*, *37*(3), 311-332. https://doi.org/10.1177/0265532219900228

Shibani, A., Knight, S., & Shum, S. B. (2019). Contextualizable learning analytics design: A Generic model and writing analytics evaluations. *Proceedings of the 9th International Conference on Learning Analytics & Knowledge* (pp. 210–219). https://doi.org/10.1145/3303772.3303785

Shibani, A., Knight, S., & Shum, S. B. (2022). Questioning learning analytics? Cultivating critical engagement as student automated feedback literacy. *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 326–335). https://doi.org/10.1145/3506860.3506912

Shirvani Boroujeni, M., & Dillenbourg, P. (2019). Discovery and temporal analysis of MOOC study patterns. *Journal of Learning Analytics*, 6, 16-33.

Song, H. S., Kalet, A. L., & Plass, J. L. (2016). Interplay of prior knowledge, self-regulation and motivation in complex multimedia learning environments. *Journal of Computer Assisted Learning*, 32(1), 31-50. https://doi.org/10.1111/jcal.12117

Sun, J. C. Y., Hwang, G. J., Lin, Y. Y., Yu, S. J., Pan, L. C., & Chen, A. Y. Z. (2018). A Votable concept mapping approach to promoting students' attentional behavior: An Analysis of sequential behavioral patterns and brainwave data. Educational Technology & Society, 21(2), 177-191.

Tee, M. Y., & Karney, D. (2010). Sharing and cultivating tacit knowledge in an online learning environment. *International Journal of Computer-Supported Collaborative Learning*, 5(4), 385-413. https://doi.org/10.1007/s11412-010-9095-3

Tenorio, T., Bittencourt, II, Isotani, S., & Silva, A. P. (2016). Does peer assessment in on-line learning environments work? A Systematic review of the literature. *Computers in Human Behavior*, *64*, 94-107. https://doi.org/10.1016/j.chb.2016.06.020

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3), 249-276. https://doi.org/10.2307/1170598

Trautwein, U., & Koller, O. (2003). The Relationship between homework and achievement - Still much of a mystery. *Educational Psychology Review*, *15*(2), 115-145. https://doi.org/10.1023/a:1023460414243

Wang, J. H., Chen, S. Y., Chang, B., & Chan, T. W. (2016). From integrative to game-based integrative peer response: High ability versus low ability. *Journal of Computer Assisted Learning*, *32*(2), 170-185. https://doi.org/10.1111/jcal.12125

Wu, C. C., & Kao, H. C. (2008). Streaming videos in peer assessment to support training pre-service teachers. *Educational Technology & Society*, 11(1), 45-55.

Wu, Y., & Schunn, C. D. (2020). From feedback to revisions: Effects of feedback features and perceptions. *Contemporary Educational Psychology*, 60, Article 101826. https://doi.org/10.1016/j.cedpsych.2019.101826

Xiong, Y., & Schunn, C. D. (2021). Reviewer, essay, and reviewing-process characteristics that predict errors in web-based peer review. *Computers & Education*, *166*, 104146. https://doi.org/10.1016/j.compedu.2021.104146

Zarzour, H., Bendjaballah, S., & Harirche, H. (2020). Exploring the behavioral patterns of students learning with a Facebookbased e-book approach. *Computers & Education*, 156, 103957. https://doi.org/10.1016/j.compedu.2020.103957

Zou, Y., Schunn, C. D., Wang, Y. Q., & Zhang, F. H. (2018). Student attitudes that predict participation in peer assessment. *Assessment & Evaluation in Higher Education*, 43(5), 800-811. https://doi.org/10.1080/02602938.2017.1409872

Tlili, A., Burgos, D., & Looi, C.-K. (2023). Guest Editorial: Creating Computational Thinkers for the Artificial Intelligence Era— Catalyzing the Process through Educational Technology. *Educational Technology & Society*, 26(2), 94-98. https://doi.org/10.30191/ETS.202304\_26(2).0007

# Guest Editorial: Creating Computational Thinkers for the Artificial Intelligence Era—Catalyzing the Process through Educational Technology

# Ahmed Tlili<sup>1\*</sup>, Daniel Burgos<sup>2</sup> and Chee-Kit Looi<sup>3</sup>

<sup>1</sup>Smart Learning Institute of Beijing Normal University, China // <sup>2</sup>UNIR iTED, Universidad Internacional de La Rioja (UNIR), Spain // <sup>3</sup>National Institute of Education, Nanyang Technological University, Singapore // ahmed.tlili23@yahoo.com // daniel.burgos@unir.net // cheekit.looi@outlook.com

\*Corresponding author

**ABSTRACT:** There is an ongoing debate in the literature about the ways of using technology to enhance students' Computational Thinking (CT). This special issue further enriches this debate by investigating how educational technology could be used, and for which purposes, to facilitate learning CT. It includes six papers demonstrating the innovative design of curricula and the use of various technologies to teach CT for students in different educational levels. Based on these papers, this special issue points out that more research is needed to investigate the best educational practices that could be used to teach CT rather than focusing on the technology itself. It also reveals that future work could cover smart learning analytics and precision education to better model students' individual differences, hence effectively supporting learning CT.

Keywords: Computational thinking, Artificial Intelligence, Educational technology, Future education, Competencies

# 1. What is computational thinking? Unveiling the ambiguity

The digital transformation and the rapid evolution of Artificial Intelligence (AI) have catalyzed the use of machines in our daily activities, where computers and their algorithms have changed the way that we think to better communicate and utilize them (Tlili et al., 2022). The world is becoming more complex and unpredictable, where students should acquire the basic skills to deal with it. The thinking processes associated with the problem-solving approach of Computational Thinking (CT) allows learners to better deal with the complexity and open-ended non-trivial problems posed by the world and its emerging technologies (e.g., AI and big data). Therefore, several research studies advocated considering CT as an essential competence that should be included in all educational levels and in every student's skill set (Grover & Pea, 2018). In his constructivist work with technology, Seymour Papert (Papert, 1980) was the first to introduce CT, which then got more popularity after the researcher Jeannette Wing (Wing, 2006) published a paper in 2006 discussing CT. She argues that:

"Computational thinking builds on the power and limits of computing processes, whether they are executed by a human or by a machine. Computational methods and models give us the courage to solve problems and design systems that no one of us would be capable of tackling alone. ... Computational thinking is a fundamental skill for everyone, not just for computer scientists. To reading, writing, and arithmetic, we should add computational thinking to every child's analytical ability. Just as the printing press facilitated the spread of the three Rs, what is appropriately incestuous about this vision is that computing and computers facilitate the spread of computational thinking." (Wing, 2006, p. 33)

CT can be confusing at it can also be related to several terms like computers and computing (Li et al., 2020) or computer science and programming (Czerkawski & Lyman, 2015). It is therefore important to further clarify these terms to readers. There is a large agreement that computing has contributed to revolutionize science. A key instance of this is the computational science movement in 1980, when several researchers claimed that computing is a new way to conduct science. Due to its significant importance, computing was considered as the "third pillar" of science (Oberkampf and Roy, 2010), the "fourth great scientific domain" (Rosenbloom, 2013), and the "most disruptive paradigm shift in the sciences since quantum mechanics" (Chazelle, 2006). Additionally, while computing has inspired researchers to look at CT, it is still not CT, *per se*. Computing is the study of natural and artificial information processing (Denning, 2007). CT, on the other hand, is much broader than that and focuses on the models and methods of processing information which could be in different formats and shapes, as well as the needed skills for that (Cansu & Cansu, 2019).

CT skills are not unique to computing and can be found in several disciplines (Li et al., 2020). In this context, Ioannidou et al. (2011) pointed out that CT skills are not the same as programming skills, but programming is a good context for helping to think computationally (Israel et al., 2015). Shute et al. (2017) further highlighted that

CT skills and concepts cover: (1) decomposition, (2) abstraction, (3) debugging, (4) iteration, (5) generalization, and (6) algorithms and their design.

Finally, while CT originates from computer science (Wing, 2006), it differs from computer science as it allows users to transfer CT skills to domains other than programming (Berland & Wilensky, 2015), such as everyday activities and problems. The misconception of CT further continued as several people considered it as "thinking like a computer" (Kite & Park, 2020), however, Wing (2006) clearly stated that this not correct raising concerns that thinking like a machine might hinder creative and divergent aspects of CT (e.g., systems thinking, problem decomposition, and abstraction). Therefore, emphasizing the difference between human thinking and computer thinking is essential in CT.

As there was no exact understanding of CT, several definitions were proposed in the literature accordingly. For instance, one of the most accepted CT definitions is that of Cuny et al. (2010, p. 1) where they considered CT as a thinking process where "...solutions are represented in a form that can be effectively carried out by an information-processing agent." This covers both well-structured problems and ill-structured problems (i.e., complicated real-life problems whose solutions are neither definite nor measurable). The National Research Council of The National Academies (NRC) considered CT as "... a fundamental analytical skill that everyone, not just computer scientists, can use to help solve problems, design systems, and understand human behavior. ... Computational thinking is likely to benefit not only other scientists but also everyone else." Berland and Wilensky (2015) further defined CT as "the ability to think with the computer-as-tool" (p. 630).

While the above-mentioned definitions (as well as those in the literature) are different to a certain extent, all of them agree that CT is a mental skill that everyone should acquire in this digital and AI era, even for non-scientists, where various tools and technologies could be used to facilitate the process.

# 2. Learning computational thinking and the role of educational technology

Since CT is very important and is considered by UNESCO (2021) as one of the five pillars for guiding AI and education, several countries have started teaching it, as well as its competencies in schools and universities. The need to enhance CT has been taken up at the policy level by national governments, for example in elementary schools in Sweden, as described by Kjällander et al. (2021). Finland and Australia have also joined this movement where they made computing/coding compulsory subjects in primary schools (Rich et al., 2019). Aligned with this international trend, China has deemed CT to be one of the core literacies of information technology curriculum and has included it in the National High School Information Technology Curriculum Standards (Zhang et al., 2023).

Several studies have relied on traditional tools (e.g., Lego) instead of technology, also known as unplugged activities, to teach CT (Zhang et al., 2023). They proved that CT could be taught with cost-effective approaches that are not technology centric. However, these approaches also revealed that there is a need for more sophisticated tools to better guide students in learning CT. For instance, Lee et al. (2020) reported that several STEM classrooms are failing to integrate CT into their curricula. De Jong and Jeuring (2020) revealed that more investigation is needed for new assessment methods to measure students' CT skills. There is, therefore, a need for technology that can keep track of how students learn in each step of the learning process, and analyze their learning log data to identify the learning obstacles faced, hence enhancing the learning process of CT. In this context, learning methods, several studies highlighted the importance of using robots (Yang et al., 2020) or visual programming tools (Fagerlund et al., 2021), among others, to teach CT.

Other studies, on the other hand, pointed out that teaching CT requires careful design of the technology which should support multiple combinations, and offer multiple ways to solve a problem (Bers, 2020). The designed technologies need to provide opportunities for creating a computational artifact that can be shared with others and support a growing range of computational literacy skills, from beginner to expert (Bers, 2020). Hamilton et al. (2020) and Pugnali et al. (2017) reported that the issue facing educational technologists is how to select the appropriate tools and practices to teach CT. Therefore, the use of educational technology to teach CT requires careful thinking in terms of the technology to be designed and used, the teaching practices and curricula, and the assessment methods.

Given the aforementioned background, this special issue aims to enrich the ongoing discussion about the use of educational technology to enhance students' CT. Through the six accepted papers, as shown in Section 3, this

special issue provides more insights about the effective ways of using educational technology to catalyze CT learning.

# 3. Contribution of papers to this special issue

The following six papers were accepted in this special issue. Each one of them elicit theoretical and practical knowledge about the use of educational technology to enhance students' CT.

To enrich the ongoing debate in the literature about CT curricula especially in primary and lower secondary education, Paper 1 conducts a systematic review on this topic, where 98 studies were covered. The obtained results revealed that while several technologies exist for age-appropriate CT development, more research is required to design and develop curricula and pedagogies for utilizing these tools effectively to foster young learners' CT skill development.

With limited tools exist about CT assessment of students at an early age, Paper 2 develops *TechCheck*, an assessment of Computational Thinking (CT) for early elementary school children consisting of fifteen developmentally appropriate unplugged challenges that probe six CT domains.

Based on four mathematics domains (arithmetic, random events and counting, number theory, and geometry), Paper 3 designed a series of programming-based learning tasks for middle school students to co-develop CT and the corresponding mathematical knowledge. The obtained results revealed that the dynamic representations and immediate visual feedback afforded by the programming tool are beneficial to student learning.

Since developing students' CT through active interactions between instructors and students is more difficult in large online than in small face-to-face classes, Paper 4 uses e-mentoring via social network services (SNS) in developing students' CT during large-scale online courses. The obtained results revealed that the most influential e-mentoring activities for students' CT development were informational and technical support in a group and informational support in a private environment. It was also found that female students benefited more from SNS-based e-mentoring than male students, and they also engaged in more types of e-mentoring activities than male students.

Inspired by the research evidence in the literature on the potential positive effects of reflection in complex CT problem-solving by regulating cognitive activities, Paper 5 designs a reflection-guided visualized mindtool strategy to address CT development challenges. Additionally, relying on the powerful insights that could be generated from behavioural analysis, it applies Lag Sequential Analysis (LSA) to analyse student's learning behaviours of CT. The results revealed that students who used the reflection-guided visualized mindtool strategy exhibited more key behaviours of facilitating CT problem-solving (e.g., generalizing the knowledge, redesigning the algorithm scheme, and evaluating the feasibility of their proposed schemes).

Finally, Paper 6 harnesses the power of learning analytics and game-based learning to develop a personalized educational game *Penguin Go* that could facilitate children's personalized learning experiences for K–5 computing education. It reveals that Sequential Data Analytics (SDA) can inform what in-game support is necessary to foster student learning and when to deliver such support in gameplay.

*Paper 1*: Integrating Computational Thinking into School Curricula of Compulsory Education: A Systematic Review of Recent Literature.

Authors: Panagiotis Kampylis, Valentina Dagienė, Stefania Bocconi, Augusto Chioccariello, Katja Engelhardt, Gabrielė Stupurienė, Vaida Masiulionytė-Dagienė, Eglė Jasutė, Chiara Malagoli, Milena Horvath and Jeffrey Earp.

*Paper 2*: A Normative Analysis of the TechCheck Computational Thinking Assessment. *Authors:* Emily Relkin, Sara K. Johnson and Marina U. Bers.

Paper 3: Integration of Computational Thinking with Mathematical Problem-based Learning: Insights on Affordances for Learning.

Authors: Zhihao Cui, Oi-Lam Ng and Morris Siu-Yung Jong.

*Paper 4*: The SNS-based E-mentoring and Development of Computational Thinking for Undergraduate Students in an Online Course.

Authors: Yeonju Jang, Seongyune Choi, Seonghun Kim and Hyeoncheol Kim.

*Paper 5*: Effect of a Reflection Guided Visualized Mindtool Strategy for Improving Students' Learning Performance and Behaviors in Computational Thinking Development. *Authors:* Xiao-Fan Lin, Wenyi Li, Jing Wang, Yingshan Chen, Zhaoyang Wang, and Zhong-Mei Liang.

*Paper 6*: A Framework for Applying Sequential Data Analytics to Design Personalized Digital Game-Based Learning for Computing Education. *Authors:* Zhichun Liu and Jewoong Moon

## 4. Conclusion and future research

This special issue revealed that developing CT curricula will facilitate the access and development of CT tools/technology that could be used in education, calling for more research in this regard, especially that each educational level has different subjects and knowledge to learn, while students in each educational level have different acquired skills.

It also revealed that while educational technology could enhance teaching and assessing students' CT, the focus should not be solely on the technology itself, but more on the educational approaches to be used with the technology. Therefore, there is a need for developing principles and guidelines about the best practices of using educational technology for enhancing students' CT. In particular, more research is needed to investigate the effective and responsible use of technology in CT education. In this context, smart learning analytics could empower both teachers and students by, for instance, revealing students' learning trajectories while addressing a particular CT topic, concept, or practice. Therefore, future research could focus on this direction, as well as on harnessing the power of big data and AI to promote the effective and safe learning of CT.

Finally, this special issue highlights that students' individual differences (e.g., age, gender, competency) should be considered when learning CT using educational technology. In this context, precision education, which aims to detect students' individual difference, could provide opportunities to overcome the one-size-fits-all approach and provide personalized experiences. In this context, Yang et al. (2023, p. 97) stated that "through precision education, teachers can understand students' learning situations by diagnostic system, extract data and establish a learning prediction model, then design adaptive learning activities for different types of students with one-of-a-kind treatment and prevention." Future studies could also investigate this line of research.

# Acknowledgment

The guest editors would like to express their gratitude to Prof. Yu-Ju Lan and Prof. Maiga Chang for their support and guidance on this special issue. Given the high volume of submissions, this special issue runs with an acceptance rate lower than 25%, and only a few papers were eventually accepted and included. We also thank the authors and reviewers for their valuable contribution and support towards the success of this special issue.

# References

Berland, M., & Wilensky, U. (2015). Comparing virtual and physical robotics environments for supporting complex systems and computational thinking. *Journal of Science Education and Technology*, 24, 628e647. http://dx.doi.org/10.1007/s10956-015-9552-x

Bers, M. U. (2020). Coding as a playground: Programming and computational thinking in the early childhood classroom. Routledge. https://doi.org/10.4324/9781003022602

Cansu, F. K., & Cansu, S. K. (2019). An overview of computational thinking. *International Journal of Computer Science Education in Schools*, 3(1), 17-30. https://doi.org/10.21585/ijcses.v3i1.53

Chazelle, B. (2006). Could your iPod be holding the greatest mystery in modern science? Math Horizons, 13(4), 14-31.

Cuny, J., Snyder, L., & Wing, J. M. (2010). *Demystifying computational thinking for non-computer scientists* [Unpublished manuscript]. http://www.cs.cmu.edu/~CompThink/resources/TheLinkWing.pdf

Czerkawski, B. C., & Lyman, E. W. (2015). Exploring issues about computational thinking in higher education. *TechTrends*, 59, 57-65. https://doi.org/10.1007/s11528-015-0840-3

de Jong, I., & Jeuring, J. (2020). Computational thinking interventions in higher education: A Scoping literature review of interventions used to teach computational thinking. In *Proceedings of the 20th koli calling international conference on computing education research* (pp. 1-10). https://doi.org/10.1145/3428029.3428055

Denning, P. J. (2007). Computing is a natural science. Communications of the ACM, 50(7), 13-18.

Fagerlund, J., Häkkinen, P., Vesisenaho, M., & Viiri, J. (2021). Computational thinking in programming with Scratch in primary schools: A Systematic review. *Computer Applications in Engineering Education*, 29(1), 12-28. https://doi.org/10.1002/cae.22255

Grover, S., & Pea, R. (2018). Computational thinking: A Competency whose time has come. *Computer science education: Perspectives on teaching and learning in school*, *19*(1), 19-38.

Hamilton, M., Clarke-Midura, J., Shumway, J. F., & Lee, V. R. (2020). An Emerging technology report on computational toys in early childhood. *Tech. Know Learn*, 25(1), 213–224. https://doi.org/10.1007/s10758-019-09423-8

Ioannidou, A., Bennett, V., Repenning, A., Koh, K. H., & Basawapatna, A. (2011, April). *Computational thinking patterns* [Paper presentation]. Annual meeting of the American educational research association, New Orleans, LA.

Israel, M., Pearson, J., Tapia, T., Wherfel, Q., & Reese, G. (2015). Supporting all learners in school-wide computational thinking: A Cross-case qualitative analysis. *Computers & Education*, 82, 263e279. http://dx.doi.org/10.1016/j.compedu.2014.11.022

Kjällander, S., Mannila, L., Åkerfeldt, A., & Heintz, F. (2021). Elementary students' first approach to computational thinking and programming. *Education Sciences*, *11*(2), 80. https://doi.org/10.3390/educsci11020080

Kite, V., & Park, S. (2020). Secondary science teachers' conceptualizations of computational thinking and perceived barriers to CT/content integration. In *Prepared for the 2020 annual meeting of the National Association for Research in Science Teaching and teacher education (NARST)*.

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. https://doi.org/10.1007/s10956-019-09803-w

Li, Y., Schoenfeld, A. H., diSessa, A. A., Graesser, A. C., Benson, L. C., English, L. D., & Duschl, R. A. (2020). Computational thinking is more about thinking than computing. *Journal for STEM Education Research*, *3*, 1-18. https://doi.org/10.1007/s41979-020-00030-2

Oberkampf, W. L., & Roy, C. J. (2010). Verification and validation in scientific computing. Cambridge University Press.

Papert, S. (1980). Mindstorms. Basic Books.

Pugnali, A., Sullivan, A., & Bers, M. U. (2017). The Impact of user interface on young children's computational thinking. *Journal of Information Technology Education. Innovations in Practice*, *16*, 171. https://doi.org/10.28945/3768

Rich, P. J., Browning, S. F., Perkins, M., Shoop, T., Yoshikawa, E., & Belikov, O. M. (2019). Coding in K-8: International trends in teaching elementary/primary computing. *TechTrends*, 63(3), 311–329. https://doi.org/10.1007/s11528-018-0295-4

Rosenbloom, P. S. (2013). On Computing: The Fourth great scientific domain. The MIT Press.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142-158. https://doi.org/10.1016/j.edurev.2017.09.003

Tlili, A., & Burgos, D. (2022). Unleashing the power of Open Educational Practices (OEP) through Artificial Intelligence (AI): Where to begin? *Interactive Learning Environments*, 1-8. https://doi.org/10.1080/10494820.2022.2101595

UNESCO. (2021). UNESCO prepares teachers and learners for 21<sup>st</sup> century challenges. https://en.unesco.org/news/unesco-prepares-teachers-and-learners-21st-century-challenges

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33-35. https://doi.org/10.1145/1118178.1118215

Yang, K., Liu, X., & Chen, G. (2020). The Influence of robots on students' computational thinking: A Literature review. *International Journal of Information and Education Technology*, *10*(8), 627-631.

Yang, S. J., Ogata, H., & Matsui, T. (2023). Guest editorial: Human-centered AI in education: Augment human intelligence with machine intelligence. *Educational Technology & Society*, 26(1), 95-98.

Zhang, X., Tlili, A., Guo, J., Griffiths, D., Huang, R., Looi, C. K., & Burgos, D. (2023). Developing rural Chinese children's computational thinking through game-based learning and parental involvement. *The Journal of Educational Research*, 1-16. https://doi.org/10.1080/00220671.2023.2167798

Kampylis, P., Dagienė, V., Bocconi, S., Chioccariello, A., Engelhardt, K., Stupurienė, G., Masiulionytė-Dagienė, V., Jasutė, E., Malagoli, C., Horvath, M., & Earp, J. (2023). Integrating Computational Thinking into Primary and Lower Secondary Education: A Systematic Review. *Educational Technology & Society*, 26(2), 99-117. https://doi.org/10.30191/ETS.202304\_26(2).0008

# Integrating Computational Thinking into Primary and Lower Secondary Education: A Systematic Review

# Panagiotis Kampylis<sup>1</sup>, Valentina Dagienė<sup>2</sup>, Stefania Bocconi<sup>1\*</sup>, Augusto Chioccariello<sup>1</sup>, Katja Engelhardt<sup>3</sup>, Gabrielė Stupurienė<sup>2</sup>, Vaida Masiulionytė-Dagienė<sup>2</sup>, Eglė Jasutė<sup>2</sup>, Chiara Malagoli<sup>1</sup>, Milena Horvath<sup>3</sup> and Jeffrey Earp<sup>1</sup>

<sup>1</sup>National Research Council, Italy // <sup>2</sup>Vilnius University, Lithuania // <sup>3</sup>European Schoolnet, Belgium // panagiotis.kampylis@itd.cnr.it // valentina.dagiene@mif.vu.lt // stefania.bocconi@itd.cnr.it // augusto@itd.cnr.it // katja.engelhardt@outlook.com // gabriele.stupuriene@mif.vu.lt // vaida.masiulionyte-dagiene@mif.vu.lt //

egle.jasute@fsf.vu.lt // chiara.malagoli@itd.cnr.it // milena.horvath@eun.org // jeffrey.earp@itd.cnr.it \*Corresponding author

**ABSTRACT:** In recent years, many countries have introduced Computational Thinking (CT) concepts into compulsory education as part of general curriculum reform efforts. A systematic review of academic and grey literature has been conducted to analyse the state of the art in implementing CT in primary and secondary education. In total, 1977 publications were identified, out of which 98 met the inclusion criteria for the review. The results show that, despite a lack of consensus on a common definition, a core set of key CT skills is addressed in primary and lower secondary education. Implementation approaches that emerged from the analysis are discussed and presented according to the European Commission's Joint Research Centre (2016) classification: (i) embedding CT across the curriculum as a transversal theme/skill set; (ii) integrating CT as a separate subject; and (iii) incorporating CT skills within other subjects such as Mathematics and Technology. New approaches to formative assessment of CT are emerging, reflecting different conceptualisations and differences in contextual and motivational aspects of CT curriculum integration. However, further investigation is needed to understand better how gender/equity/inclusion issues impact the quality of computing education integration.

Keywords: Computational thinking, Computer Science education, Compulsory education, CT skills

# **1. Introduction**

The Digital Education Action Plan 2021-2027 (European Commission, 2020b, p. 94) provides the following glossary definition of CT (along with programming and coding):

Computational thinking, programming and coding are often used in an interchangeable way in education settings, but they are distinct activities. Programming refers to the activity of analysing a problem, designing a solution and implementing it. Coding means implementing solutions in a particular programming language. Computational thinking, shorthand for "thinking as a computer scientist," refers to the ability to understand the underlying notions and mechanisms of digital technologies to formulate and solve problems.

This conceptualisation reflects the definition proposed by Wing (2017, p. 8): "Computational thinking is the thought processes involved in formulating a problem and expressing its solution(s) in such a way that a computer — human or machine — can effectively carry out." This conceptualisation is clearly connected to the concepts and practices of Computer Sciences - CS (i.e., thinking like a computer scientist) proposed as an intellectual framework for thinking. Caeli and Yadav (2020) provided the historical perspectives of CT and how initiatives today can inspire students to learn CS. A deeper historical development of CT and the intellectual ideas for development are provided by Tedre and Denning (2016).

In this paper, we use the term *Computer Science* (CS) interchangeably with *Computing* and *Informatics*. In Wing's view (2017, p. 7), "computational thinking will be a fundamental skill - just like reading, writing, and arithmetic - used by everyone by the middle of the 21st Century." This vision has been widely accepted as a basis for including CT as a key 21st century competence in compulsory education (e.g., Pérez-Marín et al., 2020). However, discussions on understanding of CT and its importance in schools have continued. As Curzon et al. (2019) note, this debate can be represented as positioning CT applicability on a scale that goes from the broad (e.g., the CT skillset overlaps with skills in other disciplines and is generally useful; computational systems exist in the natural world) to the narrow (e.g., CT is not necessarily beneficial for everyone; computational systems are confined to computers). However, most CT definitions in the literature position it somewhere between these two extremes (e.g., CT is also relevant to non-technical disciplines). Further support for this interpretation comes

from the OECD's PISA and IEA's ICILS educational assessments, which explicitly refer to CT and include key CT skills and concepts in their tests (e.g., Fraillon et al., 2019).

Generally, CT is regarded in the literature as a thought process involved in designing solutions that a computer, a human, or both can execute. While numerous definitions are currently being proposed and adopted, an agreement is nonetheless emerging on core CT concepts, namely abstraction, algorithmic thinking, automation, decomposition, and generalisation (Curzon et al., 2019). These are closely connected with a set of attitudes and skills (named as practices) that comprises creating computational artefacts, testing and debugging, collaboration and creativity, and the capacity to tackle open-ended problems (Grover & Pea, 2018). In this view, CT is seen as a fundamental competence for informed citizens to manage the ever-emerging challenges society poses. In addition, CT offers the potential to support creative problem-solving and may foster innovative approaches in other subject areas. Hence, it undoubtedly has a key role to play in compulsory education. By making CT concepts concrete, programming provides opportunities for CT education. This renders it a tool for learning, e.g., a way of exploring other domains or a means for self-expression (Resnick, 2017). Nevertheless, it is generally agreed that coding/programming is just one of the various facets of CT.

This paper is intended to provide an overview of the integration of CT skills in primary and lower secondary education, as emerging from recent research works. The systematic literature review presented herein builds upon and complements that presented by Bocconi et al. (2016), capturing developments in the CT field from 2016 to 2021. We use the term *Computer Science* (CS) for both Computing and Informatics, in line with the Digital Education Action Plan 2021-2017 (European Commission, 2020a, p. 13).

To understand how the ongoing debate and results from research on CT are influencing the implementation of CT-related curricula in schools, the following research questions were addressed:

- RQ1: How is CT defined in the context of compulsory education?
- RQ2: How is CT implemented in primary and lower secondary education?
- RQ3: How are gender and equity addressed when implementing CT in the curriculum?

Section 2 presents the methodology adopted for conducting the systematic literature review. In Section 3, we present and discuss the review results, with particular attention devoted to CT definitions in various settings, the curricular issues intertwined with CS education, gender balance and equity, pedagogical approaches adopted, and technologies employed. Some conclusions are provided in the final section.

# 2. Method

## 2.1. The PRISMA 2020 statement

A structured approach was employed to identify relevant academic and grey literature and select the publications to be analysed in-depth. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses – PRISMA 2020 statement (Page et al., 2021) was applied to increase the dependability and reliability of the data collected and analysed. Figure 1 presents the PRISMA 2020 workflow steps followed (Identification, Screening, Included) and the number of records handled in each step.

## **2.2. Identification process**

The initial search was conducted on the Scopus research database in May 2021, focusing on a selection of 24 top-tier journals and conference proceedings (Table 1) devoted to pedagogical aspects of CT and CS education.

A broad coverage of studies was sought in searching for CT occurrences and related terms. Accordingly, the following Boolean string was used to identify relevant papers published after 2016 in the selected journals: (ISSN (XXXX-XXXX) AND ALL ("computational thinking") OR ALL ("algorithmic thinking") OR ALL ("computer science education") OR ALL ("computing education") OR ALL ("informatics education") AND PUBYEAR > 2015). For the conference proceedings, the following search string was employed: (ALL ("computational thinking") OR ALL ("algorithmic thinking") AND PUBYEAR > 2015). The search string for identifying conference papers did not include the search terms "computer science education," "computing education," or "informatics education" because the selected conference proceedings are specialised in these fields, in contrast with the selected journals, which are wider in their scope.

Туре	Source	Identifie	Identified and screened			
		Step 1	Step 2	Step 3		
Journals	Education and Information Technologies	158	27	2		
	Computers & Education	135	26	13		
	ACM Transactions on Computing Education	133	25	2		
	Computer Science Education	111	22	3		
	Computers in Human Behavior	87	19	3		
	Journal of Educational Computing Research	80	15	5		
	Informatics in Education	63	14	8		
	IEEE Transactions on Education	61	7	0		
	International Journal of Child-Computer Interaction	53	16	0		
	Journal of Computer Assisted Learning	45	11	2		
	TechTrends	36	7	4		
	Journal of Research on Technology in Education	21	10	2		
	Thinking Skills and Creativity	12	1	0		
	Comunicar	7	1	0		
	Journal papers citing CompuThink 2016 study	122	23	2		
	Journal papers from author tracing and reference mining	2	2	2		
Edited books	Books chapters citing CompuThink 2016 study	17	4	2		
	Book chapters from author tracing and reference mining	3	3	3		
Conference	SIGCSE: Symposium on CS Education	327	120	11		
proceedings	ITiCSE: Innovation and Technology in CS Education	80	28	0		
	ISSEP: Informatics in School Education: Evolution and	63	29	6		
	Perspectives					
	WIPSCE: Workshop in Primary and Secondary Computing	60	33	2		
	Education					
	Koli Calling Conference on Computing Education Research	48	15	4		
	LaTiCE: Learning and Teaching in Computing and	8	5	0		
	Engineering					
	Conference papers citing Computhink 2016 study	50	8	0		
	Conference papers from author tracing and reference mining	2	2	2		
Grey literature	Grey literature citing ComputThink 2016 study	73	18	7		
	Grey literature from author tracing and reference mining	120	25	13		
	Total	1977	516	98		

Table 1. Number of publications identified, collected, and screened per source

Figure 1. PRISMA 2020 Flow Diagram for systematic literature review. Adapted from Page et al. (2021)



In addition to the Scopus search, we employed citation tracking to identify post-2016 academic and grey literature, which has become pivotal in CT research and computing education. These works include conceptually oriented and empirical studies that have (a) generated a line of investigation which has changed how problems or questions have been framed, (b) introduced new methods or concepts, or (c) generated influential debate. We also employed Google Scholar to identify and gather publications that cite the systematic literature review by Bocconi et al. (2016), referred to as the *CompuThink 2016* study.

#### 2.3. Screening and eligibility process

The screening process was performed in three steps. In Step 1, the 1977 aggregated publications were allocated among the nine researchers involved in the literature review. Screening of records involved reading the title and abstract and applying three exclusion criteria: (i) not a full article; (ii) not devoted to compulsory education; (iii) devoted to specialised topics (e.g., cybersecurity, machine learning, data analytics) outside the scope of the study.

In Step 2, the full texts of the 516 records potentially eligible for in-depth analysis were collected and subsequently screened by applying the following exclusion criteria: (i) tangential or no specific focus at all on CT/CS; (ii) pilot studies of low quality and/or conducted on small sample size; (iii) empirical papers reporting outcomes not explicitly concerning CT/CS.

In Step 3, the 98 publications (see Table 1) from academic (N = 78) and grey literature (N = 20) were distributed among the nine researchers and analysed in-depth through a review matrix (see Appendix 1) following specific guidelines and a shared understanding of the different fields, as described in the following section.

#### 2.4. Quality assurance

To ensure that the researchers analysed and coded the selected literature homogeneously and comparably, Inter-Rater Reliability (IRR) checks were carried out as a measure of quality assurance. Furthermore, guidelines were drafted for researchers to follow throughout the PRISMA screening process and for mapping against the review matrix.

To precisely evaluate IRR for multiple non-unique coders, an extension of the k statistic was used, as propounded by Hallgren (2012). Cohen's kappa was calculated through SPSS version 26.0, and the arithmetic mean of these outputs was computed. This process determined the level of agreement among all the researchers on whether 24 publications of different nature, including academic and grey literature, could be included or excluded from the literature review based on guidelines for the PRISMA screening process.

To avoid influencing the other researchers' decisions, each researcher rated the same set of 24 randomly selected publications individually. When evaluating publications, each researcher could choose only to include or exclude them for screening in Step 3. Researchers were asked to perform two rounds of evaluation. Overall, in round one, a moderate level of coder agreement was reached (Mean  $\kappa = .41$ ). In line with the approach proposed by Belur et al. (2018), this first round of coding was followed by an open discussion among members of the coding team led by two senior researchers/coders. This discussion permitted clarification of the inclusion/exclusion criteria, resulting in an excellent level of agreement in the second round (Mean  $\kappa = .98$ ).

# **3. Results and discussion**

Following the process described in the previous section, the 98 publications from 2016 to 2021 were analysed in depth. In terms of temporal spread, a significant increase in 2020 emerges (see Figure 2).

A thorough analysis of the collected works was carried out through a review matrix approach (see Appendix 1), thus facilitating a structured comparison of different sources. The research questions in the matrix are broader than those addressed in this paper, as the matrix was designed for comprehensive extraction and documentation of all the insights from each publication, thereby providing a solid basis for systematic analysis.



# Figure 2. Distribution of analysed literature by year of publication

# 5 8 2016 2017 2018 2019 2020 2021

#### 3.1. CT definition and concepts in different settings

Although this field has been the subject of intensive research for about 15 years, there is still debate on the definition of CT. Different research teams have tended to expound their own CT definition informed by their specific line of inquiry and have assumed different perspectives regarding applying, interpreting, and assessing proposed CT concepts. Taslibeyaz et al. (2020) note that CT definitions are often context-specific.

Tikva and Tambouris (2021) categorise CT definitions as domain-specific or domain-general. Domain-specific definitions indicate domain-specific knowledge or skills needed to solve problems systematically in the subject area of CS or programming. Domain-general definitions refer to competences necessary for solving problems systematically in all learning activities. This framework is similar to that proposed by Tang et al. (2020), which divides CT definitions related to (i) programming and computing concepts, and (ii) competences.

The three-type categorisation of CT definitions (i.e., generic, operational and educational / curricular definitions) proposed by Román-González et al. (2017) is used below to present examples of CT definitions emerging from the analysed literature (see Table 2).

Those who offer a precise definition of CT agree that it is a mode of thinking (thought process) for problemsolving (Grover & Pea, 2018; Hazzan et al., 2020; Zhang & Nouri, 2019). However, whatever one's view on the definition of CT, it is important to be pragmatic regarding the best ways of teaching it (Curzon et al., 2019). What all definitions have in common is that CT is more than problem-solving: the problem's solution must be expressed in such a way that permits it to be executed by a computational agent. CT is the way of thinking for developing solutions that allow a processing agent (machine) to carry it out (Corradini et al., 2017; Curzon et al., 2019; Csizmadia et al., 2019).

According to Fessakis and Prantsoudi (2019), CT-related skills commonly cited in various definitions are: algorithmic approach to problem-solving (including creativity), abstraction, logical reasoning, problem-solution transfer, generalisation, processing of data, and social impact of computation. In addition, Csizmadia et al. (2019) suggest the combination of CT with constructionism for selecting and evaluating classroom activities.

Corradini at al. (2018) classified all constitutive elements of CT into four categories: (i) mental processes or strategies useful to solve problems; (ii) methods, i.e., operational approaches used by computer scientists; (iii) practices used in the implementation of computer-based solutions; and (iv) transversal skills, e.g., general skills enhanced by CS application.

In a systematic review of empirical studies (Tang et al., 2020), the authors analyse well-cited CT definitions and notice that many are related to programming and computing. Tikva and Tambouris (2021, p. 162) observe a reciprocal association between CT and programming: "programming supports the development of CT while CT provides to programming a new upgraded role." However, Hazzan et al. (2020, p. 61) summarise that "CT is not necessarily about programming, but rather, the emphasis is on problem-solving," which fosters learning experiences. Nevertheless, programming is still the most frequently mentioned concept taught and, as Upadhyaya et al. (2020) remark, programming coupled with abstraction is becoming more commonly mentioned in conjunction with CT skills. CT is often conceptualised in a programming context and can be examined in terms of three key components: CT concepts, CT practices, and CT perspectives (Kong et al., 2020).

#### Webb et al. (2017, p. 449) state:

The distinction between computational thinking and programming is subtle; in principle computational thinking does not require programming at all, although in practice, representing a solution to a problem as a

	Table 2. The three types of CT definitions with examples
CT definition	Examples of CT definitions in the analysed literature
categories	
Generic definition	CT is the thought process entailed in formulating a problem and expressing the solution(s)
	so that a computer-human or machine can perform it effectively (Grover & Pea, 2018;
	Rich et al., 2021).
	CT regards thinking processes, so its implementation is independent of technology (Hazzan
	et al., 2020).
	CT is a thought process involving fundamental programming skills (CT skills) for solving
	problems in any domain (Zhang & Nouri, 2019).
Operational or	The CT framework employs fundamental CS concepts to solve problems, design systems,
model definition	and understand human behaviour (Jocius et al., 2020).
	CT encompasses a set of broadly applicable problem-solving skills that include abstraction,
	algorithmic thinking, decomposition, and pattern recognition (Huang & Looi, 2020).
	CT is a means to understand and solve complex problems by using CS concepts and
	techniques such as decomposition, pattern recognition, abstraction, and algorithms (Kale
	et al., 2018).
	Eight aspects at CT's core are highlighted: abstraction, algorithm design, evaluation,
	generalisation, iterative improvement, information representation, precise
	communication, and problem decomposition (Komm et al., 2020).
	CT definition relates to the operationalisation of CT practices (mainly based on Zoombinis
	gameplay) and focuses on four CT practices: problem decomposition, pattern
	recognition, abstraction, and algorithm design (Asbell-Clarke et al., 2021).
	CT is defined as a conceptual framework based on the five fundamental CT concepts:
	abstraction, decomposition, algorithmic thinking, evaluation, and generalisation (Tsai et
	al., 2020).
	The computing-based CT definition framework divided CT into general practices such as
	data, modelling & simulation, computational problem-solving, and systems thinking
	(Weintrop & Wilensky, 2019).
Educational and	CT entails addressing problem-solving systematically (e.g., algorithmically) so that the
curricular	solutions generated can be reused in various contexts (Shute et al., 2017).
definitions	According to the Australian Curriculum, Assessment and Reporting Authority, CT is a
	problem-solving approach entailing various strategies and techniques that can be
	implemented using computer systems (Australian Computing Academy, 2019).
	CT is considered a means to develop knowledge and understand concepts in CS and
	contributes significantly to general problem-solving skills (Israel-Fishelson &
	Hershkovitz, 2020).
	CT encompasses four different computational practices (problem-solving or algorithmic
	thinking, building algorithms, debugging, and simulation) and some concepts (Hooshyar
	et al., 2020).
	Thinking computationally means employing CS principles and methods to efficiently
	address and solve problems (Arfé et al., 2020) and developing algorithmic solutions to
	those problems so they can be operationalised using computers (Eickelmann et al., 2019).

program provides a perfect way to evaluate the solution, as the computer will execute the instructions to the letter, forcing the student to refine their solution so that it is very precise.

Besides the theoretical discussion on CT definition, many studies investigate CT integration in classrooms. An operational definition of CT skills (see examples in Table 2) is more suitable for everyday activities and is broadly adopted in many studies (e.g., Barendsen et al., 2016; Grgurina et al., 2018; Leonard et al., 2021). Also, it is important to mention the historical perspectives of CT, which have a strong connection to computing (Caeli & Yadav, 2020; Tedre & Denning, 2016).

#### 3.2. CT in the primary school curriculum

In a world where computing is pervasive, "CT is being recognised as a foundational competency for being an informed citizen and being successful in all STEM work, and potential for creative problem solving and innovating in all other disciplines" (Grover & Pea, 2018, p. 34).

Most researchers agree that programming should best be taught from a young age (Ching et al., 2018; Niemelä et al., 2017; del Olmo-Muñoz et al., 2020; Sáez-López et al., 2016; Wei et al., 2021). Usually, primary students are initially introduced to programming via unplugged activities, i.e., working without a computer or another digital device (del Olmo-Muñoz et al., 2020; Tonbuloğlu & Tonbuloğlu, 2019) and then move on to the use of block-based programming languages with computers (Arfé et al., 2020; Sáez-López et al., 2016; Sherwood et al., 2021). Related investigations have yielded solid evidence of the positive potential and results of developing CT when young learners program with different educational technologies (Ching et al., 2018; Kong et al., 2020). González-González et al. (2019) presented a study carried out with children aged of 3 to 6 years old with Down syndrome and showed that these pupils with cognitive disabilities can acquire basic programming and CT skills using tangible robots such as KIBO. Israel et al. (2020) examine how elementary students with autism behaved during computing instruction and concluded that these students require individualised support. Wei et al. (2021) report that partial pair programming effectively impacts the development of CT skills and self-efficacy in primary school students. Wu and Su (2021) have noticed that learning through physical robots can help students improve their CT abilities.

The primary education studies among the selected papers deal with three aspects: (i) programming in CS; (ii) programming in other disciplines like science, mathematics, art or integrated subjects; and (iii) programming used as a tool to assess CT concepts and skills. Generally, programming skills are developed in CS classes rather than in science or integrated disciplines. Based on the theoretical framework of programming-related CT proposed by Brennan and Resnick (2012), we investigated the selected papers and extracted examples of computational concepts, practices, and perspectives for primary education (see Table 3).

	FFFFFFFFF
Key CT	Examples and references
dimensions	
Computational	Sequences, directions (forward, back, left, right) and loops (del Olmo-Muñoz et al., 2020)
concepts	Algorithms, automation, coordination, creativity, data, logic, modelling and design, patterns, and problem decomposition (Fagerlund et al., 2020)
	Algorithmic thinking, creativity, collaboration, critical thinking, and problem-solving (Tonbuloğlu & Tonbuloğlu, 2019)
	Parameters, passing parameters to subprograms, sequences, simple loops, repeat, variables (Hromkovič & Lacher, 2017)
	Data, sequences, loops, parallelism (Israel-Fishelson & Hershkovitz, 2020)
	Abstraction and connected concepts (Liebe & Camp, 2019; Statter & Armoni, 2020)
	The seven big ideas: creativity, abstraction, data, algorithms, programming, internet, and
	global impact (Repenning et al., 2021)
Computational	Abstracting and modularising, algorithmic thinking, data, computational practices,
practices	experimenting and iterating tests, reusing and remixing, testing and debugging (Basu et al., 2020; Román-González et al., 2017)
	Abstracting, algorithm design, pattern recognition, and problem decomposition (Asbell-
	Clarke et al., $2021$ : Rijke et al., $2018$ : Wu & Su, $2021$ )
	Algorithmic thinking, building algorithms, and debugging (Jocius et al., 2020)
	Coding, conditionals and testing, looping, functions or debugging, nested looping, and
	sequencing (Arfé et al., 2020; Israel-Fishelson et al., 2021)
	Data representation, flow control, parallelisation, and user interactivity (Wei et al., 2021)
Computational	Better understanding of the initialisation of variables and objects (Franklin et al., 2016)
perspectives	

			• • •
Table 3 Examples of C Projected concepts	nracticae and	norchoctives in	nrimary adjugation
<i>Tuble J.</i> Examples of CI-fefalled concepts	, practices, and	DUISDUULIVUS III	
	, ,		

Games are commonly used in teaching and learning in primary schools. Several studies examined the relationship between student gameplay and the development of students' CT practices using different classroom activities (Asbell-Clarke et al., 2021; Ching et al., 2018; Hooshyar et al., 2020; Israel-Fishelson & Hershkovitz, 2020).

To reinforce CT skills in the primary grades, unplugged activities in combination with constructionist approach to learning-to-think computationally are particularly helpful (Caeli & Yadav, 2020). Also, Csizmadia et al. (2019) developed and presented a new mapping tool (the constructionism matrix) to review classroom activities in terms of both CT and constructionist learning. Pérez-Marín et al. (2020) suggested a metaphor-based methodology using Scratch to teach primary school children basic algorithmic and programming concepts. They found a statistically significant increase in participants' results and concluded that it is possible to teach children basic computer programming concepts (conditionals, loops) and improve their CT skills.

#### 3.3. CT in the lower secondary school curriculum

While in primary school CT skills are usually integrated into other subjects in primary school rather than being embedded in a specific dedicated subject, CT plays a much more complex role in secondary education. From their analysis of educational policy initiatives across the globe, Hsu et al. (2019) describe trends and rationales for including CT in the curricula. They highlight four main approaches: (i) creating a new subject; (ii) incorporating CT skills into existing subjects such as maths; (iii) embedding CT across the entire curriculum as a transversal skill set; (iv) combining the above. In many countries, introducing a new subject or even content area represents a serious challenge, as it usually entails making space in densely packed curricula. Rich, Mason and O'Leary (2021) developed the BootUp curriculum by gradually introducing coding and CT concepts and practices that are based on the K-12 CS Framework.

Jocius et al. (2020, p. 6) point out that "the value of computational thinking is not just as an isolated concept that relates to Eickelmann et al. (2019), but also as a way to enhance and support more complex discipline-specific and interdisciplinary understandings."

There is a strong emphasis on programming in CT skills development in compulsory education, especially in the secondary curriculum. Many researchers argue that programming improves understanding of CT concepts and contributes to CT practice (del Olmo-Muñoz et al., 2020; Wu & Su, 2020). Sáez-López et al. (2016) show that active pedagogical methods employing visual programming languages yield significant benefits for aspects like learning programming concepts and developing logic and computational practices. Researchers point out that programming stimulates the development of students' CT skills (Djambong et al., 2018; Corradini et al., 2018) and can be successfully employed for teaching CT in both primary and secondary schools (Hsu et al., 2019; Israel-Fishelson et al., 2021; Yağcı, 2019). Many secondary school teachers use program design courses to foster students' CT skills (Pasternak, 2016; Sáez-López et al., 2016; Zhang et al., 2020). Coenraad et al. (2021) designed Scratch Encore, a curriculum that uses Scratch and follows the Use-Modify-Create pedagogical strategy to introduce secondary school students to CS concepts in a culturally responsive way.

An increasing number of literature reviews published in the last few years investigate the broad association of CT skills with programming (Ching et al., 2018; Djambong et al., 2018; Hsu et al., 2019; Sun et al., 2021b; Zhang & Nouri, 2019). Several of these reviews concentrate on particular facets of the CT domain, such as programming; others address a spectrum of topics (e.g., Hsu et al., 2018; Ching et al., 2018). Often, teachers are expected to develop students' CT skills through program design courses based on structured assessment because algorithms and programming are recognised as ways to improve those skills (Sáez-López et al., 2016; Tikva & Tambouris, 2021; Román-González et al., 2019).

Numerous ongoing initiatives seek to foster CT skills by providing learners with programming tools and resources intended to facilitate the integration of CT in schools and thereby respond to societal needs for 21st-century skills (Hsu et al., 2019; Passey, 2017). Grover et al. (2019) developed a suite of non-programming digital activities embedded in a curriculum before students engage in Scratch block-based programming they use constructivism as a pedagogic approach. Tikva and Tambouris (2021) developed a conceptual model of CT within programming for K-12 education that is based on a systematic literature review summarising 101 studies and identifying CT areas. This model adopts a challenging holistic approach, seeking to support CT teaching and learning in K-12 education. Dagli and Sancar Tokmak (2021) emphasise development of students' CT skills through instructional design stages: analysis, design, development, implementation, evaluation and revision.

Palts and Pedaste (2020) propose an innovative model for developing CT skills This is based on three stages in CT training: (i) define the problem, including formulation and reformulation, abstraction, and decomposition; (ii) solve the problem by using data analysis, algorithm design, parallelisation, iteration and automation; and (iii) analyse the solution based on generalisation, testing and evaluation.

#### Grover and Pea (2018, p. 34) state that:

[...] learning CT, much like learning scientific and mathematical thinking, is more about developing a set of problem-solving heuristics, approaches and 'habits of mind' than simply learning how to use a programming tool to create computational artefacts.

Li (2020) stresses that good design for developing CT skills includes several important aspects, such as scaffolding, learner-centred methods, and fostering deep learning by integrating CT skills in teaching coding.

## **3.4. Integrating CT as a separate subject**

CT is increasingly becoming a central focus as countries update their school curricula. UNESCO and IFIP TC3 (2019) recommend promoting CT through the CS curriculum based on the understanding that CT is integral to problem-solving approaches in CS. Fessakis and Prantsoudi (2019) remind us that CT was introduced as a conceptual tool to promote the status of CS in general education. Accordingly, CT enables students to leverage advancements made possible by CS, from data collection to in-depth research. One important goal of today's CS curricula is improving students' CT skills (Kert et al., 2019). Tomokyio (2018) explored successes and challenges in implementing a progressive CS curriculum for K-8 schools. A new comprehensive curriculum that integrates various CS topics in middle schools in Qatar was elaborated and discussed by Razak et al. (2021). Forlizzi et al. (2018) proposed a core CS (Informatics) curriculum for all the levels of compulsory school, and outlined strategies to ensure that its implementation in schools can be effective.

The Australian Computing Academy (2019, p. 30) states that:

The development of pedagogy in computer science education lags behind that of other subjects. In contrast to CS, mathematics has been taught at schools for centuries, and there is broad consensus about teaching key concepts at different year levels, taking into account the changing cognitive capabilities as students age.

The CS (Informatics) curriculum in the Netherlands can serve as an example of incorporating CT education (Barendsen et al., 2016). The core of the curriculum comprises a skill set (comprising not just CS-specific skills but also general scientific and technical skills), together with five knowledge domains. In addition, three skill subdomains are deemed crucial aspects characterising CS as a subject: (i) design and development; (ii) focus on informatics perspective; and (iii) collaboration and interdisciplinarity. A few years following the curriculum's introduction, Grgurina et al. (2018) developed a curriculum intervention including a practical assignment and an accompanying assessment instrument consisting of grading rubrics.

When applying a strategy for a CS curriculum integrated with CT, Hromkovič and Lacher (2017) espouse the following principles:

- CS must not be taught as an isolated subject but rather as a part of Science and Technology, thus providing an in-depth contextual view;
- Do not teach the use of the latest IT products or the latest scientific discoveries. Instead, investigate the evolution of fundamental concepts and their step-by-step development;
- Teach programming and automate well-understood activities with computers.

Hromkovič et al. (2016) envisioned a broad and comprehensive CS education from primary school by introducing Logo and then carefully building through secondary school using text-based programming languages such like Python.

Dealing with approaches for CT education, Li (2020, p. 10) emphasised that:

[...] experts suggest that although CT can certainly be taught through CS, it may not be the best approach for different reasons. First, CT [skills are] best learned when they are integrated into different subjects for elementary students. Secondly, even for secondary students, depending on the student group and school environments, integrating CT into different subjects may also prove to be the most practical and useful way.

Numerous recently published studies suggest that programming skills should be considered fundamental skills that are as important as reading and writing, so CT and programming are deeply intertwined (Metcalf et al., 2021). CT shifts the focus from programming and learning to code to areas like problem-solving in various disciplines using coding or other CT skills (Basu et al., 2020). Many researchers describe the connection between programming and CT in compulsory education in the context of a CS curriculum (Sun et al., 2021b; Tikva & Tambouris, 2021; Waite et al., 2020; Webb et al., 2017). CS itself ranges from the digital skills needed to *use* technology to advanced programming skills required to *design* that technology.

When discussing unplugged pedagogy as a way to support CS-for-all and CT development, Huang and Looi (2020) point to the plethora of unplugged activities that have been developed and adapted as proof that the approach offers flexibility and suitability for teaching a broad range of learners.

#### **3.5.** Integrating CT within other subjects

Given these practical considerations, integrating CT across subjects transversally offers considerable advantages (Balanskat et al., 2018). CT skills are not necessarily confined to development within a single subject but rather can be seen as a set of essential thinking skills applicable to any STEM-related field (Sun et al., 2021b).

Non-CS-focused approaches aim to integrate CT into learning experiences within subjects other than CS, such as STEM - Science, Technology, Engineering, and Mathematics (Li, 2020; Niemelä et al., 2017). Previous studies show a significant positive correlation between STEM education and CT skills (Hsu et al., 2018; Sun et al., 2021a). Using CT and coding to solve a mathematics or science problem is quite different from using CT in a language or art lesson (Australian Computing Academy, 2019).

One non-CS-focused approach is to adopt a cross-domain teaching mode, namely allowing students to work with materials from various domains via computing. This can enable them to deepen their understanding of cross-domain knowledge, experience how cross-domain knowledge and computing can help solve complex real-world problems, and foster interest in studying STEM (Hsu et al., 2018).

While a wide range of new technologies is available for age-appropriate CT development, more research is required to design and develop pedagogies for employing these tools effectively to foster young learners' CT skill development. Furthermore, integrating the development of CT skills with the acquisition of discipline-based content knowledge should help young learners to appreciate the real-world application of CT (Ching et al., 2018; Hsu et al., 2018).

Kale et al. (2018) suggested three strategies that can help teachers make the connections between CT and their teaching of other subjects in K-12 settings: (i) use content-specific examples, (ii) recognise the similarities between CT and the problem you need to solve, (iii) use methods of teaching problem-solving (e.g., modelling).

#### 3.6. Addressing gender and equity when implementing CT in the curriculum

Research on gender balance and equity in the CT field is scarce, and the findings from those studies investigating gender and CT skills are often contradictory (Tikva & Tambouris, 2021). Some studies (Atmatzidou & Demetriadis, 2016; del Olmo-Muñoz et al., 2020; Tsai et al., 2020; Witherspoon et al., 2017; Wu & Su, 2021) find no significant relationship between gender and the acquisition of CT skills, while others conclude that there *are* gender differences in the approach to learning CT (Labusch & Eickelmann, 2020; Román-González et al., 2017; Tomokiyo, 2018; Wei et al., 2021).

The nature of gender differences in CT may depend on the type of problems, tasks or activities proposed for acquiring such skills (Israel-Fishelson et al., 2021; Román-González et al., 2017). Another possible contribution to the diversity of results might also depend on the differences in tools and instruments (i.e., self-report vs practical activities) used to assess the variety of CT-related concepts, which may complicate the current state of the art. For instance, Tsai et al. (2020) found that boys self-report a significantly higher disposition for decomposition thinking than girls. In another study, Rijke et al. (2018) found that "after the age of 9.5 years old, female students begin to outperform their male peers on the abstraction task" (p. 85); when students reach fourth grade, girls are likely to outstrip boys on abstraction. Some authors report a significant interaction effect for gender and age, hypothesising a possible link between this specificity and a gender-related developmental trajectory in consolidating this ability. Given this scenario, the availability of additional and more challenging materials would be desirable, as would adaptation that considers gender differences in the light of different developmental stages. Guggemos (2021) demonstrates that "motivation, in the form of CT self-concept and self-determined motivation, plays an important role in explaining CT level and gender differences," (p. 12) with females showing lower CT self-concept, lower computer literacy, and lower self-determined motivation – findings which translate into a negative association with CT for females.

According to Sun et al. (2021a, p. 355), who conducted an empirical study to explore the association between students' STEM learning attitude and their CT skills through a self-report survey:

[the] learning attitude of girls from primary school towards STEM was generally more positive than that of boys in the same period, while for CT skills, although the gender difference was not significant, the score of girls was slightly higher than that of boys.

According to Kong et al. (2018), this could be explained by differences in the development of boys and girls at this stage. On the other hand, boys were more interested in programming than girls were, and so "teachers might

need to pay attention to the engagement level of girls and employ strategies to enhance their interest in programming" (Kong et al., 2018, p. 188). Some studies (e.g., Balanskat et al., 2018) point out that the integration of CT skills in elective subjects such as Technology, in which usually most of the enrolled students are male, can widen the gap both in terms of gender but also between those interested in programming and those who are not. To address this challenge, the best option appears to make it compulsory for all students to engage in courses where they can learn some basics about CT/programming. In elective subjects, a special effort could be put into recruiting girls and those students who do not think they can engage in programming courses (Balanskat et al., 2018).

Another important matter in this regard, and a possible source for differences in study results, is the variety of approaches and methodologies implied in CT skills support and integration, and the ultimate focus of these, which may also vary depending on cross-cultural differences (Upadhyaya et al., 2020). Leonard et al. (2021) investigated how students choreographing dance performances involving virtual dancers utilise embodied ways of thinking within CT concepts and found that dance provides opportunities for all young people to be engaged in programming. This study showed that, with such integration, students' CT test scores increased significantly and that their embodied thought processes allowed them to enact various computational and choreographic practices. Comparing computational creativity scores regarding personal characteristics has revealed some significant differences between girls and boys. In some cases, mean levels of computational creativity are more remarkable for girls than for boys: "girls were significantly more creative than boys in terms of both creative thinking and computational creativity" (Israel-Fishelson et al., 2021, p. 1436).

Additionally, teachers' role in engaging all students is vital for gender balance and equity. Cateté et al. (2020) state that successful professional development prepares teachers to acquire the skills to teach CS and teach diverse student cohorts with different ethnicities, socioeconomic backgrounds, and genders.

In terms of curriculum enactment, specifically dealing with compulsory CT education for all students, Hsu et al. (2019, p. 268) argue that:

[...] mandatory coursework stems both from the notion that CT is a foundational skill that all students should have to be digitally competent and be active participants in a world where computing is pervasive and from a desire to motivate interest in CS and STEM, especially among girls and underrepresented minorities.

According to Hsu at al. (2019), one example of such initiatives is Code.org in the United States, which aims to increase the participation of women and minorities by implementing CS in the core curriculum. In Ghana and Burkina Faso, the Teach Need Girls mentorship program aims to teach girls how to code and create technology. Additionally, broader participation in computer science, more content responding to the need for diversity in computing, and tackling issues such as equity or accessibility are all highlighted as priorities in the K–12 Computer Science Framework. In the case of schools that already have compulsory CT curricula, implementing additional enrichment programs can enhance interest and motivation, and lead to more in-depth learning. A case study conducted on this topic in the UK found that enrichment programs, for instance Teach Future Girls and Hour of Code, had a positive effect on students, especially on girls, when it came to continuing their studies in higher-level computing courses.

Conversely, the literature review on equity carried out by Huang and Looi (2020) highlights that "unplugged activities appear in curricula that are specifically designed for girls, students of colour, students with special needs, and students in low-income communities, but there have been no studies that theorised the rationale for their inclusion" (p. 97). On the one hand, overall results in this matter point to the variety of tools and approaches as a possible explanation for inconsistencies in results. On the other hand, the results support the need for additional investigations into this matter, accounting for increased consistency in approaches and tools (i.e., use of self-reporting tools in combination with practical tasks about CT). At the same time, results also indicate the need to avoid gender biasing the proposed activities (i.e., proposing girls perform different tasks with respect to boys as a baseline).

# 4. Conclusions

This paper has discussed significant developments in integration of CT in compulsory education between 2016 and 2021, mainly focusing on CT definitions and curriculum integration approaches as emerged from the analysis of outcomes from the reported literature review. A wider range of evidence regarding CT pedagogies, assessment and professional development of teachers in computing education was also collected through this systematic literature review, which contributed to and is discussed in the study by Bocconi et al. (2022).

Results from the review show we are reaching a plateau in the debate on CT definitions, which is now mainly focusing on the set of key constituent CT concepts, including abstraction, algorithmic thinking, automation, decomposition, debugging, and generalisation. These concepts are correlated with several attitudes and skills (or practices), including creating computational artefacts, testing and debugging, collaboration and creativity, and the capacity to address open-ended problems. From this perspective, CT can be framed as a fundamental competence for a well-informed citizen capable of facing the challenges society continues to pose. CT also offers considerable potential for creative problem-solving and the adoption of innovative approaches in several other subject areas. Therefore, it has a pivotal role to play in compulsory education.

Coding/programming provides a laboratory for teaching and learning CT, for making CT concepts concrete. It can also be a learning tool for investigating other domains or self-expression. However, it is generally agreed that CT entails more than coding or programming (Basu et al., 2020; Barendsen et al., 2016). Although all key CT concepts could in principle be addressed both in primary and lower secondary education, a learning progression is emerging from the actual integration of CT in the curriculum, where the study of programming and algorithms provide the basis for developing CT skills (e.g., Zhang et al., 2020). Starting from sequences of instructions, iterations, conditionals and use of variables in primary school and proceeding to conditionals, operators and data structures in lower secondary. Particular attention is devoted to well-known learner difficulties in primary such as variables initialisation (Franklin et al., 2016), and the challenges of teaching and learning design in programming (Waite et al., 2020). In lower secondary, particular attention is devoted to the theme of abstraction in computing (Grover et al., 2019; Statter & Armoni, 2020).

Concerning curriculum integration, results highlight different implications related to the three main approaches adopted for integrating CT, namely as (i) a cross-curriculum theme, (ii) within other subjects (e.g., mathematics and tech), or (iii) as a separate subject (e.g., CS subject).

The central role played by teachers and the setting of curriculum priorities emerged as key factors from the analysis of the selected papers. The positioning of CT skills in the overall curriculum poses several demands at both policy-making and educational management/organisation levels: making space in the curriculum for including foundational CS concepts to develop CT skills; providing clear guidelines on the amount of time that teachers should devote to teaching basic CS content; allocating adequate resources for developing high-quality instructional material; and sharing examples of sound pedagogical practices. When CT skills are positioned as a cross-curricular theme, it is crucial to clarify the respective responsibilities of each subject teacher in this process.

Open questions for future investigation include (among others) how CT skills are taught and assessed when implementing CT in the curriculum, and how to pursue adequate gender balance and equity.

## 4.1. Limitations

As with any systematic review, the one presented here has its limitations. First, the search for identifying relevant academic publications was limited to 24 top-tier journals and conference proceedings. Furthermore, although two search strings with several combinations of key terms were applied, if the authors had not included these specific terms in the title, abstract and keywords of their paper/s, the respective article/s may have been excluded from this review.

Finally, certain limitations can be attributed to the screening and inclusion processes, which involved nine researchers, although several measures were taken to ensure inter-rater reliability, as described in Section 2.4 above.

# Data availability statement

The following files generated in the context of this study are available on Zenodo:

- The list of the 98 publications analysed in-depth through the review matrix: https://doi.org/10.5281/zenodo.7603313
- The structure of the complete review matrix: https://doi.org/10.5281/zenodo.7603402

# References

Arfé, B., Vardanega, T., & Ronconi, L. (2020). The Effects of coding on children's planning and inhibition skills. *Computers and Education*, 148, 1–16. https://doi.org/10.1016/j.compedu.2020.103807

Asbell-Clarke, J., Rowe, E., Almeda, V., Edwards, T., Bardar, E., Gasca, S., Baker, R. S., & Scruggs, R. (2021). The Development of students' computational thinking practices in elementary- and middle-school classes using the learning game, Zoombinis. *Computers in Human Behavior*, *115*, 1–14. https://doi.org/10.1016/j.chb.2020.106587

Atmatzidou, S., & Demetriadis, S. (2016). Advancing students' computational thinking skills through educational robotics: A Study on age and gender relevant differences. *Robotics and Autonomous Systems*, 75, 661–670. https://doi.org/10.1016/j.robot.2015.10.008

Australian Computing Academy. (2019). Coding and computational thinking—What is the evidence? Australian Computing Academy. https://education.nsw.gov.au/content/dam/main-education/teaching-and-learning/education-for-a-changing-world/media/documents/Coding-and-Computational-Report\_A.pdf

Balanskat, A., Engelhardt, K., & Licht, A. H. (2018). *Strategies to include computational thinking in school curricula in Norway and Sweden—European Schoolnet's 2018 Study Visit*. European Schoolnet. http://www.eun.org/documents/411753/817341/Computational\_thinking\_report\_2018.pdf/4d3d6fa0-dedd-4b62-a201a26bf4dfd3a0

Barendsen, E., Grgurina, N., & Tolboom, J. (2016). A New informatics curriculum for secondary education in The Netherlands. In A. Brodnik, & F. Tort (Eds.), *Lecture Notes in Computer Science: Vol. 9973* (pp. 105–117). Springer. https://doi.org/10.1007/978-3-319-46747-4\_9

Basu, S., Rutstein, D., Shear, L., & Xu, Y. (2020). A Principled approach to designing a computational thinking practices assessment for early grades. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 912-918). Association for Computing Machinery. https://doi.org/10.1145/3328778.3366849

Belur, J., Tompson, L., Thornton, A., & Simon, M. (2018). Interrater reliability in systematic review methodology: Exploring variation in coder decision-making. *Sociological Methods* & *Research*, 50(2), 837–865. https://doi.org/10.1177/0049124118799372

Bocconi, S., Chioccariello, A., Dettori, G., Ferrari, A., & Engelhardt, K. (2016). *Developing Computational Thinking in Compulsory Education - Implications for policy and practice*. Publications Office of the European Commission. https://doi.org/10.2791/792158

Bocconi, S., Chioccariello, A., Kampylis, P., Dagienė, V., Wastiau, P., Engelhardt, K., Earp, J., Horvath, M.A., Jasutė, E., Malagoli, C., Masiulionytė-Dagienė, V., & Stupurienė, G. (2022). *Reviewing Computational Thinking in Compulsory Education*. Publications Office of the European Union. https://doi.org/10.2760/126955

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American educational research association, Vancouver, Canada* (pp. 1–25). https://scratched.gse.harvard.edu/ct/files/AERA2012.pdf

Caeli, E. N., & Yadav, A. (2020). Unplugged approaches to computational thinking: A Historical perspective. *TechTrends*, 64(1), 29–36. https://doi.org/10.1007/s11528-019-00410-5

Cateté, V., Alvarez, L., Isvik, A., Milliken, A., Hill, M., & Barnes, T. (2020). Aligning theory and practice in teacher professional development for computer science. In *Proceedings of the 20th Koli Calling International Conference on Computing Education Research* (pp. 1–11). Association for Computing Machinery. https://doi.org/10.1145/3428029.3428560

Ching, Y.-H., Hsu, Y. C., & Baldwin, S. (2018). Developing Computational thinking with educational technologies for young learners. *TechTrends*, *62*(6), 563–573. https://doi.org/10.1007/s11528-018-0292-7

Coenraad, M., Palmer, J., Weintrop, D., Eatinger, D., Crenshaw, Z., Pham, H., & Franklin, D. (2021). The Effects of providing starter projects in open-ended scratch activities. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 38–44). Association for Computing Machinery. https://doi.org/10.1145/3408877.3432390

Corradini, I., Lodi, M., & Nardelli, E. (2017). Conceptions and misconceptions about computational thinking among Italian primary school teachers. In *Proceedings of the 2017 ACM Conference on International Computing Education Research* (pp. 136–144). Association for Computing Machinery. https://doi.org/10.1145/3105726.3106194

Corradini, I., Lodi, M., & Nardelli, E. (2018). An Investigation of Italian primary school teachers' view on coding and programming. In S. N. Pozdniakov, & V. Dagiene (Eds.), *Lecture Notes in Computer Science: Vol. 11169* (pp. 228–243). Springer. https://doi.org/10.1007/978-3-030-02750-6\_18

Curzon, P., Bell, T., Waite, J., & Dorling, M. (2019). Computational thinking. In S. Fincher & A. Robins (Eds.), *The Cambridge Handbook of Computing Education Research* (Cambridge Handbooks in Psychology, pp. 513-546). Cambridge University Press. https://doi.org/10.1017/9781108654555.018

Csizmadia, A., Standl, B., & Waite, J. (2019). Integrating the constructionist learning theory with computational thinking classroom activities. *Informatics in Education*, 18(1), 41–67. https://doi.org/10.15388/infedu.2019.03

Dagli, Z., & Sancar Tokmak, H. (2021). Exploring high school computer science course teachers' instructional design processes for improving students' "computational thinking" skills. *Journal of Research on Technology in Education*, 54(4), 511–534. https://doi.org/10.1080/15391523.2021.1881844

del Olmo-Muñoz, J., Cózar-Gutiérrez, R., & González-Calero, J. A. (2020). Computational thinking through unplugged activities in early years of primary education. *Computers and Education*, *150*, 1–19. https://doi.org/10.1016/j.compedu.2020.103832

Djambong, T., Freiman, V., Gauvin, S., Paquet, M., & Chiasson, M. (2018). Measurement of computational thinking in K-12 education: The Need for innovative practices. In D. Sampson, D. Ifenthaler, J. Spector, & P. Isaías (Eds.), *Digital technologies: Sustainable innovations for improving teaching and learning* (pp. 193–222). Springer. https://doi.org/10.1007/978-3-319-73417-0\_12

Eickelmann, B., Labusch, A., & Vennemann, M. (2019). Computational thinking and problem-solving in the context of IEA-ICILS 2018. In D. Passey, R. Bottino, C. Lewin, & E. Sanchez (Eds.), *IFIP TC 3 Open Conference on Computers in Education, OCCE 2018 Proceedings* (pp. 14–23). Springer. https://doi.org/10.1007/978-3-030-23513-0\_2

European Commission. (2020a). Digital education action plan 2021-2027: Resetting education and training for the digital age. https://eurlex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0624

European Commission. (2020b). Commission staff working document accompanying the document communication from the commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the regions digital education action plan 2021-2027 resetting education and training for the digital age. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020SC0209&qid=1647943853396

Fagerlund, J., Häkkinen, P., Vesisenaho, M., & Viiri, J. (2020). Assessing 4th grade students' computational thinking through scratch programming projects. *Informatics in Education*, *19*(4), 611–640. https://doi.org/10.15388/infedu.2020.27

Fessakis, G., & Prantsoudi, S. (2019). Computer science teachers' perceptions, beliefs and attitudes on computational thinking in Greece. *Informatics in Education*, *18*(2), 227–258. https://doi.org/10.15388/infedu.2019.11

Forlizzi, L., Lodi, M., Lonati, V., Mirolo, C., Monga, M., Montresor, A., Morpurgo, A., & Nardelli, E. (2018). A Core informatics curriculum for Italian compulsory education. In S. N. Pozdniakov, & V. Dagiene (Eds.), *Lecture Notes in Computer Science: Vol. 11169* (pp. 141–153). Springer. https://doi.org/10.1007/978-3-030-02750-6\_11

Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., Friedman, T. (2019). Computational thinking framework. In *IEA International Computer and Information Literacy Study 2018 Assessment Framework*. Springer. https://doi.org/10.1007/978-3-030-19389-8\_3

Franklin, D., Hill, C., Dwyer, H. A., Hansen, A. K., Iveland, A., & Harlow, D. B. (2016). Initialisation in Scratch: Seeking knowledge transfer. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (pp. 217–222). Association for Computing Machinery. https://doi.org/10.1145/2839509.2844569

González-González, C. S., Herrera-González, E., Moreno-Ruiz, L., Reyes-Alonso, N., Hernández-Morales, S., Guzmán-Franco, M. D., & Infante-Moro, A. (2019). Computational thinking and down syndrome: An Exploratory study using the KIBO robot. *Informatics*, 6(2), 1–20. https://doi.org/10.3390/informatics6020025

Grgurina, N., Barendsen, E., Suhre, C., Zwaneveld, B., & Van Veen, K. (2018). Assessment of modeling and simulation in secondary computing science education. In Q. Cutts & A. Muhling (Eds.), *Proceedings of the 13th Workshop in Primary and Secondary Computing Education* (pp. 1-10). Association for Computing Machinery. https://doi.org/10.1145/3265757.3265764

Grover, S., Jackiw, N., & Lundh, P. (2019). Concepts before coding: Non-programming interactives to advance learning of introductory programming concepts in middle school. *Computer Science Education*, 29(2-3), 106–135. https://doi.org/10.1080/08993408.2019.1568955

Grover, S., & Pea, R. (2018). Computational thinking: A Competency whose time has come. In S. Sentance, E. Barendsen, & S. Carsten (Eds.), *Computer Science Education: Perspectives on Teaching and Learning in School* (pp. 19–38). Bloomsbury. https://doi.org/10.5040/9781350057142.ch-003

Guggemos, J. (2021). On the predictors of computational thinking and its growth at the high-school level. *Computers and Education*, 161, 1–15. https://doi.org/10.1016/j.compedu.2020.104060

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An Overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8, 23–34. https://doi.org/10.20982%2Ftqmp.08.1.p023

Hazzan, O., Ragonis, N., & Lapidot, T. (2020). Computational thinking. In O. Hazzan, N. Ragonis, & T. Lapidot (Eds.), *Guide to Teaching Computer Science* (pp. 57–74). Springer. https://doi.org/10.1007/978-3-030-39360-1\_4

Hooshyar, D., Pedaste, M., Yang, Y., Malva, L., Hwang, G.-J., Wang, M., Lim, H., & Delev, D. (2020). From gaming to computational thinking: An Adaptive educational computer game-based learning approach. *Journal of Educational Computing Research*, *59*(3), 383–409. https://doi.org/10.1177/0735633120965919

Hromkovič, J., Kohn, T., Komm, D., & Serafini, G. (2016). Combining the power of python with the simplicity of logo for a sustainable computer science education. In A. Brodnik & F. Tort (Eds.), *Lecture Notes in Computer Science: Vol. 9973* (pp. 155–166). Springer. https://doi.org/10.1007/978-3-319-46747-4\_13

Hromkovič, J., & Lacher, R. (2017). The Computer science way of thinking in human history and consequences for the design of computer science curricula. In V. Dagiene & A. Hellas (Eds.), *Lecture Notes in Computer Science: Vol. 10696* (pp. 3–11). Springer. https://doi.org/10.1007/978-3-319-71483-7\_1

Hsu, T.-C., Chang, S.-C., & Hung, Y.-T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers and Education*, *126*, 296–310. https://doi.org/10.1016/j.compedu.2018.07.004

Hsu, Y.-C., Irie, N. R., & Ching, Y.-H. (2019). Computational Thinking Educational Policy Initiatives (CTEPI) across the globe. *TechTrends*, 63(3), 260–270. https://doi.org/10.1007/s11528-019-00384-4

Huang, W., & Looi, C.-K. (2020). A Critical review of literature on "unplugged" pedagogies in K-12 computer science and computational thinking education. *Computer Science Education*, 31(1), 83–111. https://doi.org/10.1080/08993408.2020.1789411

Israel, M., Chung, M. Y., Wherfel, Q. M., & Shehab, S. (2020). A Descriptive analysis of academic engagement and collaboration of students with autism during elementary computer science. *Computer Science Education*, 30(4), 444–468. https://doi.org/10.1080/08993408.2020.1779521

Israel-Fishelson, R., & Hershkovitz, A. (2020). Persistence in a game-based learning environment: the case of elementary school students learning computational thinking. *Journal of Educational Computing Research*, 58(5), 891–918. https://doi.org/10.1177/0735633119887187

Israel-Fishelson, R., Hershkovitz, A., Eguíluz, A., Garaizar, P., & Guenaga, M. (2021). A Log-based analysis of the associations between creativity and computational thinking. *Journal of Educational Computing Research*, 58(8), 1415–1447. https://doi.org/10.1177/0735633120940954

Jocius, R., Joshi, D., Dong, Y., Robinson, R., Catete, V., Barnes, T., Albert, J., Andrews, A., & Lytl, N. (2020). Code, connect, create: The 3C professional development model to support computational thinking infusion. In *Proceedings of the 51st ACM technical symposium on computer science education* (pp. 971–977). Association for Computing Machinery. https://doi.org/10.1145/3328778.3366797

Kale, U., Akcaoglu, M., Cullen, T., Goh, D., Devine, L., Calvert, N., & Grise, K. (2018). Computational what? Relating computational thinking to teaching. *TechTrends*, 62(6), 574–584. https://doi.org/10.1007/s11528-018-0290-9

Kert, S. B., Kalelioğlu, F., & Gülbahar, Y. (2019). A Holistic approach for computer science education in secondary schools. *Informatics in Education*, 18(1), 131–150. https://doi.org/10.15388/infedu.2019.06

Komm, D., Hauser, U., Matter, B., Staub, J., & Trachsler, N. (2020). Computational thinking in small packages. In K. Kori & M. Laanpere (Eds.), *Lecture Notes in Computer Science: Vol. 12518* (pp. 170–181). Springer. https://doi.org/10.1007/978-3-030-63212-0\_14

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 and Education*, *127*, 178–189. https://doi.org/10.1016/j.compedu.2018.08.026

Kong, S.-C., Lai, M., & Sun, D. (2020). Teacher development in computational thinking: Design and learning outcomes of programming concepts, practices and pedagogy. *Computers and Education*, *151*, 1–19. https://doi.org/10.1016/j.compedu.2020.103872

Labusch, A., & Eickelmann, B. (2020). Computational thinking competences in countries from three different continents in the mirror of students' characteristics and school learning. In S. C. Kong, H. U. Hoppe, T. C. Hsu, R. H. Huang, B. C. Kuo, K. Y. Li, C. K. Looi, M. Milrad, J. L. Shih, K. F. Sin, K. S. Song, M. Specht, F. Sullivan, & J. Vahrenhold (Eds.), *Proceedings of International Conference on Computational Thinking Education 2020* (pp. 2–7). The Education University of Hong Kong. http://www.eduhk.hk/cte2020/doc/CTE2020%20Proceedings.pdf

Leonard, A. E., Daily, S. B., Jörg, S., & Babu, S. V. (2021). Coding moves: Design and research of teaching computational thinking through dance choreography and virtual interactions. *Journal of Research on Technology in Education*, 53(2), 159–177. https://doi.org/10.1080/15391523.2020.1760754

Li, Q. (2020). Computational thinking and teacher education: An Expert interview study. *Human Behavior and Emerging Technologies*, *3*, 324–338. https://doi.org/10.1002/hbe2.224

Liebe, C., & Camp, T. (2019). An Examination of abstraction in K-12 computer science education. In *Proceedings of the 19th Koli Calling International Conference on Computing Education Research* (pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3364510.3364526 Metcalf, S. J., Reilly, J. M., Jeon, S., Wang, A., Pyers, A., Brennan, K., & Dede, C. (2021). Assessing computational thinking through the lenses of functionality and computational fluency. *Computer Science Education*, *31*(2), 199–223. https://doi.org/10.1080/08993408.2020.1866932

Niemelä, P., Partanen, T., Harsu, M., Leppänen, L., & Ihantola, P. (2017). Computational thinking as an emergent learning trajectory of mathematics. In *Proceedings of the 17th Koli Calling international conference on computing education research* (pp. 70-79). Association for Computing Machinery. https://doi.org/10.1145/3141880.3141885

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An Updated guideline for reporting systematic reviews. *The BMJ*, 372(71), 1–8. https://doi.org/10.1136/bmj.n71

Palts, T., & Pedaste, M. (2020). A Model for developing computational thinking skills. *Informatics in Education*, 19(1), 113–128. https://doi.org/10.15388/infedu.2020.06

Passey, D. (2017). Computer science (CS) in the compulsory education curriculum: Implications for future research. *Education and Information Technologies*, 22(2), 421–443. https://doi.org/10.1007/s10639-016-9475-z

Pasternak, A. (2016). Contextualized teaching in the lower secondary education long-term evaluation of a cs course from grade 6 to 10. In *Proceedings of the 47th ACM technical symposium on computing science education* (pp. 657–662). Association for Computing Machinery. https://doi.org/10.1145/2839509.2844592

Pérez-Marín, D., Hijón-Neira, R., Bacelo, A., & Pizarro, C. (2020). Can computational thinking be improved by using a methodology based on metaphors and scratch to teach computer programming to children? *Computers in Human Behavior*, *105*, 1–10. https://doi.org/10.1016/j.chb.2018.12.027

Razak, S., Khan, S., Hussein, N., Alshikhabobakr, H., Gedawy, H., & Yousaf, A. W. (2021). integrating computer science and ICT concepts in a cohesive curriculum for middle school-an experience report. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 966–972). Association for Computing Machinery. https://doi.org/10.1145/3408877.3432528

Repenning, A., Lamprou, A., & Basawapatna, A. (2021). Computing effect sizes of a science-first-then-didactics computational thinking module for preservice elementary school teachers. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 274–280). Association for Computing Machinery. https://doi.org/10.1145/3408877.3432446

Resnick, M. (2017). Lifelong kindergarten: Cultivating creativity through projects, passion, peers, and play. The MIT Press.

Rich, P. J., Mason, S. L., & O'Leary, J. (2021). Measuring the effect of continuous professional development on elementary teachers' self-efficacy to teach coding and computational thinking. *Computers and Education*, *168*, 1–15. https://doi.org/10.1016/j.compedu.2021.104196

Rijke, W. J., Bollen, L., Eysink, T. H. S., & Tolboom, J. L. J. (2018). Computational thinking in primary school: An Examination of abstraction and decomposition in different age groups. *Informatics in Education*, 17(1), 77–92. https://doi.org/10.15388/infedu.2018.05

Román-González, M., Pérez-González, J.-C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? Criterion validity of the computational thinking test. *Computers in Human Behavior*, 72, 678–691. https://doi.org/10.1016/j.chb.2016.08.047

Sherwood, H., Yan, W., Liu, R., Martin, W., Adair, A., Fancsali, C., Rivera-Cash, E., Pierce, M., & Israel, M. (2021). Diverse approaches to school-wide computational thinking integration at the elementary grades: A Cross-case analysis. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (pp. 253–259). Association for Computing Machinery. https://doi.org/10.1145/3408877.3432379

Shute, V. J., Sun, Ch., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142–158. https://doi.org/10.1016/j.edurev.2017.09.003

Statter, D., & Armoni, M. (2020). Teaching abstraction in computer science to 7th grade students. ACM Transactions on Computing Education, 20(1), 1–37. https://doi.org/10.1145/3372143

Sun, L., Hu, L., Yang, W., Zhou, D., & Wang, X. (2021a). STEM learning attitude predicts computational thinking skills among primary school students. *Journal of Computer Assisted Learning*, 37(2), 346–358. https://doi.org/10.1111/jcal.12493

Sun, L., Hu, L., & Zhou, D. (2021b). Which way of design programming activities is more effective to promote K- 12 students' computational thinking skills? A Meta-analysis. *Journal of Computer Assisted Learning*, *37*(4), 1048–1062. https://doi.org/10.1111/jcal.12545

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A Systematic review of empirical studies. *Computers and Education*, 148, 1–22. https://doi.org/10.1016/j.compedu.2019.103798

Taslibeyaz, E., Kursun, E., & Karaman, S. (2020). How to develop computational thinking: A Systematic review of empirical studies. *Informatics in Education*, *19*(4), 701–719. https://doi.org/10.15388/INFEDU.2020.30

Tedre, M., & Denning, P. J. (2016). The Long quest for computational thinking. In *Proceedings of the 16th Koli Calling international conference on computing education research* (pp. 120–129). Association for Computing Machinery. https://doi.org/10.1145/2999541.2999542

Tikva, C., & Tambouris, E. (2021). Mapping computational thinking through programming in K-12 education: A Conceptual model based on a systematic literature review. *Computers and Education*, *162*, 1–23. https://doi.org/10.1016/j.compedu.2020.104083

Tomokiyo, L. M. (2018). Successes and challenges in implementing a progressive K-8 computer science curriculum. In Q. Cutts, & A. Muhling (Eds.), *Proceedings of the 13th Workshop in Primary and Secondary Computing Education* (pp. 1–6). Association for Computing Machinery. https://doi.org/10.1145/3265757.3265779

Tonbuloğlu, B., & Tonbuloğlu, I. (2019). The Effect of unplugged coding activities on computational thinking skills of middle school students. *Informatics in Education*, 18(2), 403–426. https://doi.org/10.15388/infedu.2019.19

Tsai, M.-J., Liang, J.-C., & Hsu, C.-Y. (2020). The Computational thinking scale for computer literacy education. *Journal of Educational Computing Research*, 59(4), 579–602. https://doi.org/10.1177/0735633120972356

UNESCO, & IFIP TC3. (2019). Coding, programming and the changing curriculum for computing in schools. https://wcce2022.org/pubs/UNESCO%20meeting%20at%20OCCE%202018%20report%20final.pdf

Upadhyaya, B., McGill, M. M., & Decker, A. (2020). A Longitudinal analysis of K-12 computing education research in the United States: Implications and recommendations for change. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education* (pp. 605–611). Association for Computing Machinery. https://doi.org/10.1145/3328778.3366809

Waite, J., Curzon, P., Marsh, W., & Sentance, S. (2020). Difficulties with design: The Challenges of teaching design in K-5 programming. *Computers and Education*, *150*, 1–28. https://doi.org/10.1016/j.compedu.2020.103838

Webb, M., Davis, N., Bell, T., Katz, Y., Reynolds, N., Chambers, D. P., & Sysło, M. M. (2017). Computer science in K-12 school curricula of the 2lst century: Why, what, and when? *Education and Information Technologies*, 22(2), 445–468. https://doi.org/10.1007/s10639-016-9493-x

Wei, X., Lin, L., Meng, N., Tan, W., Kong, S.-C., & Kinshuk. (2021). The Effectiveness of partial pair programming on elementary school students' Computational Thinking skills and self-efficacy. *Computers and Education*, *160*, 1–15. https://doi.org/10.1016/j.compedu.2020.104023

Weintrop, D., & Wilensky, U. (2019). Transitioning from introductory block-based and text-based environments to professional programming languages in high school computer science classrooms. *Computers and Education*, *142*, 1–17. https://doi.org/10.1016/j.compedu.2019.103646

Wing, J. (2017). Computational thinking's influence on research and education for all. *Italian Journal of Educational Technology*, 25(2), 7–14. https://doi.org/10.17471/2499-4324/922

Witherspoon, E. B., Higashi, R. M., Schunn, C. D., Baehr, E. C., & Shoop, R. (2017). Developing computational thinking through a virtual robotics programming curriculum. *ACM Transactions on Computing Education*, 18(1), 1-20. https://doi.org/10.1145/3104982

Wu, S.-Y., & Su, Y.-S. (2021). Visual programming environments and computational thinking performance of fifth- and sixth-grade students. *Journal of Educational Computing Research*, 59(6), 1075–1092. https://doi.org/10.1177/0735633120988807

Yağcı, M. (2019). A Valid and reliable tool for examining computational thinking skills. *Education and Information Technologies*, 24(1), 929–951. https://doi.org/10.1007/s10639-018-9801-8

Zhang, L., & Nouri, J. (2019). A Systematic review of learning computational thinking through Scratch in K-9. *Computers and Education*, 141, 1–25. https://doi.org/10.1016/j.compedu.2019.103607

Zhang, L. C., Nouri, J., & Rolandsson, L. (2020). Progression of computational thinking skills in Swedish compulsory schools with block-based programming. In *Proceedings of the Twenty-Second Australasian Computing Education Conference* (pp. 66–75). Association for Computing Machinery. https://doi.org/10.1145/3373165.3373173

# Appendix 1. An excerpt from the review matrix focusing on the three research questions

The structure of the complete review matrix is available in Zenodo: https://doi.org/10.5281/zenodo.7603402. The complete references are also available in Zenodo: https://doi.org/10.5281/zenodo.7603313.

- RQ1 CT definition and terminology RQ2 Implementation (How things are done or proposed approach) RQ3- How can gender balance and equity be ensured when implementing CT in the curriculum?

No.	Short reference	RQ1	RQ2	RQ3
Acad	lemic literature			
1	Arfé, Vardanega & Ronconi (2020)			
2	Asbell-Clarke, Rowe, Almeda, Edwards, Bardar, Gasca, Baker & Scruggs (2021)			
3	Barendsen, Grgurina & Tolboom (2016)			
4	Basu, Rutstein, Shear & Xu (2020)			
5	Caeli & Yadav (2020)			
6	Cateté, Alvarez, Isvik, Milliken, Hill & Barnes (2020)			
7	Ching, Hsu & Baldwin (2018)			
8	Coenraad, Palmer, Weintrop, Eatinger, Crenshaw, Pham & Franklin (2021)			
9	Corradini, Lodi, & Nardelli (2017)			
10	Corradini, Lodi & Nardelli (2018)			
11	Csizmadia, Standl, & Waite (2019)			
12	Curzon, Bell, Waite & Dorling (2019)			
13	Dagli & Sancar Tokmak (2021)			
14	del Olmo-Muñoz, Cózar-Gutiérrez & González-Calero (2020)			
15	Djambong, Freiman, Gauvin, Paquet & Chiasson (2018)			
16	Eickelmann, Labusch & Vennemann (2019)			
17	Fagerlund, Häkkinen, Vesisenaho & Viiri (2020)			
18	Fessakis & Prantsoudi (2019)			
19	Forlizzi, Lodi, Lonati, Mirolo, Monga, Montresor, Morpurgo & Nardelli (2018)			
20	Franklin, Hill, Dwyer, Hansen, Iveland, & Harlow (2016)			
21	González-González, Herrera-González, Moreno-Ruiz, Reves-Alonso, Hernández-			
	Morales, Guzmán-Franco & Infante-Moro (2019)			
22	Grgurina, Barendsen, Suhre, Zwaneveld & Van Veen (2018)			
23	Grover & Pea (2018)			
24	Grover, Jackiw & Lundh (2019)			
25	Guggemos (2021)			
26	Hazzan, Ragonis & Lapidot (2020)			
27	Hooshyar, Pedaste, Yang, Malva, Hwang, Wang, Lim & Delev (2020)			
28	Hromkovič, Kohn, Komm & Serafini (2016)			
29	Hromkovič & Lacher (2017)			
30	Hsu, Chang & Hung (2018)			
31	Hsu, Irie & Ching (2019)			
32	Huang & Looi (2020)			
33	Israel, Chung, Wherfel & Shehab (2020)			
34	Israel-Fishelson & Hershkovitz (2020)			
35	Israel-Fishelson, Hershkovitz, Eguíluz, Garaizar & Guenaga (2021)			
36	Jocius, Joshi, Dong, Robinson, Catete, Barnes, Albert, Andrews & Lytl (2020)			
37	Kale, Akcaoglu, Cullen, Goh, Devine, Calvert & Grise (2018)			
38	Kert, Kalelioğlu & Gülbahar (2019)			
39	Komm, Hauser, Matter, Staub & Trachsler (2020)			
40	Kong, Chiu & Lai (2018)		•	
41	Kong, Lai & Sun (2020)			
42	Labusch & Eickelmann (2020)			
43	Leonard, Daily, Jörg & Babu (2021)			
44	Li (2020)			
45	Liebe & Camp (2019)			
46	Metcalf, Reilly, Jeon, Wang, Pyers, Brennan & Dede (2021)			
47	Niemelä, Partanen, Harsu, Leppänen & Ihantola (2017)			
48	Palts & Pedaste (2020)			
49	Pasternak (2016)			



# A Normative Analysis of the TechCheck Computational Thinking Assessment

# Emily Relkin<sup>1\*</sup>, Sara K. Johnson<sup>2</sup> and Marina U. Bers<sup>3</sup>

<sup>1</sup>Center for Children and Technology, Education Development Center, USA // <sup>2</sup>Eliot-Pearson Department of Child Study and Human Development, Tufts University, USA // <sup>3</sup>Lynch School of Education and Human Development, Boston College, USA // erelkin@edc.org // s.johnson@tufts.edu // marina.bers@bc.edu \*Corresponding author

**ABSTRACT:** *TechCheck* is an assessment of Computational Thinking (CT) for early elementary school children consisting of fifteen developmentally appropriate unplugged challenges that probe six CT domains. The first version of *TechCheck* showed good psychometric properties as well as ease of administration and scoring in a validation cohort of 768 children between 5 and 9 years of age. To increase sensitivity and reduce possible ceiling and floor effects, grade-specific versions of *TechCheck* (K, 1, 2) were subsequently created. In the present study, we explored how CT skills could be compared across grades when grade-specific versions of *TechCheck* are administered. First, we examined *TechCheck* raw score distributions and responses within CT domains in a representative sample of students from the three grades. Grade-specific Z-scores and percentile rankings were then calculated. To show utility of this normalization system, we used percentiles to compare CT outcomes between first and second graders who participated in a ScratchJr coding educational intervention. While *TechCheck* change scores suggested an unexpected 42.74% difference in CT outcomes between first and second grades. Normative analysis may provide a more meaningful way to compare results across grades when grade-specific versions of *TechCheck* are used. Implications for the future use of the *TechCheck* CT assessments are discussed.

Keywords: Assessment, Computer science, Early childhood, Coding

# **1. Introduction**

Computer Science (CS) is an integral part of early childhood education around the world (Fraillon et al., 2018; Hubwieser et al., 2015; White House, 2016). Children as young as preschool age are capable of learning to code with developmentally appropriate platforms (Clements & Gullo, 1984; Papadakis, 2021). One of the most important goals of teaching computer science (CS) to young children is to promote the development of computational thinking (CT) skills that allow for framing and solving problems using computers and other technologies. Acquiring CT skills is not limited to increasing CS knowledge but also can promote skills relevant to other disciplines, problem-solving, and self-expression in everyday life (Barr & Stephenson, 2011; Chen et al., 2017; Wing, 2010, Wing, 2006). There has been increasing interest in CT with many attempts to further define the concept, implement educational initiatives, and to create novel forms of assessment (Bakala et al., 2021; Lye & Koh, 2014; Román-González et al., 2019; Tang et al., 2020; Zhang & Nouri, 2019). Despite these efforts, most CT definitions do not take into account the context of early childhood. To address this gap, Bers (2018) developed a framework consisting of seven powerful ideas from Computer Science that are developmentally appropriate for children ages 4-9. These include the following domains: hardware/software, algorithms, modularity, control structures, representation, debugging, and design process.

There is a recognized need for well-designed and validated CT assessments for young children that can be easily administered in classroom and online settings (Grover & Pea, 2013; Lee et al., 2011; Poulakis & Politis, 2021; Román-González et al., 2019). An ideal CT assessment can be used to monitor students' progress in learning CT and allow educators to gauge the effectiveness of their lessons and CS curricula. CT assessment can be used to identify students in need of extra support as well as those with exceptional talents (Relkin et al., 2021; Román-González et al., 2019). In the context of research, it can provide new insights into how children's CT abilities develop and can assist in the development of new curricula and best practices for CS education (Zhang & Nouri, 2019). Various CT assessments for early childhood education have been created but are not always well-characterized. Tang et al. (2020) reported that of the 96 CT assessment studies analyzed (including all ages), only 45% reported reliability measures and only 18% reported validity evidence. The majority of CT assessments with validity evidence were designed for older students. Prior work with older children has helped researchers and educators identify the elements of CT amenable to assessment in early childhood (e.g., Werner et al., 2012), establish the utility of unplugged CT challenges (e.g., Román-González et al., 2018) and demonstrate

the applicability of item response theory for measuring the psychometric properties of CT assessments (e.g., Chen et al., 2017; Kong & Lai, 2022).

#### **1.1.** Assessments of CT for young children

Instruments for assessing CT in older students and adults have existed for some time (Chen et al., 2017; Fraillon et al., 2018; Werner et al., 2012). Many of these instruments are not developmentally appropriate for young children. A common assessment approach involves the use of coding exercises that are designed to elicit the same type of logic and reasoning that is involved in programming. However, coding-based assessments require prior knowledge of a coding language and can conflate coding ability with CT skills (Yadav et al., 2017). Assessments that require knowledge of coding cannot readily be used to assess baseline CT abilities in coding-naive students. In addition, research with older children has indicated that coding can become automatic and coding exercises may therefore not effectively probe CT (Werner et al., 2014).

There have been several attempts to create CT assessments for early childhood. Many CT instruments designed for early age groups utilize portfolio analysis, including interviews and/or observational methods (Bakala et al., 2021). For example, Mioduser and Levy (2010) used pre-programmed LEGO robotics construction tasks which they presented to kindergarten-age children. The children's CT level was qualitatively assessed by analyzing the terms that children used to describe the robot's actions as it navigated through a constructed environment. Children who attributed the robot's actions to magic or personification were given low CT skills ratings and those who provided mechanical explanations were considered more advanced. Wang et al. (2014) used a similar approach with 5-to-9-year-old children, who were asked open-ended questions about a tangible programming task that they created called "T-maze." "T-maze" uses TopCode to convert physical programs into digital code (Horn, 2012). The researchers identified elements of CT in the children's responses (e.g., abstraction, decomposition) as a basis for determining whether the children grasped these concepts. Bers et al., (2014) created a checklist to assess programs created by kindergarteners (ages 4.9 to 6.5 years old) exposed to a tangible and graphical programming language called CHERP (Creative Hybrid Environment for Robotics Programming). During one session, children were tasked with programming their robot to dance the "Hokey Pokey." The researchers then assessed four CT concepts by scoring children's projects on a Likert scale. Moore et al. (2020) used task and interview techniques to assess CT. Three participants were videotaped while they were interviewed and performed tasks using the Code and Go Robot Mouse Coding Activity (Learning Resources, Vernon Hills, IL). Researchers explored qualitatively how children use representations and translations to invent strategies for solving problems. Portelance and Bers (2015) conducted an exploratory study that assessed CT in young children by analyzing ScratchJr artifact-based video interviews of students in pairs. Researchers then analyzed videos of the dyads using holistic coding to identify categories.

Some effort has been put into creating activity-based CT assessments for young children. Marinus et al. (2018) created the Coding Development (CODE) Test 3–6 (for children between 3 and 6 years of age), which uses the robot Cubetto. CODE requires children to program the robot to go to a specified location on a mat by inserting wooden blocks into a "remote control." The task is to either build the program from scratch or debug an existing program. Children are given maximally three trials to complete each of the 13 items, with more points being awarded if fewer attempts are needed. Although the authors state that CODE is meant to measure CT, their assessment requires coding knowledge raising the possibility that their assessment conflates coding with CT skills. Clarke-Midura et al. (2021) are in the development stage of attempting to use evidence-centered design to develop a task-based assessment of CT for kindergarten-age children.

It is advantageous to be able to measure CT skills in children regardless of whether they have past knowledge or experience with computer programming (Grover et al., 2014). With this in mind, researchers began exploring the use of code-free instruments to assess CT skills in children. CT is exercised in the context of many "unplugged" activities (Bell & Vahrenhold, 2018; Zapata-Cáceres et al., 2020). Unplugged activities involve puzzles, games and exercises that exemplify CS concepts without requiring knowledge of coding or the use of computers. An unplugged activity typically involves a set of artifacts and procedures that are well-known to most school-age children. Unplugged activities have been used to teach CS concepts for over two decades (e.g., CSUnplugged.com; code.org), and in recent years have started to be used for the purposes of assessment. It has been argued that the unplugged assessments offer advantages because they do not rely on a particular computer language or curricula and are therefore purer reflections of CT abilities (Dagiene & Futschek, 2008).

Studies were published in 2018, 2020, 2021, and 2022 on five different unplugged CT assessments designed specifically for young children. The CTt for Beginners (BCTt) (Zapata-Cáceres et al., 2020), The Competent Computational Thinking Test (cCTt) (El-Hamamsy et al., 2022), *TechCheck* (Relkin et al., 2020), the

Computerized Adaptive Programming Concepts Test (CAPCT) (Hogenboom et al., 2021), and the Computational Thinking Assessment (CTA) (Tran, 2018). All four use unplugged challenges to probe CT domains and can be administered to children who lack prior coding experience. These instruments differ in the types of unplugged challenges they include, the CT domains assessed, the age ranges they cover, and the time required to complete and score the respective assessments (see Table 1). Some of the concepts probed by the BCTt, the CAPCT, and the CTA such as complex conditionals may be problematic for younger children on developmental grounds (Barrouillet & Lecas, 1999; Janveau-Brennan & Markovits, 1999; Muller et al., 2001). In addition, the CAPCT and the CTA require more advanced language and mathematical skills than typical K-2 students possess.

	<i>Tuble 1. H</i> <b>c</b> 0	inputison of tour unp	Jugged C1 medistres I	or young emilaren	
	The CTt for Beginners (BCTt)	The Competent Computational Thinking Test (cCTt)	TechCheck	Computerized Adaptive Programming Concepts Test (CAPCT)	Computational Thinking Assessment (CTA)
CT Concepts	Sequences, Loops (Simple, Nested), Conditionals (If-Then, If- Then-Else, While)	Sequences, Loops (Simple, Nested), Conditionals (If-Then, If-Then-Else, While)	Algorithms, Modularity, Debugging, Hardware/Software, Control Structures, Representation	Basic Sequences, Loops, Conditions (If & If-Else Statements), Debugging, Multiple Agents, Procedures, Generalization	Sequences, Algorithms, Loops, Debugging, Conditionals
Format Type	Pen and paper Multiple choice	Pen and paper Multiple choice	Pen and paper Online Multiple choice	Online Adaptive	Pen and paper Yes/No Prose responses
Items	25 items	25 items	15 items	4486 items (utilizes alternative forms of the same items)	10 items
Administrator Needed	Yes	Yes	Yes	No	No
Average Testing Time	40 minutes	30-35 minutes	13 minutes	Children play for as long as they want	6-10 minutes
Sample Age Range	299 students 5-12 (1- 6 <sup>th</sup> grade)	1519 students 7-9 (3-4 <sup>th</sup> grade)	1844 students 3-9 (PreK – 2 <sup>nd</sup> grade)	93,341 students 6-13 (1– 7 <sup>th</sup> grade)	183 students N/A (3rd grade) Not yet validated in younger children

*Table 1.* A comparison of four unplugged CT measures for young children

*Note.* The CTt for Beginners (BCTt) (Zapata-Cáceres et al., 2020), The Competent Computational Thinking test (cCTt) (El-Hamamsy et al., 2022); *TechCheck* (Relkin et al., 2020), the Computerized Adaptive Programming Concepts Test (CAPCT) (Hogenboom et al., 2021), the Computational Thinking Assessment (CTA) (Tran, 2018).

#### 1.2. Design, validation, and implementation of the original *TechCheck* assessment

*TechCheck*, an unplugged CT assessment for young children, was developed based on six of the seven powerful ideas of CS put forth by Bers (2018) (Relkin et al., 2020). The excluded powerful idea, Design Process, is an iterative and open-ended process that does not lend itself to a short multiple-choice assessment. *TechCheck* was initially tested in a cohort of 768 first and second graders (ages 5-9) participating in a research study involving the CAL-KIBO curriculum. *TechCheck* showed good reliability and validity according to classical test theory

(CTT) and item response theory (IRT), models that are commonly used to better understand the relationship of assessment items to the underlying concepts being measured (Kingsbury & Weiss, 1983). The mean difficulty index of all items was -1.25 (range = -2.63, .7), the mean discrimination index was 1.03 (range = 0.65, 1.41). The coefficient alpha indicated a moderate reliability ( $\alpha = 0.68$ ) (Hinton et al., 2004). The assessment scores were normally distributed, and the assessment readily distinguished among young children with different CT abilities. *TechCheck* scores correlated moderately (r = .53, p < .001) with a previously validated CT assessment tool called TACTIC-KIBO (Relkin & Bers, 2019).

#### 1.3. Grade-specific TechCheck versions

There are now three grade specific versions of *TechCheck TechCheck-K*, *TechCheck-1*, and *TechCheck-2* that are optimal for kindergarten, first and second/ third graders respectively (See Figure 1). When *TechCheck-1* was administered to kindergarten students, it became apparent that certain modifications were required. Previous research has shown that the working memory of children of kindergarten age (~5 years old) limits them to hold an average of three items in immediate memory, compared to children in first and second grade (~6-9 years old) who can hold an average of four items (Cowan, 2016; Simmering, 2012). This limit can potentially impact kindergartener's performance on multiple-choice assessments. Consequently, the number of response options was reduced from four to three in *TechCheck-K* (the kindergarten version). *TechCheck-K* was administered to N = 89 5-6-year-old students and the percentage of correct responses for each item on *TechCheck-K* closely paralleled that observed with *TechCheck-1*. We also noted a strong and significant correlation between the percentages correct on the two versions (r = 0.76, p < .001) (Relkin & Bers, 2021).



To create a version of *TechCheck* with improved psychometric properties for second graders, an item analysis of all the *TechCheck-1* questions was conducted. Questions that had low difficulty, discrimination, and/or point biserial correlations were modified (Relkin, 2021). *TechCheck-2* was administered to N = 63 second graders. The level of difficulty was increased in this version to mitigate a previously observed ceiling effect found when second grade students took *TechCheck-1*. *TechCheck-2* readily distinguished among young children with different CT abilities. Item equivalency to the original version of *TechCheck* was confirmed and the coefficient alpha was slightly higher (a = 0.74) than with the original assessment, (*TechCheck-1*). A paired sample *t*-test between baseline *TechCheck* and endpoint *TechCheck* was significant t = 4.01, df = 62, p < .0001.

#### 1.4. The present study

Among existing CT measures, there has been relatively little attention paid to methods for comparing CT skills across grades. The ability to perform cross-grade comparisons of CT skills is essential for assessing the applicability of CS curricula and coding platforms to specific age groups. The present study examines baseline performance on three versions of *TechCheck (TechCheck-K, TechCheck-1,* and *TechCheck-2* respectively) and applies item analysis to identify differences between these three grade-specific versions. Normalization using Z-
scores and percentile ranks was then introduced to allow comparison of performance on *TechCheck* across three grades (K,1,2). This study was conducted to answer the research question:

How can CT skills be compared across grades K, 1, and 2 when using three grade-specific versions of *TechCheck*?

## 2. Method

## 2.1. Participants

To examine the distribution of *TechCheck* scores across grades, we collected data from children in grades K-2 located in six different states across the USA. All assessments were administered prior to initiation of any formal coding instruction. Table 2 summarizes the demographic information for participants by their grade. Altogether, 1948 students were included in this analysis. A total of n = 395 kindergarteners, n = 935 first graders, and n = 618 second graders participated. The average student age was 6.64, SD = .84 with a minimum age of 4 and a maximum age of 9. There were also similar numbers of males (n = 725) and females (n = 728) in the three grades. Of the n = 1399 students from which we obtained race/ethnicity information, the most common race/ethnicity was White (58.89%) followed in frequency by Hispanic/ Latino (15.08%), Black (14.87%), Biracial/Multiracial (5.72%), Asian/Pacific Islander (4.15%%), and other (1.30%) respectively. The group characterized as "other" consisted of children identified as American Indian, Alaskan Native, Pacific Islander or Native Hawaiian.

To explore the utility of the normalized scoring technique for comparing CT performance across grades we analyzed data we collected in a longitudinal study carried out in the states of California, Minnesota, and Arkansas involving administration of a coding curriculum called CAL-ScratchJr to a total of n = 163 students in kindergarten, first, and second grade (Bers et al., in press).

	Kindergarten	First Grade	Second Grade
Number of students	395	935	618
Mean Age (SD)	5.86 (.42)	6.50 (.56)	7.81 (.35)
Missing data	55	185	299
Gender			
Male	164	399	162
Female	171	400	157
Missing data	60	136	299
Race			
Black/African American	34	154	20
Hispanic/ Latino	42	113	56
Biracial/Multiracial	23	42	15
White	220	399	206
Asian	12	30	16
Other	4	8	6
Missing data	60	190	299

Table 2. Demographics of pilot study participants by grade

Note. There is missing demographic data because some schools only shared limited information.

#### 2.2. Procedure

This study was initiated prior to the onset of the COVID-19 pandemic but was completed while the pandemic was in progress. As a consequence, different formats of administration of the *TechCheck* assessments were used over the course of this study. Some children were assessed in person while others participated virtually. Some assessments were carried out in group settings and others were conducted one-on-one. Some participants provided responses on paper while others used an online survey platform.

Regardless of the format, administrators were trained and certified to administer the assessment in a consistent fashion. Across all formats of administration, each question was read out loud to the students by an administrator who asked them to provide a single answer from a set of multiple-choice responses. There were two practice

questions that were included at the beginning of the assessment to ensure that children felt comfortable with the format of administration and knew how to indicate their answers. Students were allowed to take breaks for up to 5 minutes during the assessment. Students were instructed to guess if they did not know the answer. Administrators were instructed to indicate any abnormal issues that occurred during testing in an error log. That information was later used to clean the data.

#### 2.3. Data analysis

All statistical analyses and plots were conducted and created using R Studio version 1.2 (R Core Team, 2019) and Microsoft Excel version 16.23. Only students' baseline scores prior to receiving the CAL-ScratchJr coding curriculum were used in this analysis (Bers et al., in press). Descriptive statistics as well as data screening was conducted to examine assumptions for normality and linearity. A one-way ANOVA was conducted to explore differences between the three versions of the *TechCheck* assessments. Crossed random effects multi-level models were estimated to examine the relationship between domain specific scores and grade. Lastly, normalization was applied using Z-scores and percentile ranks.

# **3. Results**

#### **3.1.** Descriptive statistics

The average total scores were M = 7.48 (SD = 2.52) on *TechCheck-K*, M = 7.98 (SD = 2.46) on *TechCheck-1*, and M = 9.29 (SD = 2.75) on *TechCheck-2* out of a possible 15 points correct. Skewness and kurtosis values were within |2| ranging from -0.41 to 0.20 indicating the distributions were approximately normal for all three versions of the assessment (See Table 3). Density plots of each grade/assessment type also showed normal distributions with the second grade's cohort appearing to have a slightly more rightward skew than the other two grades (see Figure 2). Examination of Z- scores for kindergarten, first, and second grade revealed no extreme outliers of |3| or greater (Iglewicz & Hoaglin, 1993).

Table 3. Descriptive statistics for continuous variables

					~	
	Ν	Mean (SD)	Min	Max	Skewness	Kurtosis
TechCheck-K	395	7.48 (2.52)	0	14	0.19	-0.41
TechCheck-1	935	7.98(2.46)	0	15	0.02	-0.25
TechCheck-2	618	7.48(2.75)	0	15	0.20	-0.23

*Note.* Data in this table reflect scores on each version of *TechCheck* when administered to students prior to coding instruction.



*Note.* Density plots of total scores for *TechCheck-K* (orange, Kindergarten), *TechCheck-1* (blue, First grade), and *TechCheck-2* (green, Second grade).

#### 3.2. Differences between versions

To explore if the three grades' scores were significantly different from one another, a one-way ANOVA was conducted. This analysis showed baseline *TechCheck* scores were significantly different across grades F(2, 1841) = 7.71, p < .001. Tukey's HSD post-hoc test revealed a significant difference between first grade and kindergarteners (d = .42, p = .02) as well as between first grade and second grade (d = 0.49, p < .001).

To examine the possible basis for the observed differences between grades, item analysis was carried out by calculating percent correct responses within groups of questions corresponding to the six CT domains measured by *TechCheck*. When comparing the percentage correct at baseline for kindergarten, first, and second grade, the pattern of response is similar across the majority of domains (see Figure 3).

To establish whether the scores were statistically different within and across domains, we conducted a crossed random-effects multilevel model using REML estimation. In this model, CT domain and child were crossed and grade was a fixed effect predicting the CT domain score. First, we used an empty model with percent of questions correct within each CT domain as the outcome variable and a random effect of CT domain. The Intra-Class Correlation (ICC) was .37, which indicates that about 37% of the variation in CT domain percent correct was between domains (with the remaining percentage being differences between students across domains). A random effect of the intercept for the student variable was subsequently added to the model. The deviance significantly decreased, and the likelihood ratio test was significant, indicating the model with a random effect of grade (type of assessment administered). This addition significantly decreased deviance in the model ( $\Delta \chi^2(1) = 16.59$ , p < .0001), indicating that there is a difference across domains by grade. Upon examining the random effects, between-subjects variance attributable to the student and domain was .01 and .23 respectively. This indicates CT domain the total variance (see Table 12). ICC of the final model was .36 suggesting approximately 36% of the variation is between domains.



*Note*. Figure 3 shows the pattern of baseline responses between *TechCheck-K* (orange circle), *TechCheck-1* (blue triangle), *and TechCheck-2* (green square)

## 3.3. Normalization of scores

To permit comparison of *TechCheck* scores across the three grades, normalization techniques were carried out using Z-scores and percentile ranks. Results are shown in Figure 4. Consistent with the findings presented above, Z-scores and percentile scores appear similar in kindergarten and first grade but differ in second grade.

	Z-	Z-score			ntile Ranl	(
<i>TechCheck</i> Raw Score	Kindergarten	First	Second	Kindergarten	First	Second
1	-2.56	-2.82	-2.59	0.52	0.24	0.48
2	-2.17	-2.4	-2.2	1.5	0.82	1.39
3	-1.77	-2	-1.81	3.84	2.28	3.51
4	-1.38	-1.59	-1.42	8.38	5.59	7.78
5	-0.98	-1.18	-1.03	16.35	11.9	15.15
6	-0.59	-0.78	-0.64	27.76	21.77	26.11
7	-0.19	-0.36	-0.25	42.46	35.94	40.13
8	0.21	0.04	0.14	58.32	51.59	55.57
9	0.6	0.45	0.53	72.57	67.36	70.19
10	1	0.86	0.92	84.13	80.51	82.12
11	1.39	1.27	1.31	91.77	89.8	90.49
12	1.79	1.67	1.69	96.41	95.25	95.45
13	2.18	2.08	2.08	98.54	98.12	98.12
14	2.58	2.49	2.47	99.51	99.36	99.32
15	2.97	2.9	2.86	99.85	99.81	99.79

Figure 4. Z-scores and percentile rank norming for each of the three assessments/grades

*Note.* Z scores and Percentiles were calculated for each possible score (0-15 points) on *TechCheck-K*, *TechCheck-1*, and *TechCheck-2*. These scores can be used to compare scores from those who took the different versions of *TechCheck*.

## 3.4. Field test of normalized scoring system

In a ScratchJr pilot longitudinal study (Bers et al., in press), students were observed to improve significantly on *TechCheck* after exposure to a coding curriculum. First graders' average percentile at baseline was improved by 1.17 points on *TechCheck-1* (baseline score = 7.81, end point score = 8.98 raw points) while second graders' scores increased by 1.67 points on *TechCheck-2* (baseline score = 9.29, end post score = 10.95 raw points). Direct comparison of these mean change score results could be interpreted as showing a 42.74% greater change in second versus first graders. However, when expressed in terms of percentile changes using the normative scoring system first graders scored at the 43.40 percentile at baseline and at the 49.48 percentile after engagement in the CAL-ScratchJr curriculum (*delta* = 15.07). Second grade students scored at the 59.47 percentile at baseline and 73.02 percentile at the endpoint (*delta* = 20.24). Thus, first graders and second graders differed in outcome by only 5.17 percentile ranks when *TechCheck* score distributions were taken into account (See Figure 5).

Figure 5. Raw change score vs normalized change scores for a pilot CAL-ScratchJr pilot longitudinal study



# 4. Discussion

*TechCheck* was originally developed to fulfill the need for a well-characterized, developmentally appropriate CT assessment for early elementary school children. The original version (*TechCheck-1*) has shown considerable promise in children between the ages of 5-9. However, experience using the instrument revealed a possible ceiling effect in second graders and evidence of literacy/working memory limitations in kindergarteners. To address this, two modified versions were created (*TechCheck-K* and *TechCheck-2*) to supplement the original *TechCheck-1* which is most suitable for first grade students. In the current study, we compared baseline performance across grades using these three versions of *TechCheck*. We found that means and distributions differed across the three grade levels. Kindergarteners, first graders and second graders performed similarly by CT domain on the three versions of *TechCheck*.

Item analysis also showed small distinctions in responses across domains for the three grades-specific versions of TechCheck. The relative consistency of the pattern of responses across domains suggests these three versions of TechCheck are equivalent and developmentally appropriate for students across grades. The crossed random effects model provided evidence of differences in response patterns across domains but did not implicate particular domains as the basis for differences (Figure 3). To permit more meaningful cross-grade comparisons, we calculated Z-scores and percentile ranks for each grade from baseline data obtained from a large group of students. The creation of Z-score and percentile rank tables for the three versions of TechCheck offers certain advantages compared to raw scores in terms of understanding and communicating CT results. Raw scores can be difficult to interpret, particularly when score distributions differ across grades. When designing an assessment with the intention of comparing multiple grades/ages of students or following children longitudinally, many different techniques can be applied. One technique is to give children of different ages the same set of questions. This was our original plan for TechCheck. However, this resulted in ceiling effects in second graders and possible floor effects in kindergarten students. Another method is to create an adaptive assessment such as the approach taken by Hogenboom et al. (2021) with the CAPCT assessment. However, while this approach offers certain advantages, it does require a more complex system of administration, scoring, and interpretation. Typically, adaptive assessments employ larger numbers of questions and a broader scope of difficulty than is the case with *TechCheck*.

Standardizing scores is a common practice in large-scale educational assessment (Weiss, 2016). Although the present study does not utilize samples that are adequately representative of the populations in which they may be used, we chose to take an initial step towards normalization of the assessments for multiple reasons. The *TechCheck* data obtained to date has been normally distributed which facilitates the calculation of Z-scores. Percentile ranks based on Z-scores are familiar to many educators, parents, and administrators, making it easier for them to understand student performance and progress. Standardized scores can be used to compare students' performance to that of their peers. Percentile ranks can be conveyed to parents in a way that is easily understood. By providing a metric that can be used to evaluate students' progress from grade to grade, standardized scores can help schools evaluate the effectiveness of their programs. Standardized score benchmarks can be established to identify whose performance is significantly above or below expectations for grade. Growth norms can also be calculated, so that teachers can compare how much their students improved relative to other students (Set, 2018). Norms for different populations and cultures can be created to help researchers and practitioners compare performance cross-culturally. While the present results are a meaningful step towards normalization, data from larger, more representative populations of students will be required before the results can be considered fully standardized.

Percentile ranks take into account differences in the distributions of scores in ways that using raw score means alone cannot. This can be helpful when comparing performance across grades in longitudinal studies. For example, there was an unexpected 42.74% difference between the *TechCheck* change scores of first and second graders in the CAL- ScratchJr longitudinal study. However, when expressed in terms of percentile rank changes, this represented a relatively modest 5.17 percentile rank difference in outcome between grades. Percentile rank changes may therefore provide a more meaningful way to compare results from two or more grades when different versions of *TechCheck* are used for assessment.

# **5.** Limitations and future directions

*TechCheck* has been successfully administered in a variety of formats including in-person or remotely, online and on paper, to groups of students and individuals in many countries. The instrument has been translated into several languages in addition to English (i.e., Spanish, Turkish, Chinese, Dutch) for use in a variety of

educational and research settings around the world. Feedback from students, parents, teachers, and administrators have been remarkably positive. It is apparent that the assessment is easy to administer and score and that children enjoy taking it.

Although the multiple-choice format makes the assessment easy to administer and score, it also does not lend itself to creative self-expression and open-ended problem solving which is a significant part of CT. Thus, one of Bers' seven powerful ideas, Design Process, could not be probed in *TechCheck*. In addition, the possibility of guessing the correct answer is something that must be taken into account when interpreting multiple-choice results. Future studies should use item response theory statistical techniques such as 3pl models that take into account guessing. Román-González et al. (2019) pointed out that CT assessments often focus on "concepts "rather than "practices and perspectives," and as a consequence become "static and decontextualized" (p. 91). Other testing formats such as collection of CT telemetry data applied to real-time programming and/or individualized adaptive assessments may address these concerns.

The normalization carried out in this study is based upon data from cohorts of children in six US states constituting a relatively diverse sample. Nevertheless, our findings are subject to potential biases inherent in cohort studies including the sample not being fully representative of all children in the target age groups. While the cohorts were relatively balanced in terms of variables such as gender and race/ethnicity, other potential covariates such as socioeconomic status were not examined.

A new version of the assessment suitable for preschool age children ages 3-5 called *TechCheck-PreK* was recently validated. Future studies should explore whether the assessment can be used with neuro-diverse children and in other contexts. The goal of these efforts is to establish *TechCheck* as an assessment that can be used in a wide variety of research and real-world educational settings and assist in identifying the best CS educational practices to enhance the acquisition of CT in children.

# Acknowledgement

This work was supported by grants from the Department of Defense Education Activity (DoDEA) "Operation: Break the Code for College and Career Readiness". Unique Entity Identifier: "WORLDCL10" awarded to the Norfolk, Virginia Public Schools, the US Department of Education PR/Award Number: U411C190006 awarded to Dr. Marina Bers of Tufts University, and the Institute of Education Sciences, U.S. Department of Education, through Grant Award Number: R305A190433 awarded to the University of California, Los Angeles. We would like to thank the Norfolk Public School ITRTs, The Shaffer Evaluation Group, and the DevTech assessment team, and the participating children and teachers who made this work possible.

# References

Bakala, E., Gerosa, A., Hourcade, J. P., & Tejera, G. (2021). Preschool children, robots, and computational thinking: A Systematic review. *International Journal of Child-Computer Interaction*, 29, 100337. https://doi.org/10.1016/j.ijcci.2021.100337

Barr, V., & Stephenson, C. (2011). Bringing computational thinking to K-12: What is involved and what is the role of the computer science education community? *ACM Inroads*, 2(1), 48-54. https://doi.org/10.1145/1929887.1929905

Barrouillet, P., & Lecas, J. (1999). Mental models in conditional reasoning and working memory. *Thinking & Reasoning*, 5(4), 289–302.

Bell, T., & Vahrenhold, J. (2018). CS Unplugged—How Is It Used, and Does It Work? In H.-J. Böckenhauer, D. Komm, & W. Unger (Eds.), Adventures Between Lower Bounds and Higher Altitudes: Essays Dedicated to Juraj Hromkovič on the Occasion of His 60th Birthday (pp. 497–521). Springer International Publishing. https://doi.org/10.1007/978-3-319-98355-4\_29

Bers, M. U. (2018). *Coding as a playground: Programming and computational thinking in the early childhood classroom*. Routledge Press.

Bers, M. U., Flannery, L., Kazakoff, E. R., & Sullivan, A. (2014). Computational thinking and tinkering: exploration of an early childhood robotics curriculum. *Computers in Education*, *72*, 145–157. https://doi.org/10.1016/j.compedu.2013.10.020

Bers, M. U., Blake-West, J., Govind, M., Levinson, T., Relkin, E., Unahalekhaka, A., & Yang, Z. (in press). Coding as another language: Research-based curriculum for early childhood computer science.

Chen, G., Shen, J., Barth-Cohen, L., Jiang, S., Huang, X., & Eltoukhy, M. (2017). Assessing elementary students' computational thinking in everyday reasoning and robotics programming. *Computers in Education*, 109, 162–175. https://doi.org/10.1016/j.compedu.2017.03.001

Clarke-Midura, J., Silvis, D., Shumway, J. F., Lee, V. R., & Kozlowski, J. S. (2021). Developing a kindergarten computational thinking assessment using evidence-centered design: The Case of algorithmic thinking. *Computer Science Education*, *31*(2), 1–24. https://doi.org/10.1080/08993408.2021.1877988

Clements, D. H., & Gullo, D. F. (1984). Effects of computer programming on young children's cognition. *Journal of Educational Psychology*, 76, 1051–1058. https://doi.org/10.1037/0022-0663.76.6.1051

Cowan, N. (2016). Working memory maturation: Can we get at the essence of cognitive growth? *Perspectives on psychological science: A Journal of the Association for Psychological Science*, 11(2), 239–264. https://doi.org/10.1177/1745691615621279

Dagienė, V., & Futschek, G. (2008). Bebras international contest on informatics and computer literacy: Criteria for good tasks. In R. T. Mittermeir & M. M. Sysło (Eds.), *Informatics Education—Supporting Computational Thinking* (pp. 19–30). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-69924-8\_2

El-Hamamsy, L., Zapata-Cáceres, M., Barroso, E. M., Mondada, F., Zufferey, J. D., & Bruno, B. (2022). The Competent computational thinking test: Development and validation of an unplugged computational thinking test for upper primary school. *Journal of Educational Computing Research*, 07356331221081753. https://doi.org/10.1177/07356331221081753

 International Computer and Information Literacy

 Study:
 ICILS
 2018:
 Technical
 Report.
 https://www.iea.nl/sites/default/files/2020 

 05/ICILS%202018%20Technical%20Report-FINAL\_0.pdf
 05/ICILS%202018%20Technical%20Report-FINAL\_0.pdf
 05/ICILS%202018%20Technical%20Report-FINAL\_0.pdf

Grover, S., & Pea, R. (2013). Computational thinking in K-12: A Review of the state of the field. *Educational researcher*, 42(1), 38-43. https://doi.org/10.3102/0013189X12463051

Grover, S., Cooper, R., & Pea, R. (2014). Assessing computational learning in K-12. In *Proceedings of the 2014 conference* on Innovation & technology in computer science education (pp. 57-62). ACM. https://doi.org/10.1145/2591708.2591713

Hinton, P., Brownlow, C., McMurray, I., & Cozens, B. (2004). SPSS explained. Taylor & Francis. https://doi.org/10.4324/9780203642597

Hogenboom, S. A. M., Hermans, F. F. J., & van der Maas, H. L. J. (2021). Computerized adaptive assessment of understanding of programming concepts in primary school children. *Computer Science Education*. https://doi.org/10.1080/08993408.2021.1914461

Horn, M. (2012). TopCode: Tangible object placement codes. http://users.eecs.northwestern.edu/~mhorn/topcodes

Hubwieser, P., Giannakos, M. N., Berges, M., Brinda, T., Diethelm, I., Magenheim, J., Pal, Y., Jacova J., & Jasute, E. (2015). A Global snapshot of computer science education in K-12 schools. In *Proceedings of the 2015 ITiCSE on working group reports* (pp. 65-83). https://doi.org/10.1145/2858796.2858799

Iglewicz, B., & Hoaglin, D. C. (1993). How to detect and handle outliers. American Society for Quality Control.

Janveau-Brennan, G., & Markovits, H. (1999). The Development of reasoning with causal conditionals. *Developmental Psychology*, 35(4), 904–911. https://doi.org/10.1037/0012-1649.35.4.904

Kingsbury, G. G., & Weiss, D. J. (1983). A Comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In *New horizons in testing* (pp. 257-283). Academic Press. https://doi.org/10.1016/B978-0-12-742780-5.50024-X

Kong, S.-C., & Lai, M. (2022). Validating a computational thinking concepts test for primary education using item response theory: An Analysis of students' responses. *Computers & Education*, 187, 104562. https://doi.org/10.1016/j.compedu.2022.104562

Lee, I., Martin, F., Denner, J., Coulter, B., Allan, W., Erickson, J., Malyn-Smith, J., & Werner, L. (2011). Computational thinking for youth in practice. *ACM Inroads*, 2(1), 32–37. https://doi.org/10.1145/1929887.1929902

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. https://doi.org/10.1016/j.chb.2014.09.012

Marinus, E., Powell, Z., Thornton, R., McArthur, G., & Crain, S. (2018). Unravelling the cognition of coding in 3-to-6-year olds: The Development of an assessment tool and the relation between coding ability and cognitive compiling of syntax in natural language. In *Proceedings of the 2018 ACM Conference on International Computing Education Research - ICER '18* (pp. 133–141). https://doi.org/10.1145/3230977.3230984

Mioduser, D., & Levy, S. T. (2010). Making sense by building sense: Kindergarten children's construction and understanding of adaptive robot behaviors. *International Journal of Computers for Mathematical Learning*, *15*(2), 99–127. https://doi.org/10.1007/s10758-010-9163-9

Moore, T. J., Brophy, S. P., Tank, K. M., Lopez, R. D., Johnston, A. C., Hynes, M. M., & Gajdzik, E. (2020). Multiple representations in computational thinking tasks: A Clinical study of second-grade students. *Journal of Science Education and Technology*, 29(1), 19–34. https://doi.org/10.1007/s10956-020-09812-0

Muller, U., Overton, W. F., & Reene, K. (2001). Development of conditional reasoning: A Longitudinal study. *Journal of Cognition and Development*, 2(1), 27–49. https://doi.org/10.1207/S15327647JCD0201\_2

Papadakis, S. (2021). The Impact of coding apps to support young children in computational thinking and computational fluency. A Literature review. *Frontiers in Education*, 6. https://www.frontiersin.org/articles/10.3389/feduc.2021.657895

Portelance, D. J., & Bers, M. U. (2015). Code and tell: Assessing young children's learning of computational thinking using peer video interviews with scratchjr. In *Proceedings of the 14th International Conference on Interaction Design and Children* - *IDC '15* (pp. 271–274). https://doi.org/10.1145/2771839.2771894

Poulakis, E., & Politis, P. (2021). Computational thinking assessment: Literature review. In T. Tsiatsos, S. Demetriadis, A. Mikropoulos & V. Dagdilelis (Eds.), *Research on E-Learning and ICT in Education* (pp. 111–128). Springer, Cham. https://doi.org/10.1007/978-3-030-64363-8\_7

R Core Team. (2019). R: A Language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Relkin, E., & Bers, M. (2021). TechCheck-K: A Measure of computational thinking for kindergarten children. In 2021 IEEE Global Engineering Education Conference (EDUCON). IEEE. https://sites.tufts.edu/devtech/files/2021/05/1487.pdf

Relkin, E. (2021). TechCheck: Creation of an unplugged computational thinking assessment for young children. In M. U. Bers (Ed.), *Teaching Computational Thinking and Coding to Young Children* (pp. 250-264). IGI Global. https://doi.org/10.4018/978-1-7998-7308-2

Relkin, E., & Bers, M. U. (2019). Designing an assessment of computational thinking abilities for young children. In L.E. Cohen & S. Waite-Stupiansky (Eds.), *STEM for Early Childhood Learners: How Science, Technology, Engineering and Mathematics Strengthen Learning* (pp. 85-98). Routledge.

Relkin, E., de Ruiter, L., & Bers, M. U. (2020). TechCheck: Development and validation of an unplugged assessment of computational thinking in early childhood education. *Journal of Science Education and Technology*, *29*, 482–498. https://doi.org/10.1007/s10956-020-09831-x

Relkin, E., de Ruiter, L., & Bers, M. U. (2021). Learning to code and the acquisition of computational thinking by young children. *Computers & Education*, 169, 104222. https://doi.org/10.1016/j.compedu.2021.104222

Román-González, M., Moreno-León, J., & Robles, G. (2019). Combining assessment tools for a comprehensive evaluation of computational thinking interventions. In S. C. Kong, & H. Abelson (Eds.), *Computational Thinking Education* (pp. 79-98). Springer. https://doi.org/10.1007/978-981-13-6528-7\_6

Román-González, M., Pérez-González, J. C., Moreno-León, J., & Robles, G. (2018). Can computational talent be detected? Predictive validity of the Computational Thinking Test. *International Journal of Child-Computer Interaction*, *18*, 47-58. https://doi.org/https://doi.org/10.1016/j.ijcci.2018.06.004

Set, A. (2018). Making assessment more meaningful with norms. NWEA. https://www.nwea.org/blog/2018/making-assessment-more-meaningful-with-norms/

Simmering, V. R. (2012). The Development of visual working memory capacity during early childhood. *Journal of Experimental Child Psychology*, 111(4), 695–707. https://doi.org/10.1016/j.jecp.2011.10.007

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: a systematic review of empirical studies. *Computers in Education*, 148, 103798. https://doi.org/10.1016/j.compedu.2019.103798

Tran, Y. (2018). Computational thinking equity in elementary classrooms: What third-grade students know and can do. *Journal of Educational Computing Research*, 57(1), 3-31. https://doi.org/10.1177/0735633117743918

Wang, D., Wang, T., & Liu, Z. (2014). A Tangible programming tool for children to cultivate computational thinking. *The Scientific World Journal*, 2014, 428080. https://doi.org/10.1155/2014/428080

Weiss, L. G. (2016). *Standardized assessment for clinical practitioners: A Primer*. https://www.pearsonassessments.com/content/dam/school/global/clinical/us/assets/featured-topics/assessment-primer-whitepaper.pdf

Werner, L., Denner, J., & Campe, S. (2014). Using computer game programming to teach computational thinking skills. In K. Schrier (Ed.), *Learning, education and games: Volume 1, curricular and design considerations* (pp. 37–53). ETC Press. https://dl.acm.org/doi/10.5555/2811147.2811150

Werner, L., Denner, J., Campe, S., & Kawamoto, D. C. (2012). The Fairy performance assessment: Measuring computational thinking in middle school. In *Proceedings of the 43rd ACM Technical Symposium on Computer Science Education* (pp. 215–220). https://doi.org/10.1145/2157136.2157200

White House. (2016). Educate to innovate. https:// www.whitehouse.gov/issues/education/k-12/educate-innovate

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3),33–35. https://doi.org/10.1145/1118178.1118215

Wing, J. (2010). *Computational thinking: A Definition* (Unpublished manuscript). https://www.cs.cmu.edu/~CompThink/resources/TheLinkWing.pdf

Yadav, A., Good, J., Voogt, J., & Fisser, P. (2017). Computational thinking as an emerging competence domain. In *Technical and vocational education and training* (Vol. 23, pp. 1051–1067). https://doi.org/10.1007/978-3-319-41713-4\_49

Zapata-Cáceres, M., Martín-Barroso, E., & Román-González, M. (2020). Computational thinking test for beginners: Design and content validation. In 2020 IEEE Global Engineering Education Conference (EDUCON) (pp. 1905-1914). IEEE. https://doi.org/10.1109/EDUCON45650.2020.9125368

Zhang, L., & Nouri, J. (2019). A Systematic review of learning computational thinking through Scratch in K-9. *Computers in Education*, 141, 103607. https://doi.org/10.1016/j.compedu.2019.103607

Cui, Z., Ng, O., & Jong, M. S.-Y. (2023). Integration of Computational Thinking with Mathematical Problem-based Learning: Insights on Affordances for Learning. *Educational Technology & Society*, 26(2), 131-146. https://doi.org/10.30191/ETS.202304\_26(2).0010

# Integration of Computational Thinking with Mathematical Problem-based Learning: Insights on Affordances for Learning

# Zhihao Cui<sup>1</sup>, Oi-lam Ng<sup>1\*</sup> and Morris Siu-Yung Jong<sup>1,2</sup>

<sup>1</sup>Department of Curriculum and Instruction, The Chinese University of Hong Kong, Hong Kong SAR, China // <sup>2</sup>Centre for Learning Sciences and Technologies, The Chinese University of Hong Kong, Hong Kong SAR,

China // cuizhihao@link.cuhk.edu.hk // oilamn@cuhk.edu.hk // mjong@cuhk.edu.hk

\*Corresponding author

**ABSTRACT:** Grounded in problem-based learning and with respect to four mathematics domains (arithmetic, random events and counting, number theory, and geometry), we designed a series of programming-based learning tasks for middle school students to co-develop computational thinking (CT) and corresponding mathematical thinking. Various CT concepts and practices articulating the designated mathematical problems were involved in the tasks. In addition to delineating the design of these learning tasks, this paper presents a qualitative study in which we examined 74 students' learning outcomes and characterized their CT and mathematical thinking co-development as they accomplished the tasks. The research results demonstrate the co-development of both mathematics- and CT-related concepts and practices in the four mathematics domains. Two types of interactions are identified: (i) applying mathematical knowledge to construct CT artifacts and (ii) generating new mathematical knowledge with CT practice. The new insights provided by the present work are threefold. First, from a mathematical learning perspective, the nature of the solution processes of the designed problems should not be immediately obvious. Second, from a technology-enhanced learning perspective, the dynamic representations and immediate visual feedback afforded by the programming tool are beneficial to student learning. Third, from a pedagogical perspective, the room for customization offered by both the designed problems and programming tools can provide affordances for learning.

Keywords: Computational thinking, Mathematics education, Problem-based learning, Problem solving, STEM

# **1. Introduction**

Computational thinking (CT) can be regarded as a mode of problem solving and thinking with computational tools and as a fundamental skill required in daily life (Wing, 2006; Wing, 2011). In the current development of teaching and learning with computing, much emphasis has been placed on integration with other disciplines and fields (Guzdial & Soloway, 2003); this represents a shift away from focusing on computer science education in isolation. In light of science, technology, engineering, and mathematics (STEM) education has been a global educational focus today (Jong, Song, Soloway & Norris, 2021), CT is regarded as a kind of analytical thinking that shares close connections with all four involved disciplines (Leung, 2020), and especially with mathematics (Baldwin et al., 2013). The use of programming and the application of CT to learning mathematics can be traced back to Papert (1980), who argued that CT could have a unique effect on mathematical thinking and learning because it provides learners with a medium for exploring patterns and a logical structure for modeling and investigating mathematical relationships. More recently, a systematic illustration of the connection between mathematics and CT was proposed by Weintrop and colleagues (2016), who suggested that various CT practices, including data practices, modeling and simulation practices, computational problem-solving practices, and systems-thinking practices, can play a supportive role in mathematical practices and be mutually promoted. Recent reviews have revealed considerable literature growth around the integration of CT and mathematics in recent decades (e.g., Hickmott et al., 2018; Ye et al., 2023), arguing for the multi-faceted linkage of CT and K-12 mathematics education. Although these reviews evince the reciprocal relationship between CT and mathematical concepts, the question of how CT and mathematics can be co-developed remains underexplored (Nordby et al., 2022; Ye et al., 2023), as stated by Hickmott et al. (2018) "studies that explicitly linked the learning of mathematics concepts with computational thinking were uncommon in the reviewed literature" (p. 65). Recently, there have been studies exploring CT integration for learning in specific mathematical domains, such as combinatorics (De Chenne & Lockwood, 2022), number theory and mathematical modeling (Benton et al., 2018; Ng & Cui, 2021), and geometry (Ng & Cui, 2021; Pei et al., 2018), as well as others investigating challenges that emerge when engaging in mathematical problem solving within a programming environment (Cui & Ng, 2021; DeJarnette, 2019; Ng et al., 2021; Weng et al., 2022). However, as argued by Lockwood and De Chenne (2019), while programming seems to be effective in learning mathematics for certain topics, it cannot be concluded that it would be superior to paper-and-pencil methods in all mathematics domains. Therefore, further research is needed to understand the interplay between these two modes of thinking (i.e., mathematical and CT), especially in terms of shedding light on the affordances of simultaneously using two modes of thinking, as well as on the challenges students may experience when solving mathematical problems in programming contexts.

To this end, there is still a great deal of room to investigate the integration of CT with mathematics education, especially in K–12 contexts. The two most significant remaining questions in this regard are (i) how CT and mathematics learning outcomes can be co-developed (Ye et al., 2023) and (ii) connecting specific mathematical domains for integration with programming (Lockwood & Morken, 2021). In response to these research gaps, this study addresses the characteristics of CT-based mathematics instruction and student learning in such an environment. Our goals in this study are twofold. First, we illustrate the design elements of the CT-based mathematical tasks from four mathematical domains (i.e., arithmetic, random events and counting, number theory, and geometry) and highlight their impact on students' learning outcomes. Second, we are interested in identifying the affordances and barriers brought forward by problem-based mathematics learning in the block-based programming environment, Scratch. Specifically, we aim to address the following research questions (RQs):

- How is CT co-developed with problem-based mathematics learning in designed tasks in each of the following mathematical domains: arithmetic, random events and counting, number theory, and geometry?
- How might the design of CT-based mathematical activities provide affordances for student learning in each of these domains?

# 2. Conceptual framework

#### 2.1. Computational thinking, concepts, and practices

In the past decade, researchers have made efforts to develop conceptual and methodological frameworks for learning and teaching CT (e.g., Brennan & Resnick, 2012; Ho et al., 2021; Jong et al., 2020; Román-González et al., 2017; So et al., 2020). Among them, Brennan and Resnick (2012) proposed one of the most popular frameworks in which CT can be addressed from three dimensions: computational concepts, computational practices, and computational perspectives. They identified seven computational practices (incremental and iterative, testing and debugging, reusing and remixing, and abstracting and modularizing), and three kinds of computational perspectives on the world and oneself as a programmer (expressing, connecting, and questioning). In this study, we considered the three dimensions of CT proposed by Brennan and Resnick as the learning goals of CT.

Conversely, another group of researchers has explored the relationship between CT and thinking practices in other disciplines, such as mathematics. For example, Sneider et al. (2014) created a Venn diagram illustrating the overlap between mathematical thinking (MT) and CT, wherein the common area included problem solving, modeling, analyzing, and interpreting data, as well as skills in statistics and probability. They explained that outside the intersection of MT and CT, more distinct MT (e.g., counting and geometry) and CT (e.g., programming and data mining) practices are found. Weintrop et al. (2016) formulated a taxonomy integrating mathematics and CT into four categories: data practices, modeling and simulation practices, computational problem-solving practices, and systems-thinking practices. This mapping enabled them to produce framework statements that reflect how CT is applied—particularly in the context of mathematics and science—as a way to support integrated instruction that mutually enriches student learning in each discipline. The taxonomy was informed by the research finding that computational problem-solving practices, such as algorithm development and creating computational abstractions, can help learners develop a deep understanding of mathematical and scientific phenomena (e.g., Wilkerson-Jerde, 2014). We believe that this CT practice taxonomy could serve as a strategy for learning and problem solving, especially within the scope of integrating CT into problem-based mathematical learning.

#### 2.2. Affordance for learning

"Affordance" was introduced by Gibson (1979) to describe the relationships that exist between an object or environment and an organism. The subsidiary idea is that affordances provide both opportunities and constraints, which are not opposites but complementary. Norman (1999) proposed one of the most notable reformulations of the concept of affordance with respect to "real" affordance and "perceived" affordance, according to which real affordance refers to the physical characteristics of a device or interface that allow its operation, as described by Gibson (1979), while perceived affordance can be defined as the apparent characteristics of a device that provide clues to its proper operation. Differing from Gibson (1979), Norman (1999) recognized that the object or environment could be both symbolically and physically designed and that the term "affordance" could be used for the purpose of design. Building on the work of Gibson (1979) and Norman (1999), Hartson (2003) further defined the term "affordance" as cognitive affordance (Norman's perceived affordance) and physical affordance (Norman's real affordance); in this scheme, affordance is a "design feature" that "aids, supports, facilitates, or enables thinking, knowing, and/or doing something" (p. 319). In the context of educational research, educational affordances are "those characteristics of an artifact (e.g., how a chosen educational paradigm is implemented) that determine if and how a particular learning behaviour could possibly be enacted within a given context (e.g., project team, distributed learning community)" (Kirschner, 2002, p. 14). In the mathematics education context, Bishop and colleagues (2014) provide another example of cognitive affordance, referring to individuals' understanding or knowledge that may lead to successful learning progress or problem solving within the targeted content.

Informed by Kirschner's educational affordance and Bishop's cognitive affordance in mathematics learning, we propose a framework for understanding the potential affordance of integrating CT with mathematics in the current study (Figure 1). First, we emphasize design features, which include physical and symbolic (or intangible) aspects (Norman, 1999). For example, in the context of the present study, the block-based programming environment could be treated as a physical design affordance for learning CT, because it was something with which the students could directly interact. Moreover, we identified the learning content (i.e., the four mathematical domains of arithmetic, random events and counting, number theory, and geometry) as an intangible design affordance for supporting learning behavior (Kirschner, 2002). For instance, certain geometrical content is suitable to accompany visual representations to support programming practices and thus to construct, explore, and verify the properties of geometric figures; this represents a case of the selection of learning content to afford students' CT-based mathematics learning. Second, we highlight the support feature of affordance in terms of providing both opportunities and constraints. This can be exemplified by the use of computing to make arithmetic calculations with ease (i.e., opportunity); however, the students must correctly program to perform the calculations (i.e., constraint). In summary, CT-based mathematics instruction may provide a unique affordance for student learning from the perspectives of design and support.

Design feature	Support feature	
<i>Physical</i> : e.g., programming environment	<i>Opportunity</i> : e.g., make arithmetic calculations of large numbers accessible	
<i>Symbolic (Intangible)</i> : e.g., selection of content	<i>Constraint</i> : e.g., programming skills required, and programming	

challenges experienced

*Figure 1.* The two features of affordance in the current study **Affordance** 

# 3. Methodology

#### 3.1. Research design, participants, and context

This study is situated in a series of programming-based teaching interventions that address various mathematical domains. It employed a design-based research (DBR) methodology consisting of three iterative cycles of implementations to achieve its aims. DBR is conducted "with the intent of producing new theories, artifacts and practices that account for and potentially impact learning and teaching in naturalistic settings" (Barab & Squire, 2004, p. 2). During the three cycles of implementation, the researchers designed and refined a total of eight programming-based mathematical tasks in partnership with mathematics schoolteachers and computer science experts.

A total of 74 participants (57 male and 17 female) ranging from fifth to eighth grade (ages 10 to 14) were recruited from various primary and secondary schools in Hong Kong and provided informed consent to participate in the study. According to their self-reported previous experience in programming before the study, some of the participants had experienced some very basic functions related to programming (such as motion control and simple conditions) in Scratch. They had never been engaged in using programming to solve mathematical problems or in learning more comprehensive CT concepts (e.g., variables and iteration) and practice (e.g., modeling and remixing). Hence, their prior knowledge was considered to have no significant influence on the learning outcomes of the current study.

#### 3.2. Selected mathematical domains and tasks

Table 1 lists the mathematical domains and tasks developed and implemented in the study, labeled (1)–(8). Specifically, the first cycle of implementation employed tasks (1) and (6); the second cycle of implementation addressed tasks (1), (2), (3), and (8); and the third cycle of implementation involved tasks (4), (5), (7), and (8). All the tasks were designed with authentic contexts with real-life relevance and were open-ended in nature, which required knowledge from the respective mathematical domains to solve. In summary, the tasks can be categorized into four major mathematical domains (arithmetic, random events and counting, number theory, and geometry), thus allowing the research questions to be explored. Noting that most of the selected mathematical content (i.e., geometric sequences, probability, and fractal geometry) had not been introduced in formal lessons prior to the study, it can be inferred that both mathematical and CT concepts were developed by the students in the current study. More information about the tasks implemented in this study is provided in Appendix 1.

Domoin	Toolt nome	Mathematical concent involved	Evenented anodust
Domain	Task name	Mathematical concept involved	Expected product
Arithmetic	Two Savings problem	Sequence and series	Numerical output and their
	(1)		visual representations
	Fibonacci Sequence (2)	Recursive sequence	
Random events	Dice Rolling problem	Random events with equally	Value output of variables/visual
and counting	(3)	likely outcomes	representation of distributions
	Dart Throwing problem	Random events with unequally	
	(4)	likely outcomes	
Number theory	Count to 21 or 100	Counting, inductive and	A math game with inputs and
	problem (5)	deductive reasoning	computer auto-reactions
	Prime Detector (6)	Divisibility rules, factors and	Text and/or list output
		multiples	_
Geometry	Drawing Polygons (7)	Exterior and interior angles	Multiple polygons
-	Drawing Fractals (8)	Fractal and recursion	Fractal geometry

Table 1. Selected domains and tasks with corresponding mathematical and CT concepts

#### **3.3. Procedures**

The three cycles of implementation employed a similar set of procedures. Participants attended three to five programming sessions involving various mathematical problem-based learning in increasing order of complexity. Each session took approximately two hours. In the first part of each session, the course instructor conducted whole-class instruction with the goal of scaffolding essential prerequisite mathematical and programming knowledge for solving the target problems in the respective sessions. Afterward, the students would follow demonstrations led by the course instructor, answer questions posed by the instructor, and complete some guided activities. The remainder of each session (around one hour) was devoted to students' individual and collaborative problem solving, in which teaching assistants, with a teacher–student ratio of roughly 1:6, provided the necessary assistance. Participants were encouraged to communicate with peers about their thoughts and plans to solve the problem, while this process was video-recorded. After each session, participants wrote reflections on the tasks; this included critically discussing what they had learned and the challenges they had met during the session. By the end of the implementation, selected students were invited to participate in semi-structured post-course interviews. The semi-structured questions included: What was something new you learned in the course? Were there any challenges or difficulties that you met, and how did you overcome them? How did you come up with ideas for solving the problem?

## 3.4. Data analysis

We adopted a case study as the analytic methodology. Case study complements in-depth analyses of learning "given the need for extended, open, and careful consideration of data" (Parnafes & diSessa, 2013, p. 7). It takes into account the intriguing parts and significant components of the subject, which is suitable for answering research questions such as those proposed in the current study. During the DBR research, we collected data, including programming artifacts, classroom observation notes, video recordings, field notes, and student interviews. The researchers first reviewed all the artifacts constructed by the students, as well as the video and audio recordings of the class. Then, by combining these with the classroom observation field notes, the researchers selected episodes and artifacts that characterized the students' cognitive development of both CT and mathematics in each mathematical domain. For the selected episodes, we employed a constant comparative strategy (Corbin & Strauss, 2015) to narrow down the selection of episodes and artifacts so that they demonstrated and characterized the students' learning outcomes from both mathematical and CT perspectives. The student interviews served as supplemental evidence for triangulating the results. With regard to the nature of affordance according to the proposed framework, we identified how the present instructional design could provide affordances for the co-development of CT and mathematics.

To examine students' CT development in this study, we adopted two influential frameworks—those of Brennan and Resnick (2012) and Weintrop et al. (2016)—to analyze the students' CT development as encompassing a set of CT concepts and practices (Table 2). The shortened list of CT concepts and practices served as the coding criteria for demonstrating students' development of CT when reviewing the data. For example, the CT concepts in use could be identified by the programming codes used by the students, such as "if ... then" and "repeat" with respect to the CT concepts of conditionals and loops, respectively. For CT practices, we referred to the students' programming processes over a period of time in terms of what kinds of subtasks they were tackling within the CT environment, e.g., modeling, testing, and debugging (Weintrop et al., 2016) or reusing and remixing (Brennan & Resnick, 2012).

1 44	<i>The 2.</i> CT concepts and practices involved in the current study
CT concepts	Description
Loop	A mechanism for running the same sequence multiple times
Sequence	A particular activity or task expressed as a series of individual steps or instructions that can be executed by the computer
Condition	Make decisions based on certain conditions, which supports the expression of multiple outcomes
Iteration	The outcome of each iteration is the starting point of the next iteration
Variable	Value that contains some known or unknown quantity
Subroutine	A complete executable packaged program instruction that can be used in other programs at any time
Boolean logic	A form of algebra in which all values are either True or False. These values are used to test the conditions.
CT practices	Description
Modelling and simulation	Using computational models to understand a concept, to find and test solutions; assessing, designing, and constructing computational models
Algorithmic thinking	A series of steps to solve a problem
Reusing and remixing	Building on others' work (i.e., ideas and code) to create things that are much more complex
Testing and debugging	Developing strategies (e.g., by trial and error) to deal with and anticipate problems

Table 2. CT concepts and practices involved in the current study

To investigate students' mathematical development, we designed tasks targeting certain mathematical concepts in a particular domain, as illustrated in Table 1. The list served as the coding criteria to select artifacts and extract episodes to provide evidence for students' mathematical development from the data. For example, the concept of random events of equal likelihood can be identified by how they inductively infer the law of large numbers from simulating dice rolls in the reflection questions. In addition, we consistently employed transcripts to analyze the students' discourse while engaging with the mathematical problems in the programming environment to triangulate data from different sources and ensure the credibility of the qualitative results.

# 4. Results

The following subsections present the results with respect to the designed tasks in four mathematical domains— (i) arithmetic, (ii) random events and counting, (iii) number theory, and (iv) geometry—with representative artifacts and episodes to detail, provide evidence for, and situate the students' development in both mathematics and CT. In each domain, to respond to RQ1, we first demonstrate how each type of mathematical content was co-developed with CT concepts and practices during the designed CT-based mathematics activities (intangible affordance). Then, in response to RQ2, we explain the importance of the design feature of the tasks that provided affordances for students' knowledge or skill acquisition.

# 4.1. Arithmetic

## 4.1.1. Co-development of CT and mathematics

Arithmetic thinking was co-developed with the CT concepts of variables and iterative operations. As the first problems tackled by the students, the CT-based arithmetic tasks were meant for the students to begin translating their arithmetic procedures, such as computing 3 + 222 + 222 = 447 and 3 + 6 + 12 = 21 in the Two Savings problem, and finding the next term in a Fibonacci sequence by adding the previous two terms using the Scratch programming language. Given the programming environment's ability to take care of the arithmetic procedures effortlessly, the respective problems stimulated the students to use effective strategies to ensure that their programs displayed the correct sequence. For this, the use of variables was called for, where (i) a variable was something that took on different values, and (ii) variables could be operated iteratively by using codes such as "set balance to balance + 222." In other words, the students linked their mathematical thinking, which involved searching for patterns and determining the next term in the sequence, with variables in a CT sense, knowing that as long as something changed, they could use a variable to represent this changing quantity. Moreover, the use of variables was complemented with iterative operations by using the [repeat] code in Scratch, which enabled a quantity to change by the repeated use of a certain rule. As shown in Figure 2a, a typical solution to the Fibonacci Sequence problem was to operate three variables by setting "the third number" equal to the sum of the "first number" and "second number." In mathematical language, this was precisely  $T_n = T_{n-1} + T_{n-2}$ . The concept of iterations came into play in coding the next few lines because the "second number" and "third number" become the new "first number" and "second number," respectively, generating an iteration process, as illustrated in Figure 2b. Hence, we regard the students' arithmetic thinking as co-developing with the CT concepts of variables and iterative operation.



Figure 2. Programming codes of the Fibonacci Sequence and the iteration idea

4.1.2. Affordance from the design feature of the tasks

The same two CT-based mathematical activities afforded the students the opportunity to visualize and simulate the arithmetic (e.g., 3, 225, 447, ...) and geometric progressions (e.g., 3, 6, 12, ...) posed in the problems. Compared to a paper-and-pencil environment, which would likely have prompted the students to use a static formula, such as 3 + 222(n - 1) to describe the  $n^{\text{th}}$  term of the sequence, the students in this study used the programming environment to simulate each term dynamically, using the codes mentioned above to visualize the growth of various sequences one term at a time. Furthermore, some students used visual representations to show the amount of growth from one term to the next, which can be significant in improving their understanding of the differences between arithmetic and geometric sequences. On the other hand, the programming techniques required were considered constraints in solving the problems. As one student commented, "The numbers are too big, so it's nearly impossible for a human brain to do it, but I don't know how a computer thinks in this program. It took me three days to complete that task." This suggests that, although the student recognized the affordances

of computing in dealing with large numbers, he struggled with solving the problem in a computational context due to being unfamiliar with the programming tools.

Regarding the Two Savings problem, it was observed that the design of the problem provided affordances that supported the students in modeling a real-life scenario that involved arithmetic operations. After successfully creating a program that simulated the two saving plans, a final challenge remained: namely, modeling the situation in which a deposit was to be made every day for the first saving plan, as opposed to making deposits only on weekends for the second saving plan. With this type of problem design, we observed that the students used various nonroutine strategies to model the situation successfully. For example, some students used different keys to denote different parameters, such as using the "D" key to denote the passing of weekdays and "W" for the passing of weekends. Other strategies included (1) using seven days or a week as a unit, that is, within each week, repeating the deposit seven times for the first plan and two times for the second plan (Figure 3a), and (2) creating a new variable (i.e., day counter) to serve as a hint regarding the day of the week, and then operating the deposit accordingly (Figure 3b). This indicated that students experienced and developed the skills of modeling and simulation in computational practice.





#### 4.2. Random events and counting

#### 4.2.1. Co-development of CT and mathematics

The concept of randomness, as appropriated by the randomize function in Scratch, was codeveloped in the students' probabilistic thinking in mathematics. In the first stage of the Dice Rolling problem, the students were guided to simulate the situation of rolling six dice at once by using the randomize function and calculating their sums. As shown in the online chat record, the students initially held certain common-sense expectations regarding the concept of randomness: "I found that Scratch's random is fake ... it has a pattern." This comment was in agreement with other students' observations: "I got 22 five times (in 20 clicks)," "21 never happened for me," and "I got three 28s in a row (20 clicks)." These expressions suggest that the students had wrongly related the computer-generated randomized results to their expectations of a uniform distribution within only a small number of trials. In other words, they thought that when the dice were rolled randomly, a certain number should not appear at such a high frequency (i.e., five instances of 22 in 20 rolls), or a certain number should have appeared (i.e., 21 never happened but "1 + 2 + 3 + 4 + 5 + 6 = 21"). These conversations suggested that the students were rethinking the meaning of "randomness," both in a mathematical and computational sense: as they obtained more and more trials with the help of loops in Scratch, the students began to see that the observed frequency would mirror the expected distribution when performing random events.

At the same time, the students' concept of experimental probability was found to co-develop with the CT practice of simulation. In both the Dice Rolling problem and the Dart Throwing problem, the students were encouraged to simulate the process a large number of times to observe the distribution of the results. As illustrated in Figure 4a, in order to design a fair scoring system, one student ran the dart-throwing simulation 20,000 times to find the frequency distribution with which the dart hit the squares. As such, he proposed a scoring system that incorporated his observed frequency distribution. The square that was hit most frequently should be scored the lowest, and so on. Furthermore, using the data obtained from 20,000 simulations, the student designed a scoring system (A = 22, B = 12, C = 6, D = 2.5, E = 1) according to the ratio of the number of times each square was hit (i.e., the scoring system should be inversely proportional to the ratios). Another student

simulated the dart-throwing situation 1,000 times, and by observing the experimental outcomes of the darts' landing points with the area of the dartboard, he inferred that the two quantities were proportional. A similar observation was also found in the Dice Rolling problem. By visualizing the outcomes of the targeted sums in Scratch (Figure 4b), one student discovered that the outcomes were expected to be symmetrical around the median when obtaining the four-dice sum. These examples suggest that concepts of experimental probability co-emerged with the CT practice of simulation.



#### 4.2.2. Affordance from the design feature of the tasks

The most significant affordance provided by the tasks in this domain is the opportunity for students to simulate a large number of random events, which would be nearly impossible to do when performed manually. The combination of experimental probability with the programming environment was an instrumental affordance allowing students to develop their probabilistic thinking alongside computational concepts and practices. For example, the students' discourse about their expected dice rolling results reflected that they had made sense of what the frequency distribution would look like. With the ability to process a large number of trials by programming, the students tested their hypothesis, which filled the gap between experimental and theoretical probability—which is one of the main challenges for learning probability. The second stage of the Dice Rolling problem required students to generate the outcome space (e.g., 1-1-1-1-1, 1-1-1-2, ... 6-6-6-6-6), as well as the frequency of the sums. We note that students who successfully programmed to find the outcome space inferred that the obtained sums would be symmetrical around the median sum (i.e., 21 in a six-dice situation) and that the median sum would appear with the greatest frequency. With the comparison of the counting results and the experimental simulation over a large number of trials, the students found the relationship between theoretical and experimental probability, which was co-developed with the CT concept and practice of loops and simulations.

#### 4.3. Number theory

#### 4.3.1. Co-development of CT and mathematics

The mathematical concepts of divisibility, factors, and multiples were co-developed with the CT concept of conditions and Boolean logic. In the Prime Detector problem, the students used the code "mod" to determine whether a number was divisible by another number (i.e., A mod B results in the remainder of A divided by B). Starting from the definition of a prime as a whole number greater than 1 and divisible only by the number 1 and itself, a typical model in computational language was "check the mod of the target number (N) repeatedly from 2 to N – 1; if all the mod results are non-zero, then the target number is a prime," as illustrated in Figure 5a. The codes used by some students seemed to meet the logic of "checking all the divisors starting from 2 and then using the condition 'if' to determine whether all the remainders were greater than zero." However, this was incorrect because the computer did not store the results within the loop. The condition "if" only checked the result of the last divisor. We observed that the students who experienced this unsuccessful attempt turned to the alternative model of "checking for a composite number" (Figure 5b). In other words, whenever the remainder yielded a result of zero, this indicated that the number had a factor other than 1 or itself and was therefore a composite number. Here, the students needed to combine Boolean logic, which returned a value of true or false and the conditional statement of "if … then …" to conclude that a number was a prime "if all remainders were non-zero."

We note that the question of when to stop the loop came into play for some students, who questioned how they might optimize the condition in the program to make it "more efficient." The students knew that based on the definition of a prime number, one could perform divisions from N/2 to N/(N - 1) to check for factors within the interval [2, N - 1]. However, some students found that if they changed the stopping condition to "repeat until the divisor > target number/2" (Figure 5b), the results would remain the same. We suggest that this particular meaning of primes was situated in the students' CT practice in that the students were made aware of how to think like a computer as well as of ways to make the program more efficient (i.e., by taking half of the calculation time).





## 4.3.2. Affordance from the design feature of the tasks

The task provided an affordance supporting students in systematically testing and debugging their programs. For example, the student who programmed the codes in Figure 5a claimed that his program worked because he had tested the numbers 14, 16, and 110, and the program had returned the result of a composite number. Meanwhile, when he tested the numbers 5, 7, and 11, the program returned the result of a prime number. Consequently, he incorrectly claimed that the number 7,081 was a prime number because he had failed to test the program with an odd composite number, such as 15, which the program would incorrectly detect as a prime. This phenomenon raised the question of how to test the program offectively. Unlike the Two Savings problem, in which the students could compare their program output with their hand calculations to test whether the program had worked as desired, the prime detector problem was more sophisticated in that it prompted the students to decide what numbers to use to test their program. Without adequate consideration of the properties of primes and composites, their choice of testing numbers potentially limited their judgment of the correctness of the program. Therefore, the task was considered meaningful for developing the CT practice of testing and debugging in tandem with number properties.

## 4.4. Geometry

#### 4.4.1. Co-development of CT and mathematics

We observed the co-development of the students' geometrical thinking along with the CT concept of sequence, as well as the use of parameters. The mechanism of drawing in the Scratch program requires making a path for the Pen tool to move along. In this case, the programming sequence becomes critical because it differs from spontaneous drawing with paper and pencil. For example, drawing a line and then coming back to the starting point could be easily done by hand, but it would require the sequence of "move X steps – turn 180 degrees – move X steps" in Scratch, and the direction of the Pen would remain reversed. In the problem of Drawing Polygons, students were first guided to draw a triangle. Initially, some students encountered difficulties due to confusing the Pen tool's movement with the turning angle in Scratch. Using trial and error, they then successfully determined how to turn properly and went on to draw various (regular) polygons. In doing so, they also inferred that the sum of the exterior angle of any polygon would be 360 degrees (as reflected in the code "turn [360/number of sides] degrees" in Figure 6a) because the Pen tool would have rotated exactly one round

after drawing the polygon. The use of the parameter "number of sides" was especially pertinent in helping the students observe this relationship, which demonstrated that the students' geometrical thinking was closely supported by the CT environment.

The mathematical concept of the limit was co-developed with the CT concept of subroutines and the CT practices of reusing and remixing as the students continued to advance their usage of parameters. Unlike other tasks in the course in which the output was in numerical form rather than strictly visual, the students needed to think about the size and aesthetics of their drawings. As such, the students learned to use subroutines to duplicate drawings with varied sizes or customizable features using parameters. In drawing multiple polygons, such as the ones in Figure 6a, the students learned to create a function with two parameters (i.e., the number of sides and the lengths of sides). The "draw polygons" function now served as a subroutine that students could reuse repeatedly, which was a significantly different experience from drawing with paper and pencil. As one student commented regarding using subroutines: "The main code will be shorter and more efficient, and if you wanted to change some parts of the codes, you would only have to change it one time." Meanwhile, in the process of drawing polygons with varying numbers of sides (Figure 6b), one student asked, "If I draw a 360-sided polygon, will I get a circle?" This comment was derived from his observation that a polygon with many sides resembles the shape of a circle, so he set 360 as the parameter for the number of sides, which is a relatively large number of sides, making each turn angle 1 degree (very small). Regardless of what he believed to be the limit value, his exploratory thinking could be considered a limiting process that supported the construction of the limit concept. Importantly, this was uniquely contextualized in the CT environment, particularly with subroutines and the practices of reusing and remixing.





## 4.4.2. Affordance from the design feature of the tasks

The design features of "multiple" and "regular differences" were the key affordances supporting students' development of both geometrical concepts and the CT practice of reusing and remixing. For example, in this task, the students were encouraged to draw multiple polygons of different sizes and shapes, which would be difficult to perform with paper and pencil. To complete it efficiently, the task prompted students to observe and think about the similarities and differences among these polygons—that is, how the number of sides related to the exterior angles. In addition, during the process of exploring and drawing multiple polygons, the students came to appreciate the use of "functions" to demonstrate reusing and remixing. Figure 7 illustrates one student's work in drawing complex figures. After creating and using the "draw polygon" function as a basic element repeatedly, the student created a new function "T2" based on it, and then used "T2" to create another function "T3" (Figure 7a). The reusing and remixing of existing functions finally yielded a complex drawing (Figure 7b).





## 5. Discussion and conclusion

Based on the four mathematical domains, we presented the findings for the teaching interventions in which the learning tasks were designed and articulated with mathematical problem-based learning in programming environments. The results suggest that these domains, as exemplified by the eight tasks used in the study, provided affordances for the co-development of computational and mathematical aspects of learning. In the following, we explicate some factors that are critical for future research and practice in K-12 mathematics and programming education.

Apart from the intangible affordance provided by the selected content, the use of technology (i.e., Scratch) should be noted, especially in the context of integration with mathematical problems. First, we highlight the functions of "operators" built into the programming tools. The operators contain blocks that provide support for mathematical, logical, and string expressions, enabling the programmer to perform numerical and string manipulations. Some participants had experience using Scratch for programming; however, most only used functions focusing on interactive or narrative projects, such as animations and games (e.g., Kafai & Peppler, 2011). With the codes in the operators, students were able to perform the necessary calculations in order to solve mathematical problems they had not encountered in formal classes. For instance, with the code "mod," a concept that most students knew about but which was new to them in the programming context, students were able to explore advanced mathematical domains, such as number theory. Thus, this study demonstrated that applying mathematical knowledge to construct CT artifacts plays an important role in solving CT-based mathematical tasks (Bouck & Yadav, 2020; Grizioti & Kynigos, 2021; Miller, 2019).

The second technological aspect we wish to spotlight is the "sensing" function. The capabilities of the sensors included detecting the position of the sprite and mouse and any key input, as well as providing the affordance of human–computer interaction (HCI) via the "ask and answer" code. As suggested by Kafai and Burke (2014), the programming environment should move "beyond the computer screen to meld the digital with the tangible" (p. 91), thus providing additional sensory input and output. Although Scratch's sensors and HCI functions are virtual, they allow learners to see the actual movement and outcomes of the program visually, similar to the physical world (Ching et al., 2018). Based on our observations, the students showed enthusiasm for making something that was "more like a real product" by using the sensing function, while many students attempted to change the sprite and background to customize the problem's context. The students' pursuit of making products is in line with how "learning as making" (Ng & Chan, 2019) pedagogy supports mathematics learning as hands-on and goal-oriented. Making allows learners to actively construct knowledge instead of passively receiving information.

Third, we would like to highlight the feature of the stage area in Scratch. Unlike other text-based programming tools with value or text outputs, many codes were specifically intended to produce visual output in the Scratch stage area. In the current study context, the stage allowed the students to visualize abstract mathematical problems as "authentic." For instance, in the Two Savings problem, students could program the sprite to report their deposit and even change the images of the sprites to match the context of the problem; in the Dice Rolling problem, the students created sprites to determine the frequency of a target sum and various dynamic representations to visualize the frequency. Importantly, visualization became both the product and process of the

students' creation, and this afforded opportunities for the students to engage in mathematical thinking because visualization is one of the vital processes in mathematics learning (Barmby et al., 2007). It is argued that the stage area in Scratch is able not only to prompt the co-development of mathematics and CT knowledge but also to improve the students' ability to visualize in general.

Regarding the selection of mathematical domains and tasks, we identified two types of interaction in the codevelopment of mathematics and CT: (i) applying mathematical knowledge to construct CT artifacts and (ii) generating new mathematical knowledge along with CT practice (see also Ye et al., 2023). For the first type, students developed CT concepts, including variables, loops, iterations, and conditions, supported by existing mathematical knowledge. Previous research has suggested that the concept of looping in programming is difficult for students to master (Grover et al., 2015; Robins et al., 2003; Zur-Bargury et al., 2013). Also, students experience challenges in creating multiple variables and applying conditions (Cui & Ng, 2021). Given the mathematical context of the current study, the students were prompted to draft an algorithmic solution or create pseudocodes using their mathematical knowledge, which was considered helpful in overcoming difficulties in programming, thus supporting the development of CT concepts (Futschek, 2006; Grover et al., 2014). In addition, the mathematical ideas or relationships related to the problems were constructed based on the students' reflections on the CT outputs. As suggested by Wilkerson-Jerde (2014), students "explore[d] important mathematical properties of [fractal] structures, and offered more ways to construct fractals with particular mathematical properties" (p. 118) by observing a collection of fractals produced by a computer. Pei and colleagues (2018) also found that students could reason and generalize regarding patterns from the data of a number of polygons. The current study provides empirical evidence that the students developed new mathematical ideas, such as experimental probability and limits, through CT practice and outputs. They also constructed mathematical ideas and relationships by working with and reflecting on the CT outputs they created with Scratch.

The current study has both theoretical and practical implications for the integration of CT and mathematics. Theoretically, we clarified and highlighted the meaning of "affordance" in the instructional design. Moreover, we argue that the intangible affordance is even more important in technology-rich learning environments, given that the settings of these environments can vary for different programming languages and hardware. According to the results, the four mathematical domains provided unique opportunities for different CT concepts and practices (i.e., arithmetic – variables; random events – loops; number theory – testing and debugging; geometry – reusing and remixing). Meanwhile, within the given domains, the design features of the tasks played an important role. For example, the design of drawing multiple and different polygons in the domain of geometry prompted higher-order thinking in students regarding the mathematical concept of the limit. On the other hand, the constraints of programming-rich environments for mathematics learning should be acknowledged. In line with the findings of Cui and Ng (2021), the students encountered two types of constraints of affordance, namely mathematics-related constraints (e.g., the challenge of using appropriate code to perform certain functions). We suggest that programming-related constraints could influence students' learning outcomes in significant ways, and they should be minimized when teaching and learning mathematics in programming-rich environments.

In terms of practical contributions, we suggest three areas in which computational problem solving may enrich mathematics learning, as informed by the findings. The first is that the problem should be stated such that the solution and/or the solution process are not immediately known. From this perspective, some of the mathematicsrelated problems motivated students more than others when presented in the programming context. For example, when first presented, the solution process (e.g., strategies for counting to 21), the solution itself (e.g., a prime detector for large numbers), or both (e.g., experimental probability and fractal geometry) were not immediately obvious to the students. This element of the unknown provided opportunities for students to explore and inquire about new concepts, both in mathematics and programming. Second, computing with screen-based artifacts afforded dynamic visual representation and immediate feedback (such as the movement of the sprite, outputting a certain number, and a figure to be drawn), which significantly engaged the students who participated in this study. As such, the students were more likely to continue, regardless of the complexity and difficulty; thus, the visualization also prompted higher-order thinking toward the CT and mathematics concepts involved (Barmby et al., 2007). It is also worth noting that feedback enables students to manage their learning and mental processes through metacognition (Hesse et al., 2015). Conversely, a task requiring a long procedure with no outputs or feedback should be avoided because students might easily become stuck at some point, as was the case in the "fractal geometry" problem. Third, we consider room for autonomy to be an important characteristic to engage students by providing problems with choices and the customization of solutions. This might be explained by the fulfillment of autonomy, which is regarded as one of the key psychological requirements to support students' motivation to engage in learning tasks (Hsu et al., 2019).

To conclude, this paper described and discussed affordances provided by activities for teaching and learning mathematics in computationally enhanced ways, drawing on selected mathematical domains and tasks. This initial research is highly promising, but more research is warranted to further investigate the design of learning materials for computationally enhanced mathematical teaching and learning.

# Acknowledgment

The work described in this paper was fully supported by the Research Grants Council (Hong Kong), General Research Fund (Ref. No. 14603720).

## References

Baldwin, D., Walker, H. M., & Henderson, P. B. (2013). The Roles of mathematics in computer science. ACM Inroads, 4(4), 74–80. https://doi.org/10.1145/2537753.2537777

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences*, *13*(1), 1–14. https://doi.org/10.1207/s15327809jls1301\_1

Barmby, P., Harries, T., Higgins, S., & Suggate, J. (2007). How can we assess mathematical understanding? In *Proceedings* of the 31st Conference of the International Group for the Psychology of Mathematics Education (Vol. 2, pp. 41–48).

Benton, L., Saunders, P., Kalas, I., Hoyles, C., & Noss, R. (2018). Designing for learning mathematics through programming: A case study of pupils engaging with place value. *International Journal of Child-Computer Interaction*, *16*, 68–76. https://doi.org/10.1016/j.ijcci.2017.12.004

Bishop, J. P., Lamb, L. L., Philipp, R. A., Whitacre, I., Schappelle, B. P., & Lewis, M. L. (2014). Obstacles and affordances for integer reasoning: An analysis of children's thinking and the history of Mathematics. *Journal for Research in Mathematics Education*, 45(1), 19–61. https://doi.org/10.5951/jresematheduc.45.1.0019

Bouck, E. C., & Yadav, A. (2020). Providing access and opportunity for computational thinking and computer science to support mathematics for students with disabilities. *Journal of Special Education Technology*, 37(1), 151–160. https://doi.org/10.1177/0162643420978564

Brennan, K., & Resnick, M. (2012). New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American educational research association, Vancouver, Canada* (Vol. 1, p. 25). http://scratched.gse.harvard.edu/ct/files/AERA2012.pdf

Ching, Y.-H., Hsu, Y.-C., & Baldwin, S. (2018). Developing computational thinking with educational technologies for Young Learners. *TechTrends*, 62(6), 563–573. https://doi.org/10.1007/s11528-018-0292-7

Corbin, J., & Strauss, A. (2015). Basics of qualitative research: Techniques and procedures for developing. Sage.

Cui, Z., & Ng, O.-L. (2021). The Interplay between mathematical and computational thinking in primary school students' mathematical problem-solving within a programming environment. *Journal of Educational Computing Research*, *59*(5), 988–1012. https://doi.org/10.1177/0735633120979930

De Chenne, A., & Lockwood, E. (2022). A Task to connect counting processes to lists of outcomes in combinatorics. *The Journal of Mathematical Behavior*, 65, 100932. https://doi.org/10.1016/j.jmathb.2021.100932

DeJarnette, A. F. (2019). Students' challenges with symbols and diagrams when using a programming environment in mathematics. *Digital Experiences in Mathematics Education*, 5(1), 36–58. https://doi.org/10.1007/s40751-018-0044-5

Futschek, G. (2006). Algorithmic thinking: the key for understanding computer science. In *Informatics Education–The* Bridge between Using and Understanding Computers: International Conference in Informatics in Secondary Schools– Evolution and Perspectives, ISSEP 2006, Vilnius, Lithuania, November 7-11, 2006. Proceedings (pp. 159–168). Springer.

Gibson, J. J. (1979). The Ecological approach to visual perception. Houghton Mifflin Co.

Grizioti, M., & Kynigos, C. (2021). Code the mime: A 3D programmable charades game for computational thinking in MaLT2. *British Journal of Educational Technology*, 52(3), 1004–1023. https://doi.org/10.1111/bjet.13085

Grover, S., Cooper, S., & Pea, R. (2014). Assessing computational learning in K-12. In *Proceedings of the 2014 Conference* on Innovation & Technology in Computer Science Education (pp. 57–62). ACM.

Grover, S., Pea, R., & Cooper, S. (2015). Designing for deeper learning in a blended computer science course for middle school students. *Computer Science Education*, 25(2), 199–237. https://doi.org/10.1080/08993408.2015.1033142

Guzdial, M., & Soloway, E. (2003). Computer science is more important than calculus. *ACM SIGCSE Bulletin*, 35(2), 5–8. https://doi.org/10.1145/782941.782943

Hartson, R. (2003). Cognitive, physical, sensory, and functional affordances in interaction design. *Behaviour & Information Technology*, 22(5), 315–338. https://doi.org/10.1080/01449290310001592587

Hesse, F., Care, E., Buder, J., Sassenberg, K., & Griffin, P. (2015). A Framework for teachable collaborative problem-solving skills. In P. Griffin & E. Care (Eds.), *Assessment and teaching of 21st century skills: Methods and approach* (pp. 37–56). Springer.

Hickmott, D., Prieto-Rodriguez, E., & Holmes, K. (2018). A Scoping review of studies on computational thinking in K–12 mathematics classrooms. *Digital Experiences in Mathematics Education*, 4(1), 48–69. https://doi.org/10.1007/s40751-017-0038-8

Ho, W. K., Lool, C. K., Huang, W., Seow, P., & Wu, L. (2021). Computational thinking in mathematics: To be or not to be, that is the question. In *Mathematics—Connection and Beyond: Yearbook 2020 Association of Mathematics Educators* (pp. 205–234). https://doi.org/10.1142/9789811236983\_0011

Hsu, H., Wang, C., & Levesque-Bristol, C. (2019). Reexamining the impact of self-determination theory on learning outcomes in the online learning environment. *Education and Information Technologies*, 24(3), 2159–2174.

Jong, M. S. Y., Geng, J., Chai, C. S., & Lin, P. Y. (2020). Development and predictive validity of the computational thinking disposition questionnaire. *Sustainability*, *12*(11), 4459. https://doi.org/10.3390/su12114459

Jong, M. S. Y., Song, Y., Soloway, E., & Norris, C. (2021). Teacher professional development in STEM education. *Educational Technology & Society*, 24(4), 81–85.

Kafai, Y. B., & Burke, Q. (2014). Connected code: Why children need to learn programming. MIT Press.

Kafai, Y. B., & Peppler, K. (2011). Youth, technology, and DIY. Review of Research in Education, 35(1), 89–119. https://doi.org/10.3102/0091732x10383211

Kirschner P. A. (2002). Can we support CCSL? Educational, social and technological affordances for learning. In *Three worlds of CSCL: Can we support CSCL*? (pp. 7–34). The Open Universiteit Nederland.

Leung, A. (2020). Boundary crossing pedagogy in STEM education. *International Journal of STEM Education*, 7(1). https://doi.org/10.1186/s40594-020-00212-9

Lockwood, E., & De Chenne, A. (2019). Enriching students' combinatorial reasoning through the use of loops and conditional statements in Python. *International Journal of Research in Undergraduate Mathematics Education*, 6(3), 303–346. https://doi.org/10.1007/s40753-019-00108-2

Lockwood, E., & Mørken, K. (2021). A call for research that explores relationships between computing and mathematical thinking and activity in Rume. *International Journal of Research in Undergraduate Mathematics Education*. 7(3), 404–416. https://doi.org/10.1007/s40753-020-00129-2

Miller, J. (2019). STEM education in the primary years to support mathematical thinking: Using coding to identify mathematical structures and patterns. ZDM, 51(6), 915–927. https://doi.org/10.1007/s11858-019-01096-y

Ng, O.-L., & Chan, T. (2019). Learning as making: Using 3D computer-aided design to enhance the learning of shape and space in STEM-integrated ways. *British Journal of Educational Technology*, 50(1), 294–308. https://doi.org/10.1111/bjet.12643

Ng, O.-L., & Cui, Z. (2021). Examining primary students' mathematical problem-solving in a programming context: Towards computationally enhanced mathematics education. *ZDM – Mathematics Education*, *53*(4), 847–860. https://doi.org/10.1007/s11858-020-01200-7

Ng, O.-L., Liu, M., & Cui, Z. (2021). Students' in-moment challenges and developing maker perspectives during problembased digital making. *Journal of Research on Technology in Education*, 1–15. https://doi.org/10.1080/15391523.2021.1967817

Nordby, S. K., Bjerke, A. H., & Mifsud, L. (2022). Computational thinking in the primary mathematics classroom: A Systematic review. *Digital Experiences in Mathematics Education*, 8(1), 27–49. https://doi.org/10.1007/s40751-022-00102-5

Norman, D. A. (1999). Affordances, conventions, and design. *Interactions*, 6(3) 38–43. https://doi.org/10.1145/301153.301168

Papert, S. A. (1980). Mindstorms: Children, computers, and powerful ideas. Basic books.

Parnafes, O., & diSessa, A. A. (2013). Microgenetic learning analysis: A Methodology for studying knowledge in transition. *Human Development*, *56*(1), 5–37. https://doi.org/10.1159/000342945

Pei, C., Weintrop, D., & Wilensky, U. (2018). Cultivating computational thinking practices and mathematical habits of mind in Lattice Land. *Mathematical Thinking and Learning*, 20(1), 75–89. https://doi.org/10.1080/10986065.2018.1403543

Robins, A., Rountree, J., & Rountree, N. (2003). Learning and teaching programming: A review and discussion. *Computer Science Education*, *13*(2), 137–172. https://doi.org/10.1076/csed.13.2.137.14200

Román-González, M., Pérez-González, J. C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? criterion validity of the computational thinking test. *Computers in Human Behavior*, 72, 678–691. https://doi.org/10.1016/J.CHB.2016.08.047

Sneider, C., Stephenson, C., Schafer, B., & Flick, L. (2014). Exploring the science framework and NGSS: Computational thinking in the science classroom. *Science Scope*, *38*(3), 10–15. https://doi.org/10.2505/4/ss14\_038\_03\_10

So, H. J., Jong, M. S. Y., & Liu, C. C. (2020). Computational thinking education in the Asian Pacific region. *The Asia-Pacific Education Researcher*, 29(1), 1–8.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, 25(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Weng, X., Cui, Z., Ng, O. L., Jong, M. S. Y., & Chiu, T. K. F. (2022). Characterizing students' 4C skills development during problem-based digital making. *Journal of Science Education and Technology*, *31*(3), 372–385. https://doi.org/10.1007/s10956-022-09961-4

Wilkerson-Jerde, M. (2014). Construction, categorization, and consensus: Student generated computational artifacts as a context for disciplinary reflection. *Educational Technology Research and Development*, 62(1), 99–121. https://doi.org/10.1007/s11423-013-9327-0

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, 49(3), 33–35. https://doi.org/10.1145/1118178.1118215

Wing, J. M. (2011). Research notebook: Computational thinking-What and why. The Link Magazine, 6, 20-23.

Ye, H., Liang, B., Ng, O.-L., & Chai, C. S. (2023). Integration of computational thinking in K-12 mathematics education: A systematic review on CT-based mathematics instruction and student learning. *International Journal of STEM Education*, *10*(1). https://doi.org/10.1186/s40594-023-00396-w

Zur-Bargury, I., Parv, B., & Lanzberg, D. (2013). A Nationwide exam as a tool for improving a new curriculum. In *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education* (pp. 267–272). ACM.

## Appendix 1. Brief description of the tasks included in the study

*Two Savings problem*. There are two money-saving strategies. The first plan starts with \$3 and then deposits \$222 every day thereafter. The second plan starts with \$3 and then deposits double the amount of the last deposit (i.e., \$6, \$12) on each subsequent weekend (i.e., Saturday and Sunday). The problem asks students to determine which saving strategy is more optimal given different saving periods.

The mathematical concepts involved were arithmetic and geometric sequences. In solving the problem, the students were asked first to sketch a graph of the bank balance over time for each saving plan. Then, they were to model the respective saving processes with programming.

*Fibonacci Sequence Generator*. Observe the following number sequence: 1, 1, 2, 3, 5, 8, 13, ..., what is the next term? What is the 50<sup>th</sup> term? Create a program that could find any term in this sequence.

The mathematical concept involved was sequences. The students were prompted to create a program that included input and output. The input indicates the number of items in the sequence, and the program outputs the value of the corresponding term.

*Dice Rolling problem*. When we roll six dice together, we can obtain the sum of the results. If we have to guess the sum of the six dice given four choices, 19, 20, 21, or 22, which one should we choose?

This was a two-part problem. The first part was programming a dice-rolling simulator with computer-generated random numbers and observing the sum upon a certain number of simulations. The second part was generating the outcome space (e.g., 1-1-1-1-1, 1-1-1-1-2, ... 6-6-6-6-6-6), as well as the frequency of obtaining a given sum theoretically. For example, to obtain the sum of 6, the frequency was only one, that is, 1-1-1-1-1; but to obtain the sum of 7, the frequency was six (1-1-1-1-2, 1-1-1-2, ..., 2-1-1-1-1). The mathematical concept involved in the task was classic probability. However, because the students had not yet learned the concept of probability, we avoided using the term and let the students experiment with the concept on their own.

**Dart Throwing problem**. There is a rectangular dartboard made of square regions of different sizes (i.e., side lengths of 1, 1, 2, 3, 5, and 8). When throwing darts at the board, the darts will either hit one of the square regions or miss the board completely. Design a fair game scoring system to indicate the number of points a player should gain when the darts land on the different squares.

The mathematical concept involved was unequal likelihood outcome space. To solve the problem, the students would simulate throwing darts at this specific board a large number of times and record the frequencies of the darts hitting the various squares using programming.

*Prime Detector*. Is 7081 a prime number? Create a device that could determine whether a number, such as 7081, is a prime or composite number.

The mathematical concepts involved were divisibility rules, factors, and multiples. The students were introduced to the code "mod" to determine the remainder of a division operation.

*Counting to 21 (or 100)*. This is a game with two players. The players take turns calling either 1 or 2 (or 1 to 9 in the game of counting to 100), and the program will record and add all the numbers being called. For example, at the beginning, if Player A calls 2, the program will show 2; then, if Player B calls 1, the program will show 3, and so on. The player who gets the program to show 21 wins the game. The students were invited to, first, create the program and, then, play the game with their partners. Then, they would design a program with a computerized player in which the human player goes first such that the computerized player will always win.

In this two-part task, the mathematics involved observing that a winning strategy was to ensure that, upon the human opponent taking the first turn to call a number, the computer will call a number such that the sum is a multiple of three.

Drawing Polygons. Draw different regular polygons with the Pen function in Scratch.

The mathematical concepts involved were exterior and interior angles and the number of sides. The students were first shown how to draw an equilateral triangle, and then, they were to explore drawing various (regular) polygons.

*Drawing Fractals*. Observe the following geometric figures: a Sierpinski triangle, a fractal tree, and a Koch curve. What are some common features between them? Create a program to draw one of these fractal geometries and, then, design your own fractal geometry.

The mathematical concept involved in this task was recursion. The students were not introduced to the mathematical definition of recursion (for example, using a typical factorial example, which can be represented by the recursive formula,  $f(n) = f(n-1) \times n$ ); rather, they were instructed to conceptualize a recursion as "a function calling the function itself" in programming.

# The SNS-based E-mentoring and Development of Computational Thinking for Undergraduate Students in an Online Course

# Yeonju Jang<sup>1</sup>, Seongyune Choi<sup>1</sup>, Seonghun Kim<sup>2</sup> and Hyeoncheol Kim<sup>3\*</sup>

<sup>1</sup>Dept. of Computer Science and Engineering, The Graduate School of Korea University, Republic of Korea // <sup>2</sup>Dept. of Education, Gachon University, Republic of Korea // <sup>3</sup>Dept. of Computer Science and Engineering, Korea University, Republic of Korea // spring0425@korea.ac.kr // csyun213@korea.ac.kr // ryankim@gachon.ac.kr // harrykim@korea.ac.kr

\*Corresponding author

ABSTRACT: Given the importance of digital technology in daily life, computational thinking (CT) has become a necessary skill for everyone, not just for computer scientists. For CT development, students need to receive appropriate social learning support. However, instructors find it difficult to provide such support to many students in online courses. This study aimed to examine the effectiveness of e-mentoring via social network services (SNS) in developing students' CT during large-scale online courses. A total of 327 undergraduate students volunteered to participate in this study, which included 16 weeks of lectures aimed at developing CT. The effects of SNS-based e-mentoring on CT development, the influences of each e-mentoring activity on CT development, and gender differences were analyzed using data on participants' CT assistance scores and their utilization of e-mentoring activities. The findings indicated that SNS-based e-mentoring was effective in developing the CT of undergraduate students' engagement in a large-scale online course. The most influential ementoring activities for students' CT development were informational and technical support in a group and informational support in a private environment. Female students benefited more from SNS-based e-mentoring than male students, and they also engaged in more types of e-mentoring activities than male students. Participation in SNS-based e-mentoring was found to lower the gap in CT between students with and without prior learning experience. Our study findings can be used by educational institutions and instructors when designing courses for students' CT development in large-scale online courses or when developing strategies to close the gender gap in CT ability.

**Keywords:** Computational thinking, e-Mentoring, Social Network Service (SNS), Gender difference, Computational thinking and prior learning experience

# **1. Introduction**

Ever since its mention by Wing (2006), the interest in "computer thinking (CT)" has been growing steadily. Wing (2006) described CT as a "fundamental skill for everyone, not just for computer scientists" (p. 33). People now live in a world where digital technologies are used in various fields such as health care and education (Jung et al., 2022; Jang et al., 2022b; Choi et al., 2022). In a world where digital technology is critical for performing essential daily tasks, individuals must have the skills necessary to both understand critically the technological systems they use and solve problems when things go wrong (Czerkawski, 2015). Consequently, numerous studies on CT education have been conducted across a range of subjects, from K–12 (Angeli et al., 2016; Li et al., 2022) to higher education (Lyon & Magana, 2020; Jocius et al., 2021).

In CT development, problem-based or project-based learning strategies have been primarily used (Hsu et al., 2018), and instructor–student interaction is crucial for the learning process (Kwame Boateng, 2020). Through interactions with students, instructors can positively influence students' CT development by providing just-in-time instructions, role modeling, and other social learning support (Gong et al., 2020; Lye & Koh, 2014).

However, developing students' CT through active interactions between instructors and students is more difficult in large online than in small face-to-face classes. It is difficult for instructors and students to interact actively online (Drange et al., 2015), but it is even more difficult for instructors to interact with multiple students when there are many students to manage.

E-mentoring can be a solution to this problem. E-mentoring refers to a pairwise relationship between a more experienced individual (mentor) and a less experienced individual (mentee), primarily through electronic communication. E-mentoring can provide mentees with informational, psychosocial, and instrumental benefits (Single & Single, 2005), as well as alleviate the problem of lack of interaction between instructors and learners online (Dahalan et al., 2012). For e-mentoring to be effective, users should be comfortable using the mentoring tool (Sánchez et al., 2014). Additionally, when synchronous tools capable of real-time dialogue are used, the

effect of e-mentoring is enhanced by making communication more comfortable (Jacobs et al., 2015; Tanis & Barker, 2017). Therefore, it is important to select an appropriate e-mentoring tool to maximize its effect (Chong et al., 2020).

In this study, Social Networking Sites (SNS) were used to help students develop CT through e-mentoring in large-scale online classes. As SNS has become more common in daily life, attempts to use SNS for education have emerged (Lee & Kim, 2016; Rutten et al., 2016; Son et al., 2016). The advantage of using SNS for education is that users are already familiar with it, can readily share various data, and can interact in real time (Sánchez et al., 2014).

Several previous studies have tried to develop students' CT using e-mentoring or SNS. However, there are some research gaps with regard to how e-mentoring using SNS in large online courses affects CT. First, there are studies on how e-mentoring improves CT (Dlab et al., 2019); however, studies on how e-mentoring activities affect CT enhancement are lacking. In addition, there is a limitation that the e-mentoring process proceeded asynchronously as e-mail was used as a tool for mentoring. Therefore, it is necessary to conduct additional research on the effects of real-time e-mentoring via SNS on CT development. Second, a study that developed CT using SNS (Tsutsui & Takada, 2018) was conducted in an offline class with a small number of students. Consequently, the impact of using SNS for e-mentoring in large online courses needs to be investigated further. Third, one study used SNS for e-mentoring (Lee & Mehta, 2015), but it is unclear whether this helps students develop CT in large online courses. Finally, to the best of our knowledge, no research has been conducted to determine whether the effect of SNS-based e-mentoring on CT development differs depending on the gender or prior learning experience of students. Gender gaps in CT education are frequently mentioned (Angeli & Giannakos, 2020; Bati, 2022). Analyzing gender differences in the method or effect of using SNS-based ementoring in courses for CT development can provide insight into how to reduce the gender gap when designing CT education classes in the future. However, it is well known that when programming practice is included in a CT development course, students' prior learning experiences have a significant impact on their learning success (Bergersen & Gustafsson, 2011; Lau & Yuen, 2011; Jegede, 2009). Therefore, analyzing whether there is a difference in the effect of CT development through SNS-based e-mentoring based on prior learning experiences can be used as a reference when designing an e-mentoring program in the future while taking students' educational backgrounds into account. Consequently, we designed and conducted research on the following questions.

- RQ 1: Is SNS-based e-mentoring useful for college students' CT development?
- RQ 2: Which e-mentoring activities influence CT development?
- RQ 3: Is there a gender difference in the effects of SNS-based e-mentoring and e-mentoring activities?
- RQ 4: Is there a difference in the CT enhancement effect of SNS-based e-mentoring considering prior learning experience? Is there an interaction effect between prior learning experience and SNS-based e-mentoring?

To answer the research questions and accomplish our research goals, we investigated the effect of SNS-based ementoring on the CT development of college students in this study. An informatics course was conducted for 16 weeks with the goal of developing CT for students, and SNS-based e-mentoring was also conducted during this period. Data on students' CT abilities and their utilization of e-mentoring were collected during this process. Through data analysis, the effect of SNS-based e-mentoring on CT development, the effect of e-mentoring activities on CT development, and gender differences were investigated.

# 2. Literature review

## 2.1. Computational thinking

Although numerous attempts have been made to integrate CT into various fields of education, there are various opinions on its definition. Wing (2006) described CT as "solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science" (p. 33). Following that, she clarified CT as "the thought processes involved in formulating problems and their solutions so that the solutions are represented in a form that can be effectively carried out by an information-processing agent" (Wing, 2011, p. 20).

Aho (2012) defined CT as "the thought processes involved in formulating problems so their solutions can be represented as computational steps and algorithms" (p. 832) The Royal Society (2012) described CT as "the

process of recognizing aspects of computation in the world that surrounds us and applying tools and techniques from Computer Science to understand and reason about both natural and artificial systems and processes" (p. 29). Meanwhile, CT has been also defined as "reformulating a seemingly difficult problem into one we know how to solve, perhaps by reducing, embedding, transforming, or simulating" (Wing, 2006, p. 33).

Although there is currently no universally accepted definition of CT, researchers have come to accept that it is a thought process that incorporates elements of abstraction, generalization, decomposition, algorithmic thinking, and debugging (Angeli et al., 2016). Abstraction is the ability to strip away features or attributes from an object or entity to reduce it to a set of fundamental characteristics (Wing, 2011). While abstraction reduces complexity by concealing unessential details, generalization reduces complexity by substituting a single construct for multiple entities that perform similar functions (Thalheim, 2000). Abstraction and generalization are frequently used in combination, with abstracts generalized via parameterization to increase utility. Decomposition is the ability to reduce complex problems to their simplest components (National Research Council, 2010). Algorithmic thinking is a problem-solving skill that entails formulating a problem solution step-by-step (Selby, 2014). Debugging is the ability to identify when actions do not correspond to instructions and correct errors (Selby, 2014).

Table 1 shows the elements of CT as these have been discussed and defined in this section. This conceptual framework was referenced by Angeli et al. (2016). Accordingly, this conceptual framework was adopted for designing an informatics curriculum for undergraduate students to develop CT.

Table 1. The elements of CT			
Element	Definition		
Abstraction (AB)	The ability to determine which data about an entity/object to retain and which to discard (Wing, 2011).		
Generalization (GN)	The ability to formulate a solution in generic terms for it to be applicable to a variety of problems (Selby, 2014).		
Decomposition (DC)	The ability to decompose a complex problem into smaller, more manageable components (National Research Council, 2010; Wing, 2011).		
Algorithms (AL)	The ability to create a step-by-step sequence of operations/actions for resolving a problem (Selby, 2014).		
Debugging (DB)	The ability to identify, eliminate, and correct errors (Selby, 2014).		

## 2.2. E-mentoring

With the advancement of technology, especially the improvement of electronic communication, the concept of mentoring has been developed without face-to-face elements (Risquez, 2008; Single & Single, 2005). Single and Muller (2001) defined e-mentoring as a relationship or pairwise relationship that occurs naturally within the program, established between a more experienced individual (the mentor) and a less experienced individual (the mentee), mainly using electronic communication. Methods such as e-mail, threaded discussions through learning management systems (LMSs) and SNSs can be used for e-mentoring (Rowland, 2012).

According to Single and Single (2005), e-mentoring has informational, psychosocial, and instrumental benefits. Informational benefits refer to the exchange of knowledge and subject matter beneficial to a newcomer. Psychosocial benefits refer to mentees gaining self-esteem, confidence, and encouragement to take risks as a result of effective mentoring relationships. Instrumental benefits refer to relationships that provide mentees with opportunities for increased visibility and advancement. Instrumental benefit can also be defined in terms of behaviors targeted toward facilitating the mentee's goal attainment (Eby et al., 2013), or practical contributions (Gafni-Lachter et al., 2021). A previous study reported that students felt confused, anxious, and frustrated because of the lack of prompt feedback from instructors and vague instructions on websites (Hara & Kling, 2001). E-mentoring can alleviate this problem because students (mentees) and their mentors can interact regardless of location through email, chat rooms, bulletin boards, forums, and discussions (Dahalan et al., 2012). Several studies have shown that e-mentoring can help improve student performance (de Janasz & Godshalk, 2013; Jacobs et al., 2015).

E-mentoring has also been applied to the development of CT in students. For example, Kahraman and Abdullah (2016) used an online forum and e-mail-based communication tool to conduct e-mentoring, which facilitated the CT development of undergraduate students. In addition, Dlab et al. (2019) demonstrated that the CT of primary school students was developed as a result of e-mentoring using an LMS.

## 2.3. Social network service

Social network services (SNS) are a collection of web technologies that enable users to create, share, communicate, and interact with one another. SNS users can interact with "friends" or other users or members on and offsite who are invited to connect to their profile. Other connected users, referred to as "friends," "contacts," or "followers," can be anyone who is granted access to the user's profile (to view and share information), and friends can range from close family members to complete strangers (Weber, 2012).

As the use of social media continues to grow, attempts to use it for educational purposes have emerged. The use of SNS in learning involves the advantage of real-time, information sharing, simple posting, and reliable feedback from friends (Du et al., 2013; Popescu, 2014). SNS is effective for writing education (Lee & Kim, 2016) and can help adolescents develop their online career skills (Rutten et al., 2016). In addition, Son et al. (2016) proposed an LMS that enables real-time and reliable feedback for incorrect answers by incorporating an SNS. Tsutsui and Takada (2018) created an SNS platform for programming education and used it in a class for the CT development of students.

Considering the benefits of integrating SNS into education, this study incorporated SNS into the e-mentoring process. KakaoTalk used in this study was released in 2010 and is used by more than 93% of smartphone owners in South Korea (Lee & Kim, 2016).

# 3. Methods

## 3.1. The course implementation for developing undergraduate students' CT

The researchers in this study developed a CT program for undergraduate students through a 16-week online course. The course was implemented for students at Korea University in the Republic of Korea. Since 2014, the course has been open annually, and it was a relatively large course, with an average of 400 students enrolling each year. Due to the coronavirus (COVID-19) outbreak, the course was conducted entirely online and twice weekly for 90 min each. The course curriculum for 14 weeks is shown in Table 2, excluding the 2 weeks for the midterm/final exams. Each session included a problem-solving activity for CT development based on basic informatics concepts.

The course included lectures on the fundamental concepts of informatics as well as lectures on programming. All classes were conducted via video conferencing in real time (Zoom). At the beginning of the course, the students' programming experience was investigated. According to the survey, 48.0% (N = 157) of the students had programming experience (including block-based programming), but only 17.8% (N = 58) had text-based programming experience. Therefore, a cloud-based programming environment (Google Colaboratory) that does not require complicated environment settings was selected for the programming lectures. In addition, Python was chosen as the programming language because it is simple for beginners to learn. Programming assignments were given after each lecture. A lecture on training a machine learning model using Google Teachable Machine was held for a week. Additionally, by creating a webpage, students could check the results of the trained model directly. Figure 1 illustrates examples of students' primary activity outputs and a researcher-created web page.



Figure 1. Screenshots of students' activity results and a researcher-created web page

(b) A website designed to visualize machine learning modeling results.

		2. Curriculum of the implemented online c			~~~
Week	Topic	Main concept/contents	Main activity	Related element	CT .s
1	Computing machine	<ul> <li>Data, information, and knowledge</li> <li>Automation of information processing process</li> <li>Automation</li> <li>Problem solving with machines: abstraction, decomposition, algorithm</li> <li>History of computing tools</li> </ul>	<ul> <li>Distinguish between data and information</li> <li>Designing an automated machine that recognizes handwriting</li> </ul>	AB, AL	DC,
2-3	Data representation	<ul> <li>Representation of information using code</li> <li>Information theory, entropy</li> <li>Data encoding: number, text, image, sound</li> </ul>	<ul> <li>Create code to communicate using five fingers</li> <li>Calculating entropy</li> <li>Encoding characters, numbers, images, and sounds</li> </ul>	AB, DC, AL	GN,
4	Problem solving	<ul> <li>Problem solving: IPO (input-process- output), problem representation, problem decomposition</li> <li>Data modeling: decision table, entity- relation diagram (ERD), state machine, data flow diagram</li> </ul>	• Expressing a problem as a decision table, ERD, etc.	AB, DC, DB	GN, AL,
5	Algorithmic thinking	<ul><li>Algorithm: flowchart, pseudo-code, sequence, flow control</li><li>Algorithm and program</li></ul>	• Solve problems by expressing them as flowcharts and pseudo- code	AB, DC, DB	GN, AL,
6-7	Algorithm	<ul> <li>Data structure</li> <li>Sorting: selection, bubble, insertion, quick</li> <li>Searching: sequential, binary</li> </ul>	<ul> <li>Display your favorite soccer teams as an array</li> <li>Solving sorting problems</li> <li>Solving searching problems</li> </ul>	AB, DC, AL	GN,
8	Functional world	•Function, recursive function	• Representing a problem as a function	AB, DC, AL	GN,
9-11	Programming	•Python programming	• Creating basic programs in Python	AB, DC, DB	GN, AL,
12-13	Alternative computing	<ul> <li>Greedy algorithm</li> <li>Intelligent model: knowledge based, data based</li> <li>Evolutionary computing, genetic algorithm</li> <li>Game theory</li> </ul>	<ul> <li>Solve the problem by expressing it with a greedy algorithm</li> <li>Representing and solving problems with genetic algorithms</li> </ul>	AB, DC, AL	GN,
14	Machine learning	•Training classification models	• Create image classification and sound classification models using teachable machine	AB, DC, DB	GN, AL,

C /1 C. CT 4

# **3.2.** Participants

Participants in this study were undergraduate students from Korea University in South Korea with 50 different majors (e.g., computer science, philosophy, architecture) enrolled in the same informatics course. Participants were recruited using a voluntary response sampling method (Murairwa, 2015; Tiit, 2021; Jang et al., 2022a) that targeted the students who took this course. At the beginning of the course, the researchers investigated whether students desired e-mentoring and consented to participate in the research. Participation in e-mentoring was optional, but participants were required to write a mentoring report at the end of the course. A total of 380 students attended the course and 327 volunteered to participate in the study. Among the participants, 189 students engaged in e-mentoring and 138 did not.

We categorized the participants into two groups (control and group), depending on whether they engaged in ementoring. Therefore, students who did not participate in e-mentoring were assigned to the control group, whereas those who did were assigned to the treatment group.

Table 3 presents the demographics of the participants (mentees). Among the participants, 167 (51.1%) were male and 160 (48.9%) were female. Most of the participants were freshmen (N = 131; 40.1%), followed by seniors (N = 88; 26.9%) and sophomores (N = 78; 23.9%), with the least number of participants being juniors (N = 30; 9.1%). The total number of participants' majors was 43, with the largest number of participants majoring in computer science (N = 33), followed by mechanical engineering (N = 28) and new materials engineering (N = 24). In contrast, sociology (N = 3), architecture (N = 5), and psychology (N = 8) were the three majors with the fewest participants. Table 3 shows the characteristics of the participants by group according to their engagement in e-mentoring.

			<u> </u>
		Engaged in e-mentoring	Not engaged in e-mentoring
Gender	Male	<i>N</i> = 82; 49.1%	<i>N</i> = 85; 50.9%
	Female	<i>N</i> = 107; 66.9%	<i>N</i> = 53; 33.1%
The top three	majors the most	Business $(N = 18)$	Computer science $(N = 22)$
students have		Biology ( $N = 15$ )	Mechanical engineering $(N = 16)$
		Electronic engineering $(N = 14)$	Mathematics $(N = 13)$

Table 3. Participant demographics by group according to the engagement in e-mentoring

## **3.3. E-mentoring process**

## 3.3.1. Recruitment

Mentors and mentees were recruited concurrently during the first week of the course. First, students enrolled in the course were invited to apply for e-mentoring via Google Forms. E-mentors were recruited from among the students who took this course in the prior semester through an e-mail to students who received A0 or A+ grades. Out of a total of 127 students, 19 hoped to participate as mentors. Twelve mentors were selected through online interviews. Of the mentors, seven were male and five were female. Mentors mainly majored in computer science (N = 5), and pre-medical (N = 3). The remaining mentor majors were electronic engineering, industrial management engineering, economics, and psychology (each N = 1). A total of 189 mentees (the treatment group) were assigned to the mentors, with each mentor assigned 15–16 mentees. All mentors agreed to participate in the study.

## 3.3.2. SNS-based e-mentoring environment

All e-mentoring processes were conducted online through an SNS platform. Figure 2 shows an SNS-based ementoring environment. The mentors were included in three chat rooms. The first was a private chat room with individual mentees, the second was a group chat room with matched mentees, and the third was a group chat room with instructors and other mentors. In the chat rooms, both mentors and mentees had to use an account with their real name and could not join anonymously. Figure 3 shows screenshots of the mentor–mentee group chat and private chat conducted in this study.





## 3.3.3. E-mentoring activities

E-mentoring can provide informational, psychosocial, and instrumental support to mentees (Single & Single, 2005). In this study, practical benefits provided by instrumental support were limited to "technical benefits." The curriculum in this study was designed to use various web tools when conducting programming tasks. Consequently, it was planned that e-mentors could assist students with any problems they might encounter while using these tools.

Accordingly, mentors provide informational, psychosocial, and technical support to their mentees. Table 4 shows the e-mentoring activities used in this study. To begin with, with regard to informational activities, e-mentors provided knowledgeable assistance in responding to mentees' inquiries about their comprehension of class content. Mentors were not allowed to directly answer the assignment questions. Instead, when students encountered difficulties completing assignments, hints or supplementary materials to assist with problem solving were provided via SNS. In addition, e-mentors responded to questions seeking general information about course attendance (e.g., assignment submission form). Second, mentors performed psychosocial activities. When mentees felt frustrated while taking a course or wanted to give up, mentors provided emotional support. For example, they said words that encouraged students, inspired confidence, and shared the difficulties and overcoming processes they had experienced while taking the course. Third, mentors carried out the technical activities. When mentees asked for assistance with using Google Colaboratory, Teachable Machine, and other tools, mentors suggested appropriate solutions.

Students who participated in SNS-based e-mentoring (treatment group) received informational, psychosocial, and technical support through an e-mentor, either in a group or privately. In contrast, students who did not participate in e-mentoring (control group) contacted the instructor directly via e-mail when informational, psychosocial, and technical assistance were required.

Table 4. E-mentoring activities					
Interaction target	Category of e-mentoring activity	Details of activity			
Mentee	Informational activity (IA)	Provide hints or supplementary materials for problem solving			
		Answer questions about class content			
	Psychosocial activity (PA)	Encouragement, role modeling			
	Technical activity (TA)	Help with the skills or tools mentees need on an assignment			

Mentors also interacted with other mentors. Figure 4 shows the overall activity of the e-mentors according to the interaction target. To begin, mentors shared useful materials that would aid mentees in their learning. Mentors also interacted with the instructor, and mentors sought answers from the instructor to mentees' difficult-to-answer questions. In addition, they were responsible for submitting a weekly report to the instructor detailing their interactions with the mentee.

Figure 4. E-mentoring activities by interaction target



#### 3.4. Data collection and data analysis

#### 3.4.1. Assessment of CT

For the CT assessment, course assignments, quizzes, and midterm/final exam scores were used. After converting homework, quizzes, and midterm/final exam scores into a scale of 10, the average value of the total score was used. The purpose of the course was to develop CT for undergraduate students by incorporating the fundamental concepts of informatics. Therefore, all assignments, quizzes, and exams were designed to assist students in developing CT through the resolution of problems related to the lecture's topic. The assignment included nine problem-solving tasks involving informatics and seven programming tasks. The quiz was conducted a total of three times using the quiz function of LMS. The midterm/final exam is not just a test of students' knowledge of fundamental informatics concepts; it is designed to assess their overall CT ability. Table 5 shows examples of the data used for the CT ability measurement.

Table 5. Examples of resources used for CT assessment

Content Related CT elements Assignment -Create code to communicate using five fingers AB, GN, DC, AL -Proposing a structure to efficiently organize photos in a AB, GN, DC smartphone photo album -Sorting: Select, Insert, Bubble, Quick Sort AB, GN, DC, AL -Python programming: find the cause of the error and fix it AL, DB correctly -Python programming: Creating a fractal pattern using the AB, GN, DC, AL, DB turtle module and nested loops -Creating Image and Sound Classification Models with AB, GN, DC, AL, DB **Teachable Machines** AB, GN, DC, AL Quiz -Data representation, problem solving -Algorithmic thinking, algorithm AB, GN, DC, AL, DB -Alternative computing AB, GN, DC, AL Midterm Exam - Imagine making a swing hanging on a tree, and explain step-AB, GN, DC, AL, DB by-step how to make a wooden swing so that someone else can make a swing exactly the way you imagined it. - Using "nodes" and "links" to represent the operational form of this course, which is being taught in a non-face-to-face format because of the coronavirus disease. - Expressing the algorithm for finding the same mate in a pile of socks in pseudocode Final Exam - Expressing the Fibonacci sequence in the form of a recursive AB, GN, DC, AL, DB function in pseudo-code - Structuring and expressing how a valet parking agent stores customers' cars and quickly finds and delivers the right car when the customer wants it

given data

## 3.4.2. Utilization of e-mentoring

The frequency of use of each mentee's e-mentoring activity was measured using the weekly activity report submitted by e-mentors. Screenshots of all conversations each mentor had with the mentors via group chat or private chat for a week were attached to the weekly activity report submitted by e-mentors. The conversation between e-mentors and e-mentees included questions and answers about incomprehensible parts of the class, questions and answers about the format of the assignment, questions and answers about error handling during python programming, questions and answers about Blackboard LMS access errors, and how to overcome programming as a non-major. The researchers classified and labeled the activities depicted in the report into three types of e-mentoring activities (IA, PA, and TA), noting their frequency. Content analysis was used for labeling (Clark et al., 2018). First, using weekly activity reports, two researchers independently classified each mentee's conversations with e-mentors on SNS as IA, PA, or TA and recorded the frequency. Afterward, they discussed their classification results to reach a consensus on all e-mentoring activities.

## 3.4.3. Demographics

Researchers collected demographic data, which included 10-digit students' IDs, gender, grades/year, and majors.

## 3.4.4. Students' educational background

In this study, we focused on whether students had programming experience (including block or text-based) in their educational backgrounds. An online survey was conducted at the start of the course using Google Forms to determine whether students had programming experience.

The following is why, among the students' educational backgrounds, we focused on programming experience rather than major. First, programming is known to be difficult for many undergraduate students (Ambrósio et al., 2011; Askar & Davenport, 2011; Hawi, 2010), and previous programming experience plays an important role in programming success (Bergersen & Gustafsson, 2011; Lau & Yuen, 2011; Jegede, 2009). Second, based on the demographics of the students, only approximately 10% (N = 33) majored in computer science, with the majority of students not majoring in computer science. Furthermore, approximately 40% (N = 131) of the students were freshmen who had just started college. Therefore, we determined that prior programming experience was the most important educational background factor given the nature of the course in which this study was conducted.

## 3.4.5. Data analysis

To determine the effect of SNS-based e-mentoring on CT, the CT assessment scores of those in the mentoring group and those in the comparison group (not engaged in e-mentoring) were compared using an independent sample t-test. Additionally, multiple linear regression analysis was used to determine the influence of each e-mentoring activity conducted in a group and private environment on CT. Finally, we examined whether the effect of SNS-based e-mentoring on CT differed by gender. Two-way analysis of variance (ANOVA) and two multiple linear regression analyses were conducted for this purpose. Data were statistically analyzed using SPSS 26.0, and the alpha level was set at 0.05.

# 4. Results

#### 4.1. Effects of SNS-based e-mentoring on CT development

RQ1 was to explore potential differences between groups of students who have attended SNS-based e-mentoring and those who have not. As described in Table 6, descriptive statistics showed that the group of students who participated in e-mentoring acquired a mean CT score of 8.475 with 0.531 *SD*, while their counterparts acquired a mean value of 7.186 with 0.637 *SD*.

An independent *t*-test was conducted to identify whether the differences were significant. First, we examined the normality of the data distribution with skewness and kurtosis. As a multivariate normal distribution, all items satisfied the absolute values of skewness (< 3) and kurtosis (< 8) (Kline, 2005). The independent sample *t*-test

showed a significant difference in CT scores between the two groups of students. Levene's test did not assume homogeneity of variance (F = 4.689, p = .031); the t-value was 19.32, and the *p*-value was < .001.

Table 0. Means and standard deviations for the students CT score				
	Ν	Mean	SD	
Engaged in E-mentoring	189	8.475	0.531	
Not engaged in E-mentoring	138	7.186	0.637	

Table 6 Means and standard deviations for the students' CT score

#### 4.2. Influences on CT by e-mentoring activity

RQ2 was to investigate how each e-mentoring activity affects students' CT. First, each e-mentoring activity was categorized based on the environment in which the interaction took place (group chat or private chat). Thus, six independent variables were considered. As depicted in Table 7, GI was found to be the most utilized activity among participants (M = 5.047, SD = 3.826), followed by PI (M = 3.968, SD = 4.034) and GT (M = 3.074, SD = 3.074, 2.508). In contrast, psychosocial activity showed relatively less utilization compared to other activities with GP (M = 0.021, SD = 0.144) and PP (M = 0.238, SD = 0.506).

Then, multiple linear regression was conducted to analyze the influence of each activity. Tolerance and VIF were assessed to exclude multicollinearity, and the values of all constructs were acceptable (Hair et al., 2010). The values of Durbin–Watson have an upper limit of four and a lower limit of zero (Niresh & Thirunavukkarasu, 2014). In addition, the data were found to be normally distributed (Kline, 2005).

As Table 8 shows, the result of multiple linear regression analysis, all variables were found to be statistically positive with the model explaining 50.7% of the variance in the CT. In addition, the model acquired an acceptable Durbin-Watson value (1.098), indicating that there were no independent errors caused by the residuals (Field, 2013). The three most influential determinants were GI ( $\beta = 0.540$ ), PI ( $\beta = 0.436$ ), and GT ( $\beta =$ 0.244). In contrast, PP ( $\beta = 0.119$ ), GP ( $\beta = 0.127$ ), and PT ( $\beta = 0.132$ ) were the three least influential determinants.

Table 7. Means and standard deviations for the utilization of e-mentoring activities

E-mentoring activity	Ν	Mean	SD
Group-Informational activity (GI)	189	5.047	3.826
Group-Psychosocial activity (GP)	189	0.021	0.144
Group-Technical activity (GT)	189	3.074	2.508
Private-Informational activity (PI)	189	3.968	4.034
Private-Psychosocial activity (PP)	189	0.238	0.506
Private-Technical activity (PT)	189	2.153	1.523

Table 8. Results of multiple linear regression									
Dependent	Independent	В	SE	β	t	Tolerance	VIF		
variable	variable								
СТ	(Constant)	7.57	0.075		100.364***				
	GI	0.075	0.008	0.540	$9.056^{***}$	0.739	1.353		
	GP	0.469	0.193	0.127	$2.427^{*}$	0.955	1.047		
	GT	0.052	0.013	0.244	4.139***	0.755	1.324		
	PI	0.058	0.008	0.436	7.281***	0.731	1.368		
	PP	0.125	0.057	0.119	$2.203^{*}$	0.901	1.11		
	PT	0.046	0.019	0.132	$2.48^{*}$	0.923	1.084		
$R(723)$ $R^2(522)$ adjusted $R^2(507)$ $E(33,159)$ $n < 001$									

*Note.* \*\*\* *p* < .001, \**p* < .05.

#### 4.3. Gender differences regarding the effect of e-mentoring using SNS on CT

RQ3 aimed to determine whether the effect of SNS-based e-mentoring on CT differs by gender. The researchers used a two-way ANOVA and two multiple linear regression analyses. First, participants in this study were classified into four groups according to their gender and whether they engaged in e-mentoring. Descriptive statistics showed that e-mentoring engaged males achieved the highest score on CT assessment (M = 8.545, SD =0.535), followed by e-mentoring engaged females (M = 8.421, SD = 0.524), not engaged males (M = 7.335, SD =

0.598), and not engaged females (M = 6.948, SD = 0.632). The two groups with the highest CT scores were those who engaged in e-mentoring. Table 9 describes the results of the descriptive statistics.

<i>Tuble 9.</i> Means and standard deviations for each group					
Gender	E-mentoring	N	Mean	SD	
Male	Engaged	82	8.545	0.535	
	Not engaged	85	7.335	0.598	
Female	Engaged	107	8.421	0.524	
	Not engaged	53	6.948	0.632	
	Total	327	7.931	0.572	

Table 9. Means and standard deviations for each group

Because there were two independent variables, a two-way ANOVA was performed to investigate the main and interaction effects on the dependent variables. The data were normally distributed according to Kline (2005). As demonstrated in Table 10, both participation in e-mentoring (p < .001) and gender p < .001) had a significant effect on students' CT. As depicted in Figure 5, an interaction effect was also observed between gender and participation in e-mentoring (p < .05).

Table 10. Result of two-way ANOVA								
Source	Type III sum	df	Mean square	F	Partial eta			
	of squares				squared			
Corrected model	137.997	3	45.999	143.789**	0.572			
Intercept	18717.22	1	18717.22	$58508.462^{**}$	0.995			
E-mentoring	137.96	1	137.96	431.252**	0.572			
Gender	5.031	1	5.031	15.728**	0.046			
E-mentoring * Gender	1.316	1	1.316	4.113*	0.013			
Error	103.33	323	0.32					
Total	20810.875	327						
Corrected total	241.327	326						
<i>Note.</i> $R^2 = .572$ (adjusted $R^2 = .568$ ). ** $p < .001$ , * $p < .05$ .								

Figure 5. Interaction effect plot of e-mentoring and gender Estimated Marginal Means of CT



Two multiple linear regression analyses were conducted to determine whether the effect of each e-mentoring activity on CT differed by gender. First, we investigated whether the two sets of data were suitable for regression analysis. First, based on the Durbin–Watson value, both models showed no multicollinearity problems (Niresh & Thirunavukkarasu, 2014). Second, according to Kline (2005), the data were normally distributed.

Descriptive statistics showed that male students utilized informational activity through group chat the most (M = 7.817, SD = 3.916), followed by technological activity through group chat (M = 4.512, SD = 2.911). In contrast, female students mostly used informational activity through private chat (M = 6.514, SD = 3.673), followed by informational activity through group chat (M = 2.925, SD = 2.894). Table 11 demonstrates the results of the descriptive statistics.
As Table 12 indicates, the result of multiple linear regression analysis for male students, three variables were statistically positive with the model explaining 50.9% of the variance in the CT. The three influential determinants were GI ( $\beta = 0.395$ ), PI ( $\beta = 0.370$ ), and GT ( $\beta = 0.180$ ). In the case of female students, the results of multiple linear regression analysis showed that all variables were statistically positive, except for GP. The model explained 56.9% of the variance in CT. The most influential determinant was PI ( $\beta = 0.356$ ), followed by GI ( $\beta = 0.339$ ), GT ( $\beta = 0.272$ ), PP ( $\beta = 0.189$ ), and PT ( $\beta = 0.173$ ).

Gender	E-mentoring activity	N	Mean	SD
Male	GI	82	7.817	3.916
	GP	82	0.024	0.155
	GT	82	4.512	2.911
	PI	82	0.646	0.616
	PP	82	0.122	0.329
	PT	82	1.682	1.142
Female	GI	107	2.925	1.941
	GP	107	0.018	0.136
	GT	107	1.972	1.362
	PI	107	6.514	3.673
	PP	107	0.327	0.595
	PT	107	2.514	1.678

. . . ..

Table 11. Means and standard deviations for e-mentoring activity utilization of both genders

	<i>Table 12.</i> Results of multiple linear regression							
Dependent	Gender	Independent	В	SE	β	t	Tolerance	VIF
variable		variable						
СТ	Male	(Constant)	7.754	0.118		65.472***		
		GI	0.054	0.012	0.395	$4.401^{***}$	0.754	1.326
		GP	0.336	0.284	0.097	1.181	0.895	1.117
		GT	0.033	0.015	0.180	$2.153^{*}$	0.872	1.147
		PI	0.321	0.078	0.370	4.095***	0.745	1.342
		PP	-0.043	0.13	-0.026	-0.33	0.956	1.046
		РТ	0.006	0.037	0.012	0.159	0.994	1.006
	$R(.738), R^{2}($	.545), adjusted	$R^{2}(.509), F$	r(14.968), p	<.001, Dur	bin–Watson =	1.042	
	Female	(Constant)	7.417	0.094		78.747***		
		GI	0.092	0.018	0.339	$5.006^{***}$	0.885	1.13
		GP	0.334	0.253	0.087	1.321	0.945	1.058
		GT	0.105	0.026	0.272	$4.046^{***}$	0.899	1.113
		PI	0.051	0.01	0.356	5.063***	0.823	1.215
		PP	0.167	0.059	0.189	$2.849^{**}$	0.921	1.085
		PT	0.054	0.021	0.173	$2.625^{*}$	0.933	1.071
	$R(.770), R^{2}($	.593), adjusted	$R^{2}(.569), F$	(24.289), p	< .001, Dur	bin–Watson =	1.375	

*Note.*  $^{***}p < .001$ ,  $^{**}p < .01$ ,  $^{*}p < .05$ .

### 4.4. Differences in the effect of e-mentoring via SNS on CT based on previous learning experience

RQ4 was to determine whether the effect of SNS-based e-mentoring on CT differed according to previous programming experience. The researchers used two-way ANOVA. First, participants were classified into four groups according to their previous programming experience and whether they engaged in e-mentoring. Descriptive statistics showed that e-mentoring engaged students with programming experience achieved the highest score on CT assessment (M = 8.789, SD = 0.435), followed by e-mentoring engaged students with no programming experience (M = 8.259, SD = 0.484), not engaged students with programming experience (M = 7.634, SD = 0.347), and not engaged students with no programming experience (M = 6.569, SD = 0.612). The two groups with the highest CT scores were those who engaged in e-mentoring. Table 13 lists the results of the descriptive statistics.

Because there were two independent variables, a two-way ANOVA was performed to investigate the main and interaction effects on the dependent variables. The data were normally distributed according to Kline (2005). As shown in Table 14, both participation in e-mentoring (p < .001) and programming experience (p < .001) were found to have a significant effect on students' CT. An interaction effect was also observed between programming

experience and participation in e-mentoring (p < .05). Figure 6 shows the interaction effect plot of e-mentoring and programming experience.

Table	Table 13. Means and standard deviations for each group								
Previous Learning Experience	E-mentoring	Ν	Mean	SD					
Have programming experience	Engaged	77	8.789	0.435					
	Not engaged	80	7.634	0.347					
No programming experience	Engaged	112	8.259	0.484					
	Not engaged	58	6.569	0.612					
	Total	327	7.931	0.572					

Table 13. Means and standard deviations for each group

Table 14. Result of Two-way ANOVA									
Source	Type III sum	df	Mean square	F	Partial eta				
	of squares				squared				
Corrected model	164.939a	3	54.98	$198.113^{**}$	0.648				
Intercept	18972.083	1	18972.083	68363.756**	0.995				
Programming experience	32.542	1	32.542	117.262**	0.266				
E-mentoring	149.193	1	149.193	537.599**	0.625				
E-mentoring * Programming experience	1.231	1	1.231	$4.434^{*}$	0.014				
Error	89.638	323	0.278						
Total	20824.125	327							
Corrected total	254.577	326							
<i>Note.</i> $R^2 = .648$ (adjusted $R^2 = .645$ ). ** $p < .645$	.001, *p < .05.								







# 5. Discussion

The independent sample *t*-test revealed a significant difference in students' CT scores depending on whether they participated in SNS-based e-mentoring. This result demonstrates the possibility of SNS-based e-mentoring. This result is consistent with the findings of previous studies. For example, Dlab et al. (2019) showed that utilizing an LMS as an e-mentoring environment is an efficient way of fostering participants' CT. However, an LMS cannot provide real-time interaction. As Grant et al. (2020) stressed, e-mentoring could be advantageous when mentormentee interaction occurs anytime and anyplace. In addition, Tsutsui and Takada (2018) applied a real-time SNS platform and showed that it is an effective interaction tool to develop CT for K–12 students, but the study was restricted only to small classes with five to seven students.

In this study, the e-mentoring method was used to assist students in developing CT during large online courses, and an SNS tool was introduced to facilitate quick interaction with them. Additionally, by concurrently facilitating group and individual interactions between mentors and mentees, students can utilize mentoring at

their convenience and inclination. It was found that SNS-based e-mentoring was helpful for students' CT development.

Mentees primarily engaged in informational activities, and this e-mentoring activity was most frequently used in both group and private environments. Informational activities also had the greatest influence on CT development. This result is in line with previous studies. In a study using e-mentoring in education, mentees required the most informational support (Cassiani, 2017). Additionally, as a result of analyzing the textual data from the discussion forums of the mentoring group, the only activity identified was informational (Cassiani et al., 2020).

Mentees' psychosocial activity utilization was low in both group interactions and the private environment. In this regard, Kaufman (2017) asserted that e-mentoring necessitates the ability to disclose and share emotions online, and psychosocial activity is difficult to achieve without these abilities. Psychosocial activity benefits role modeling, self-esteem, and learning motivation and has a positive effect on CT development (Lye & Koh, 2014; Gong et al., 2020). Therefore, it was determined that an e-mentoring program should be designed with this point in mind. To facilitate active psychosocial activity in e-mentoring, it is helpful to engage in a personal acquaintance process that includes introductions and searching for mutual interest (Shpigelman et al., 2009).

Gender gaps in CT education are an issue that has consistently been addressed (Angeli & Giannakos, 2020; Bati, 2022). Previous research has produced conflicting findings regarding whether there is a gender difference in CT ability. According to some research, males demonstrated greater CT ability than females, or females required more time to achieve the same level of CT ability (Atmatzidou & Demetriadis, 2016; Jenson & Droumeva, 2016). However, some studies assert that there is no gender difference in CT ability and that females demonstrate superior ability in certain elements of CT (Lee et al., 2017; Wu & Su, 2021). In this study, we found an interaction effect between gender and e-mentoring on students' CT. The mean difference in CT scores between males and females was greater in the group that did not engage in e-mentoring (mean values of male students = 7.335, mean values of female students = 6.948). However, as illustrated in Figure 5, the effect of SNS-based e-mentoring on female students' CT development was greater. This finding implies that SNS-based e-mentoring can contribute to closing the gap in CT abilities between male and female students.

Female students appeared to benefit more from SNS-based e-mentoring because they engaged in various ementoring activities more frequently than male students. In particular, female students utilized the psychosocial activity of e-mentoring better than male students, which was consistent with previous research findings (Elliott et al., 2010) And this result is significant because psychosocial activities such as role modeling also contribute to the development of CT (Gong et al., 2020; Lye & Koh, 2014).

Among the various educational activities for developing CT, it is well known that students' prior programming experience has a significant impact on their learning success (Bergersen & Gustafsson, 2011; Jegede, 2009; Lau & Yuen, 2011). Similar to previous studies' findings, there was a statistically significant difference in students' CT scores based on prior learning experience (i.e., programming experience) in this study. It seems that programming is an activity that requires all of the CT elements (AB, GN, DC, AL, DB), and thus, students with programming experience encountered more CT elements.

This study discovered that students' prior learning experience and SNS-based e-mentoring had an interaction effect on their CT. In other words, the difference in CT based on prior learning experience narrowed when students participated in SNS-based e-mentoring. Based on this finding, it is suggested that introducing SNS-based e-mentoring can bridge the gap between students' prior learning experiences when running courses for students from various educational backgrounds, with the goal of developing CT.

# 6. Conclusion and implication

As CT has become an essential basic skill, several studies have been conducted on its development. The interaction between the instructor and the student is critical in the development of CT. Through interaction with the instructor, students can receive a variety of support, including explicit instruction and role modeling. However, instructors find it difficult to actively interact with individual students in large-scale online courses. Consequently, this study examined the effect of e-mentoring in a large-scale online course aimed at developing students' CT through the use of SNS, which is capable of real-time interaction.

An independent sample *t*-test and multiple linear regression analysis were performed based on the participants' CT assessment scores and the utilization data of e-mentoring activities. The analysis determined that SNS-based e-mentoring is effective in assisting undergraduate students' CT development during a large-scale online course. The most beneficial e-mentoring activities for students' CT development were informational and technical support in the group environment, as well as informational support in the private environment.

To investigate whether there were any gender differences regarding the effect of SNS-based e-mentoring on CT development, a two-way ANOVA analysis, and two multiple linear regression analyses were performed. It was found that the effect of SNS-based e-mentoring was higher for female students than for male students. Additionally, female students engaged in more types of e-mentoring activities than male students.

A two-way ANOVA was used to determine whether the effect of SNS-based e-mentoring on CT development differed depending on students' prior learning experiences. The analysis revealed that participation in SNS-based e-mentoring could narrow the CT gap based on prior learning experience.

Our findings have practical implications for higher education institutions and instructors. First, when planning a course for students' CT development in a large-scale online course, a method of utilizing SNS-based e-mentoring can be considered. Second, the method used in this study can be applied when developing strategies to close the gender gap or the gap in students' prior learning experiences in CT ability.

# References

Angeli, C., & Giannakos, M. (2020). Computational thinking education: Issues and challenges. *Computers in Human Behavior*, 105, 106185. https://doi.org/10.1016/j.chb.2019.106185

Ambrósio, A. P., Costa, F. M., Almeida, L., Franco, A., & Macedo, J. (2011). Identifying cognitive abilities to improve CS1 outcome. In 2011 Frontiers in Education Conference (FIE) (pp. F3G-1). IEEE. https://doi.org/10.1109/FIE.2011.6142824

Askar, P., & Davenport, D. (2009). An Investigation of factors related to self-efficacy for Java Programming among engineering students. *Turkish Online Journal of Educational Technology*, 8(1), Article 3.

Atmatzidou, S., & Demetriadis, S. (2016). Advancing students' computational thinking skills through educational robotics: A Study on age and gender relevant differences. *Robotics and Autonomous Systems*, *75*, 661-670.

Aho, A. V. (2012). Computation and computational thinking. Computer Journal, 55, 832-835.

Angeli, C., Voogt, J., Fluck, A., Webb, M., Cox, M., Malyn-Smith, J., & Zagami, J. (2016). A K-6 computational thinking curriculum framework: Implications for teacher knowledge. *Educational Technology & Society*, 19(3), 47-57.

Bati, K. (2022). A Systematic literature review regarding computational thinking and programming in early childhood education. *Education and Information Technologies*, 27(2), 2059-2082.

Bergersen, G. R., & Gustafsson, J. E. (2011). Programming skill, knowledge, and working memory among professional software developers from an investment theory perspective. *Journal of individual Differences*, 32(4). https://doi.org/10.1027/1614-0001/a000052

Cassiani, C. M. (2017). A Peer e-mentoring employment program for youth with physical disabilities: What are the social supports that are needed and provided? (Unpublished doctoral dissertation). University of Toronto, Canada.

Cassiani, C., Stinson, J., & Lindsay, S. (2020). E-mentoring for youth with physical disabilities preparing for employment: A Content analysis of support exchanged between participants of a mentored and non-mentored group. *Disability and Rehabilitation*, 42(14), 1963-1970.

Chong, J. Y., Ching, A. H., Renganathan, Y., Lim, W. Q., Toh, Y. P., Mason, S., & Krishna, L. K. (2020). Enhancing mentoring experiences through e-mentoring: a systematic scoping review of e-mentoring programs between 2000 and 2017. *Advances in Health Sciences Education*, 25(1), 195-226.

Clark, R., Kaw, A., Lou, Y., Scott, A., & Besterfield-Sacre, M. (2018). Evaluating blended and flipped instruction in numerical methods at multiple engineering schools. *International Journal for the Scholarship of Teaching and Learning*, *12*(1), 1–16. https://doi.org/10.20429/ijsotl.2018.120111

Choi, S., Jang, Y., & Kim, H. (2022). A deep learning approach to imputation of dynamic pupil size data and prediction of ADHD. *International Journal on Artificial Intelligence Tools*. https://doi.org/10.1142/S0218213023500203

Czerkawski, B. (2015). Computational thinking in virtual learning environments. In *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2015* (pp. 993-997). Association for the Advancement of Computing in Education (AACE).

Dahalan, N., Hassan, H., & Atan, H. (2012). Student engagement in online learning: Learners attitude toward e-mentoring. *Procedia-Social and Behavioral Sciences*, 67, 464-475.

de Janasz, S. C., & Godshalk, V. M. (2013). The Role of e-mentoring in protégés' learning and satisfaction. Group & Organization Management, 38(6), 743-774.

Dlab, M. H., Hoić-Božić, N., Anđelić, M., & Botički, I. (2019). Digital games and tools for development of computational thinking in primary school. In *Proceedings of the International Conference on Management, Economics & Social Science-ICMESS* (pp. 1-7).

Drange, T., Sutherland, I., & Irons, A. (2015). Challenges of interaction in online teaching: A Case study. In *Proceedings of the International Conference on E-Technologies and Business on the Web* (pp. 35-42). The Society of Digital Information and Wireless Communications (SDIWC).

Du, Z., Fu, X., Zhao, C., Liu, Q., & Liu, T. (2013). Interactive and collaborative e-learning platform with integrated social software and learning management system. In W. Lu, G. Cai, W. Liu, & W. Xing (Eds.), *Proceedings of the 2012 International Conference on Information Technology and Software Engineering* (pp. 11-18). https://doi.org/10.1007/978-3-642-34531-9\_2

Eby, L. T. D. T., Allen, T. D., Hoffman, B. J., Baranik, L. E., Sauer, J. B., Baldwin, S., Morrison, M. A., Kinkade, K. M., Maher, C. P., Curtis, S., & Evans, S. C. (2013). An Interdisciplinary meta-analysis of the potential antecedents, correlates, and consequences of protégé perceptions of mentoring. *Psychological Bulletin*, *139*(2), 441-476. https://doi.org/10.1037/a0029279

Elliott, B. A., Dorscher, J., Wirta, A., & Hill, D. L. (2010). Staying connected: Native American women faculty members on experiencing success. *Academic Medicine*, 85(4), 675-679.

Field, A. (2013). Discovering statistics using IBM SPSS statistics. Sage.

Gafni-Lachter, L., Niemeyer, L., Doyle, N., Norcross, J., & Jacobs, K. (2021). Equal peer e-mentoring for online graduate students: a case study and mediation model. *Mentoring & Tutoring: Partnership in Learning*, 29(5), 545-564.

Gong, D., Yang, H. H., & Cai, J. (2020). Exploring the key influencing factors on college students' computational thinking skills through flipped-classroom instruction. *International Journal of Educational Technology in Higher Education*, 17(1), 1-13.

Grant, M. A., Bloom, G. A., & Lefebvre, J. S. (2020). Lessons learned: Coaches' perceptions of a pilot e-mentoring programme. *International Sport Coaching Journal*, 7(1), 22-30.

Hara, N., & Kling, R. (2001). Student distress in web-based distance education. Educause Quarterly, 24(3), 68-69.

Hair, J. F., Anderson, R. E., Babin, B. J., & Black, W. C. (2010). *Multivariate data analysis: A Global perspective* (Vol. 7). Pearson.

Hawi, N. (2010). Causal attributions of success and failure made by undergraduate students in an introductory-level computer programming course. *Computers & Education*, 54(4), 1127-1136.

Hsu, T. C., Chang, S. C., & Hung, Y. T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education*, *126*, 296-310.

Jacobs, K., Doyle, N., & Ryan, C. (2015). The Nature, perception, and impact of e-mentoring on post-professional occupational therapy doctoral students. *Occupational therapy in health care*, 29(2), 201-213.

Jang, Y., Choi, S., & Kim, H. (2022a). Development and validation of an instrument to measure undergraduate students' attitudes toward the ethics of artificial intelligence (AT-EAI) and analysis of its difference by gender and experience of AI education. *Education and Information Technologies*, 27(8), 11635-11667.

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022b). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies*, 27, 12855–1288. https://doi.org/10.1007/s10639-022-11120-6

Jegede, P. O. (2009). Predictors of java programming self-efficacy among engineering students in a Nigerian University. https://doi.org/10.48550/arXiv.0909.0074

Jenson, J., & Droumeva, M. (2016). Exploring media literacy and computational thinking: A Game maker curriculum study. *Electronic Journal of e-Learning*, *14*(2), 111-121.

Jocius, R., O'Byrne, W. I., Albert, J., Joshi, D., Robinson, R., & Andrews, A. (2021). Infusing computational thinking into STEM teaching. *Educational Technology & Society*, 24(4), 166-179.

Jung, H., Jang, Y., Kim, S., & Kim, H. (2022). KPCR: Knowledge graph enhanced personalized course recommendation. In *Proceedings of AI 2021: Advances in Artificial Intelligence: 34th Australasian Joint Conference, AI 2021* (pp. 739-750). Springer International Publishing.

Kahraman, M., & Abdullah, K. U. Z. U. (2016). E-mentoring for professional development of pre-service teachers: A Case study. *Turkish online journal of distance education*, *17*(3). https://doi.org/10.17718/tojde.19973

Kaufman, M. (2017). E-mentoring. National Mentoring Resource Center.

Kline, T. J. (2005). Psychological testing: A Practical approach to design and evaluation. Sage Publications.

Kwame Boateng, J. (2020). Using SAKAI LMS for enhanced instructor-student interactions and effective student learning: Perceptions of students from the University of Ghana. *Ubiquitous Learning: An International Journal, 13*(4). https://doi.org/10.18848/1835-9795/CGP/v13i04/1-10

Lau, W. W., & Yuen, A. H. (2011). Modelling programming performance: Beyond the influence of learner characteristics. *Computers & Education*, 57(1), 1202-1213.

Lee, H. M., & Mehta, M. R. (2015). E-Mentoring through a network of practice on Facebook. International Journal of Knowledge-Based Organizations (IJKBO), 5(1), 34-45.

Lee, J., Jung, Y., & Park, H. (2017). Gender differences in computational thinking, creativity, and academic interest on elementary SW education. *Journal of The Korean Association of Information Education*, 21(4), 381-391.

Lee, K. S., & Kim, B. G. (2016). Cross space: The Exploration of SNS-based writing activities in a multimodal learning environment. *Educational Technology & Society*, 19(2), 57-76.

Li, J., Liu, J., Yuan, R., & Shadiev, R. (2022). The Influence of socially shared regulation on computational thinking performance in cooperative learning. *Educational Technology & Society*, 25(1), 48-60.

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.

Lyon, J. A., & J. Magana, A. (2020). Computational thinking in higher education: A Review of the literature. *Computer Applications in Engineering Education*, 28(5), 1174-1189.

Murairwa, S. (2015). Voluntary sampling design. International Journal of Advanced Research in Management and Social Sciences, 4(2), 185-200.

National Research Council. (2010). Committee for the workshops on computational thinking: Report of a workshop on the scope and nature of computational thinking. National Academy Press. https://doi.org/10.17226/12840

Niresh, A., & Thirunavukkarasu, V. (2014). Firm size and profitability: A Study of listed manufacturing firms in Sri Lanka. *International journal of business and management*, 9(4). https://ssrn.com/abstract=2422441

Popescu, E. (2014). Providing collaborative learning support with social media in an integrated environment. *World Wide Web*, 17, 199-212.

Risquez, A. (2008). E-mentoring: An Extended practice, an emerging discipline. In F. J. Garcia- Penalvo (Ed.), Advances in e-learning: Experiences and methodologies (pp. 61-82). Information Science Publishing.

Rowland, K. N. (2012). E-mentoring: An innovative twist to traditional mentoring. *Journal of technology management & innovation*, 7(1), 228-237.

Royal Society. (2012). Shut down or restart: The Way forward for computing in UK schools. http://royalsociety.org/education/policy/computing-in-schools/report/

Rutten, M., Ros, A., Kuijpers, M., & Kreijns, K. (2016). Usefulness of social network sites for adolescents' development of online career skills. *Educational Technology & Society*, 19(4), 140-150.

Selby, C. C. (2014). How can the teaching of programming be used to enhance computational thinking skills? (Unpublished doctoral dissertation). University of Southampton, Southampton, UK.

Shpigelman, C. N., Weiss, P. L. T., & Reiter, S. (2009). E-mentoring for all. Computers in Human Behavior, 25(4), 919-928.

Single, P. B., & Muller, C. B. (2001). When emailing and mentoring unite: The Implementation of a nationwide electronic monitoring program. In L. K. Stromei (Ed.), *Creating mentoring and coaching programs* (pp. 107-122). American Society for Training and Development.

Single, P. B., & Single, R. M. (2005). Mentoring and the technology revolution: How face-to-face mentoring sets the stage for e-mentoring. In F. K. Kochan, & J. T Pascarelli (Eds.), *Creating successful telementoring programs* (pp. 7-27). Information Age Press.

Son, J., Kim, J. D., Na, H. S., & Baik, D. K. (2016). A Social learning management system supporting feedback for incorrect answers based on social network services. Journal of *Educational Technology & Society*, 19(2), 245-257.

Sánchez, R. A., Cortijo, V., & Javed, U. (2014). Students' perceptions of Facebook for academic purposes. *Computers & Education*, 70, 138-149.

Tanis, H., & Barker, I. (2017). E-mentoring at a distance: An Approach to support professional development in workplaces. *Turkish Online Journal of Distance Education*, 18(3), 135-155.

Thalheim, B. (2000). Fundamentals of entity-relationship modeling. Springer.

Tiit, E. M. (2021). Impact of voluntary sampling on estimates. Papers on Anthropology, 30(2), 9-13.

Tsutsui, K., & Takada, H. (2018). A Classroom SNS to promote reflective activity in programming learning for children. *Research and Practice in Technology Enhanced Learning*, 13(1), 1-18.

Weber, A. S. (2012). Considerations for social network site (SNS) use in education. *International Journal of Digital Information and Wireless Communications*, 2(4), 37-52.

Wing, J. (2006). Computational thinking. Communications of the ACM, 49(3), 33-36.

Wing, J. (2011). *Research notebook: Computational thinking—What and why? The Link Magazine, Spring*. Carnegie Mellon University, Pittsburgh. http://link.cs.cmu.edu/article.php?a=600

Wu, S. Y., & Su, Y. S. (2021). Visual programming environments and computational thinking performance of fifth-and sixth-grade students. *Journal of Educational Computing Research*, 59(6), 1075-1092.

Lin, X.-F., Wang, J., Chen, Y., Zhou, Y., Luo, G., Wang, Z., Liang, Z.-M., Hu, X., & Li, W. (2023). Effect of a Reflection-Guided Visualized Mindtool Strategy for Improving Students' Learning Performance and Behaviors in Computational Thinking Development. *Educational Technology* & *Society*, 26(2), 165-180. https://doi.org/10.30191/ETS.202304\_26(2).0012

# Effect of a Reflection-Guided Visualized Mindtool Strategy for Improving Students' Learning Performance and Behaviors in Computational Thinking Development

# Xiao-Fan Lin<sup>1,2,3\*</sup>, Jing Wang<sup>3,4</sup>, Yingshan Chen<sup>3</sup>, Yue Zhou<sup>3</sup>, Guoyu Luo<sup>3</sup>, Zhaoyang Wang<sup>3</sup>, Zhong-Mei Liang<sup>5</sup>, Xiaoyong Hu<sup>3,6</sup> and Wenyi Li<sup>2</sup>

 <sup>1</sup>Guangdong Provincial Philosophy and Social Sciences Key Laboratory of Artificial Intelligence and Smart Education, Guangdong Engineering Technology Research Center of Smart Learning, South China Normal University, Guangzhou, P.R. China // <sup>2</sup>Guangdong Provincial Institute of Elementary Education and Information Technology, Guangzhou, P.R. China // <sup>3</sup>School of Education Information Technology, South China Normal University, Guangzhou, P.R. China // <sup>4</sup>Teacher Education College of Guangdong-Hong Kong-Macao Greater Bay Area, South China Normal University, Guangzhou, P.R. China // <sup>5</sup>Zhixin South Road Primary School, Guangzhou, P.R. China // <sup>6</sup>Institute of Artificial Intelligence in Education, South China Normal University, Guangzhou, P.R. China // linxiaofan@m.scnu.edu.cn // jjw15683078248@163.com // 2021020874@m.scnu.edu.cn // 15107152939@163.com // 2475317832@qq.com // m17754831067@163.com // 351632147@qq.com // 472275060@qq.com // liwenyi@pku.edu.cn

\*Corresponding author

**ABSTRACT:** Computational thinking (CT) is an imperative competency in the 21st century. Mindtools can assist students in understanding concepts and decomposing tasks during CT development through programming. However, students may encounter challenges in complex CT problem-solving tasks due to being confused when using mindtools without proper guidance. Research evidence shows the potential of reflection in complex CT problem-solving by regulating cognitive activities. Accordingly, this study designed a reflection-guided visualized mindtool strategy to address CT development challenges. A quasi-experiment and lag sequential analysis were conducted by recruiting 97 junior high school students to examine the effects of the proposed strategy on CT development and to explore students' behavior patterns. Results revealed that the proposed approach improved students' CT performance, CT disposition, meta-cognitive awareness, and learning motivation. Students learning with the proposed strategy exhibited more key behaviors of facilitating CT problem-solving (e.g., generalizing the knowledge, re-designing the algorithm scheme, and evaluating the feasibility of their proposed schemes) than students in the control group, revealing the essential process of CT development and enlightening teachers on guiding students to produce such learning processes when cultivating CT.

Keywords: Reflection, Mindtool, Computational thinking, Behavior, Junior high school students

# 1. Introduction

Computational thinking (CT) is considered an imperative competency for everyone in the 21st century (Denning, 2017; Hsu et al., 2018). Existing studies show that learning CT benefits students' high-level thinking abilities, such as problem-solving, critical thinking, creativity, and collaboration (Denner et al., 2019; Scherer et al., 2019). CT has been incorporated into the national curricula of many countries, especially in programming courses, which are regarded as one of the most effective approaches to developing CT (Ezeamuzie & Leung, 2021; Shute et al., 2017; Wing, 2006). Nevertheless, there are challenges for novice CT learners to break down problems into small subproblems in the problem-solving process, generalize what they have learned in programming courses, and transfer it to solve problems in authentic and complex contexts (Zhao et al., 2022). Furthermore, it was found that students might continually encounter certain CT challenges related to programming concepts due to the abstraction and complexity of concepts such as event handling, conditionals, and manipulation of variables (Grover et al., 2016; Mouza et al., 2020). Thus, it is suggested that additional guidance or tools be provided for students to better grasp complex CT concepts (Lye & Koh, 2014).

Mindtools have been widely recognized in CT development through computer programming courses (Zhang et al., 2021). Concept maps are one of the mindtools that are believed to help students understand complex knowledge and to promote higher-order thinking, such as problem-solving (Jonassen & Carr, 2020). With the support of concept maps, students can easily divide an entire task into smaller subtasks in programming (Zhao et al., 2022). Although existing evidence has shown the role of concept maps in CT development (Krieglstein et al., 2022), students face several challenges when dealing with complex CT problem-solving tasks with concept maps (Wang et al., 2017). Specifically, self-generated concept maps may be incomplete and incorrect (e.g., unable to create meaningful connections) because of the lack of teachers' guidance (Eshuis et al., 2021; Wong et al., 2021).

Besides, students may also be confused about identifying gaps and misconceptions in their knowledge even when they have constructed a concept map (Eshuis et al., 2021). CT is not only a cognitive but also a meta-cognitive thinking process that regulates one's cognitive activities (Chen et al., 2021). Therefore, it is necessary to introduce appropriate guidance for students to help them review knowledge and regulate the meta-cognitive thinking process for CT development during problem-solving.

Reflection, which is a meta-cognitive strategy (Medina et al., 2017), has the potential to improve CT because students can be aware of their cognitive process and use this awareness to regulate their problem-solving (Colbert et al., 2015). According to Schön (1987), the mastery of a subject depends on a person's ability to reflect on the spot. Previous studies have indicated that reflection increased students' self-efficacy, learning motivation, meta-cognitive awareness, and CT performance by comparing new and old knowledge and integrating different opinions to understand errors and misconceptions (Chen et al., 2021; He et al., 2021; Lin et al., 2022a). Moreover, several scholars have also reported the effectiveness of reflection in figuring out connections between theory and practice (Radović et al., 2021), identifying misconceptions and deficiencies in knowledge (Cavilla, 2017), and developing a deeper understanding (Ghanizadeh, 2017). Therefore, the integration of mindtools and reflection may offer potential advantages for students to address complex learning problems (Chang & Hwang, 2022).

### 1.1. Research gaps

Although previous studies (e.g., Chang & Hwang, 2022) have noted the role of mindtools and reflection in complex problem-solving contexts, they did not regard the incorporation of mindtools and reflection as a teaching approach and measure its effect on CT education. In addition, there are conflicting findings regarding whether the use of mindtools and reflection enhances students' learning outcomes. For example, Chang and Hwang (2022) argued that reflection facilitated students' structural knowledge when they reflected on feedback generated by a mindtool, while Eshuis et al. (2021) indicated that integrating reflection prompts into a mindtool did not work as expected and could not help students improve their learning. Although there is a lack of relevant studies on adopting the combination approach of mindtool and reflection in CT development, this controversial situation may also be transferred to students' CT development. A general challenge is that CT is usually developed through programming in diverse operationalized ways, which may ignore cultivating the complex problem-solving ability of CT, potentially accounting for some debatable issues and conflicting findings (Ezeamuzie & Leung, 2021). Despite a plethora of CT studies in the general area of programming (e.g., Wu et al., 2019), robust empirical research investigating the effectiveness of a mental process to cultivate CT is more limited and often focused on the role of cognitive skills for CT development (Ezeamuzie & Leung, 2021). Rather mechanical understandings of CT development formerly prevailed, leading to students' frustration and interest reduction when they were faced with programming challenges (Sun et al., 2021). More recently, effective studies on CT development have conceptualized it as processes in which students have to actively participate in the cognitive and meta-cognitive process and be aware of regulating their problem-solving (Chen et al., 2021). In order to clarify what differentiates more effective from less effective reflective practices, there is a need to incorporate reflective approaches with mindtool-based CT development in relation to the thinking process (Lin et al., 2022a). To sum up, reflection could be considered an effective approach to promote students' active participation in the meta-cognitive process, and visualized mindtools could be regarded as a useful support strategy for facilitating cognitive development in CT learning. Therefore, it seems necessary to move beyond operationalized CT processes in various contexts of research and take into consideration a novel instructional strategy for problem-solving with a more simplified procedure (e.g., discover, extract, create, and assemble) (Ezeamuzie, 2022). In addition, few studies have used Lag sequential analysis (LSA) to focus on how students participate in the process of CT development and to what extent learning is facilitated from the perspective of behaviors (He et al., 2021).

#### **1.2. Research questions**

To address these gaps, the present study proposed a reflection-guided visualized mindtool strategy and assessed its effect on CT development. We experimented with the strategy and investigated its impact on students' performance and CT disposition. Measuring CT disposition is important for assessing whether students will work consistently in the CT development process despite being frustrated and failing due to challenges posed by complex problems (Jong et al., 2020). Besides, we were also interested in students' motivation as the result of the strategy, considering that reflection often plays a role in enhancing learners' CT learning motivation (Fang et al., 2022). It is also worth considering that students' CT can be developed by improving their metacognition (Chen et al., 2021). Measuring meta-cognitive awareness provides a good indicator that reflects changes in students' metacognition after incorporating reflection into the mindtool-based CT development process (Lin et al., 2022b). To understand how students' CT works through programming in a comprehensive and specific way, it is necessary for us to observe the details of students' behavior in the understanding process of the CT concept and problem-solving from a process-oriented perspective and at a micro level (Sun et al., 2021). LSA has been noted as an effective method to explore students' behavior patterns and learning performances with a contextualized reflective mechanism (Lin et al., 2022b). Accordingly, we conducted a quasi-experiment in a Fire Extinguishing AI Robot task of a programming course and examined students' CT development behaviors with LSA to explore their learning processes. We aimed to visualize the patterns and detect the sequential relationships between each behavior. This is one of the few studies that has used behavior analysis to explore the essential process of students' CT development behaviors in an experimental study with the reflection-guided visualized mindtool strategy. Specifically, the current study addressed the following research questions:

- RQ1. Are there any significant differences in students' CT performance (RQ1.1) and CT disposition (RQ1.2) in the Fire Extinguishing AI Robot task of a programming course which adopted the reflection-guided visualized mindtool strategy (RVMS) and the visualized mindtool strategy (VMS)?
- RQ2. Are there any significant differences in students' learning motivation (RQ2.1) and meta-cognitive awareness (RQ2.2) in the Fire Extinguishing AI Robot task of a programming course which adopted RVMS and VMS?
- RQ3. Are there any differences between the behavior patterns of those learning with RVMS and VMS?

# 2. Literature review

### 2.1. Computational thinking and programming

Computational thinking was first defined by Wing (2006) as a way of thinking to solve problems, design systems, and understand human behaviors with the use of fundamental computer science concepts. Although CT has been defined from different perspectives since Wing (2006), common points include that CT is the ability to use computer science concepts to solve problems, including computational concepts and computational practices in programming (Ezeamuzie & Leung, 2021). Buitrago Flórez et al. (2017) claim that students could better develop solutions for complicated problems in the real world with CT. Besides, several previous studies have shown that CT could benefit students' academic achievement (Lei et al., 2020), cognitive benefit (Scherer et al., 2019), problem-solving, and computer science attitudes (Denner et al., 2019).

Researchers have indicated that CT could be developed and facilitated through proper approaches (e.g., programming, robotics, and simulations) (Shute et al., 2017). It is generally believed that programming is a prominent and effective way to cultivate CT (Ezeamuzie & Leung, 2021; Lye & Koh, 2014), as it involves breaking down a problem into smaller problems and expressing a solution in the form of computational steps and algorithms (Merino-Armero et al., 2022). However, there is a lack of in-depth analysis of the way of thinking used in CT, which leads to students' difficulty in developing CT when engaged in programming learning (e.g., understanding of logics and semantics) (Shute et al., 2017). Additionally, previous research has pointed out the challenges in CT development practices, including difficulties understanding and applying complex CT concepts (Mouza et al., 2020), failure to decompose problems, and being unable to solve problems effectively by using CT and implementing solutions practically due to a lack of detailed mental models (e.g., mental maps, Venn diagrams) (Buitrago Flórez et al., 2017). To address the above difficulties of CT development, some studies have proved that scaffolding from appropriate visualized tools is essential for developing students' programming knowledge and thinking skills (Buitrago Flórez et al., 2017; Lye & Koh, 2014; Omer et al., 2020). Researchers have noted that mindtools play an important role in representing and organizing knowledge in computer programming courses (Zhang et al., 2021), which could help students understand complex knowledge and decompose tasks (Jonassen & Carr, 2020; Zhao et al., 2022).

### 2.2. Mindtool

Mindtools have been suggested as an effective method to engage students in organizing and presenting their knowledge through computer application programs (Jonassen et al., 1998; McAleese, 1998). Concept maps, as mindtools, have been widely applied in educational settings (Chang et al., 2022). With concept maps, individuals can generate meaningful learning by representing concepts with nodes, and the relationships between concepts with links (Yue et al., 2017). Moreover, students engage in a deep cognitive process while combining and representing conceptual knowledge structures (Novak & Gowin, 1984).

Chang's et al. (2022) study shed light on the potential of concept maps to improve computer science conceptual understanding. To help students understand complex knowledge, researchers have applied concept maps in complex problem-solving processes, such as breaking down the whole programming task, presenting logical thinking, and applying knowledge to practice, which might be difficult for students (Wang et al., 2017; Zhao et al., 2022). There is significant interest in two instructional methods of concept maps: self-generated concept maps and constructing on a scaffold (also known as the fill-in-the-blank construct) (Chang et al., 2001). Although self-generated concept maps have been applied less than fill-in-the-blank concept maps in studies, they offer a high degree of flexibility that can benefit students with different learning styles (Oliver, 2008). While fillin-the-blank concept maps are helpful for short-term learning, their restriction of freedom of content and structure may become a new constraint during long-term learning or in the case of complex knowledge (Wong et al., 2021). In a similar context for conceptual understanding and complex problem-solving, it is reasonable to apply the above-mentioned pedagogies (e.g., self-generated and fill-in-the-blank concept maps) to students' CT development. Although studies have confirmed that self-generated concept maps are more conducive to students' permanent development, constructing concept maps is challenging for students with low prior knowledge without extra support (Wong et al., 2021; Chuang et al., 2018). To improve students' programmable logic controllers knowledge performance, previous literature has noted the importance of promoting students' active participation in both cognitive and meta-cognitive learning processes for their CT development (Chen et al., 2021). Otherwise, students have difficulties eliminating the misunderstandings of the concepts by improving concept maps, which still existed even after completing concept maps (Eshuis et al., 2021).

### 2.3. Reflection

Reflection is regarded as an imperative activity in education research, including work by Dewey (1933). He defined reflection as continuously evaluating one's performances or behaviors to gain a deeper understanding of one's experiences (Dewey, 1933). Rodgers (2002) explained Dewey's concept of reflection as a four-stage process: presence to experience, description of experience, analysis of experience, and intelligent action, and noted that the action stage was often overlooked. Schön (1987) further studied reflective practice and divided the reflective practice into reflection-before-action, reflection-in-action, and reflection-on-action. Among these practices, Schön (1987) regarded reflection-in-action (occurring while performing the task) as the centre of art through which practitioners can cope with troublesome practices.

The effectiveness of reflection has been well recognized in the literature. Reflection is regarded as an essential meta-cognitive strategy for obtaining meaningful learning from specific experiences (Medina et al., 2017). Research has shown that reflection could facilitate complex problem-solving by critiquing the initial understanding of phenomena and constructing new descriptions (Schön, 2017). Besides, reflection is also helpful when cognitive knowledge is lacking. With the help of reflection, learners can observe and evaluate themselves, examine the gaps in their understanding, and their thinking and behavior can evolve (He et al., 2021; Lin et al., 2022a). There are many approaches to support and direct students' reflection, including prompting and guiding questions, think-aloud protocols, and peer assessment with assessment criteria (Radović et al., 2021). Studies have identified the benefits of these methods in students' learning performance (Fang et al., 2022; Radović et al., 2021). Meta-cognitive prompts were found to direct students' attention to important aspects of CT during their problem-solving processes and trigger their self-reflection (Chen et al., 2021). The adoption of think-aloud protocols has been suggested as students' cognitive processes can be verbalized and they can better understand CT practices (Lye & Koh, 2014). Fang et al. (2022) indicated that students could be encouraged to improve their work and tendency to solve complicated problems more systematically when receiving constructive comments from their peers.

Regarding CT, scholars have indicated that students are mainly situated in learning contexts that focus on the passive use of syntax and algorithms while lacking opportunities to engage in in-depth thinking about systematic approaches to problem-solving, which could lead to difficulty in facing real problems (Buitrago Flórez et al., 2017). Reflection can inspire students to have a stronger sense of motivation to engage in CT problem-solving and can develop their confidence (Chen et al., 2021). Besides, reflection has the potential to trigger students' metacognition in the CT development process (Buitrago Flórez et al., 2017). Previous studies suggested the likely effect of metacognition on directing students' attention to the critical aspects of CT and on helping students evaluate and question their solutions from different perspectives and discovering the limitations of their thinking during their problem-solving process (Chen et al., 2021; He et al., 2021), which could support the generalization of CT concepts and practices in more complex problem-solving contexts. Accordingly, there is a necessity to fill the research gap due to the lack of relevant studies on adopting the combination approach of mindtools for cognition and reflection for meta-cognitive thinking in CT development. In addition, there are conflicting findings concerning the effectiveness of integrating reflection prompts into mindtools (Chang &

Hwang, 2022; Eshuis et al., 2021). In order to clarify what more effective CT practices are, there is a need to incorporate a reflective approach with mindtools for CT development (Lin et al., 2022b). Therefore, we proposed a reflection-guided visual mindtool strategy to engage students in learning CT from a cognitive to meta-cognitive perspective.

# 3. Programming teaching design within a reflection-guided visualized mindtool strategy

# 3.1. System structure

In this study, we developed a reflection-guided visualized mindtool strategy system. The system structure includes a reflection-guided strategy, a mindtool promoting CT algorithm design mechanism, and several databases. The reflection-guided strategy enabled students to complete the reflection activity with the guidance of reflective prompts, reflective evaluation rubrics, and stimulated recall reflections. With the guidance of directions to create a better CT outcome. With the guidance of reflective evaluation rubrics, every student is responsible for the judgment results of peers' projects. With the guidance of stimulated recall reflections, students with an overall view of the cognitive knowledge structure and their cognitive process by providing self-generated and fill-in-the-blank concept maps. With the guidance of fill-in-the-blank concept maps, students can easily construct CT knowledge maps and better discover the connection between knowledges. With the guidance of self-generated concept maps, students can visualize their thinking process, helping them break down the task and clarify the step of the algorithm design. The mindtool promoting CT algorithm module also provides materials to connect the knowledge and thinking process for CT development with the open source software build module for completing CT tasks.

### 3.2. The context of the fire extinguishing AI robot task

In this self-developed online inquiry CT development environment, students are required to solve problems. They need to decompose the fire extinguishing AI robot task of a programming course (Figure 1) into several big steps, with one-by-one detailed solutions, in this process to develop their CT. The whole learning process is completed with the guidance of a reflection prompt. Reflective prompts may set up initial boundaries, trigger students to think more deeply, and make the most appropriate connection of concepts during the CT online inquiry activity. It can be inferred that when students are given reflective prompts such as specific content-based concept mapping prompts corresponding to CT development materials for the fire extinguishing AI robot, they can develop a more elaborated CT disposition than those without prompts. Also, reflective prompts provide an anchoring structure to which new information can be attached to existing schemas.

### Figure 1. Operating scenario of the fire extinguishing AI robot

The fire extinguishing robot should start from the starting point of the door, advance along the channel, enter the room, approach the fire source and extinguish it. Key dimensions in the room have been marked ( mm ).

The site represents a room in an ordinary residential house, where there is a fire point (burning candle) and the flame is 15 cm above the ground.



During the cruise, the robot cannot collide with or contact the wall. Fire is merciless. Fire extinguishing time should be controlled within 3 minutes, the shorter the better. Fire extinguishing methods are not limited. In addition, due to the limited budget of the fire brigade, your team must control the material cost of the fire fighting robot within 100,000 yuan.

Figure 2 shows the algorithm design. In this process, students used a mindtool to plan and design a structured and complex algorithm script to implement the critical actions that robots needed to include in the fire extinguishing process. Besides, it might also help students identify and correct individual mistakes they have made in the individual design, construct a more logical structure, and rearrange their thinking logically. Students used Arduino 1.0.3 to write code to realize the algorithm design program. Figure 3 shows the open source software. Students used Corona SDK to develop mobile phone software and generalize to solve problems in similar situations. Figure 4 shows the hardware. Students used this hardware to build robots and complete the key actions designed in the program.

# Figure 2. Program code design interface ink | Arduino 1.0.3 h 🛨 🔶 Blink Turns on an LED on for one second, then off for one second, repeatedly. This example code is in the public domain. // Pin 13 has an LED connected on most Arduino boards. aive a name: int led = 13; // the setup routine runs once when you press reset: void setup() { // initialize the digital pin as an output. pinMode(led, OUTPUT); routine runs over and over again forever: void loop()]{ digital delay(1000); digitalWrite(led, LOW); delay(1000); Arduino Mega (ATmega 1280) on /dev/tty.usbserial-A600enb: Using Arduino 1.0.3 to write code to implement the key program of the robot designed by the algorithm

The students completed the preliminary model building, found the model's shortcomings through the evaluation: Can the prototype of the fire extinguishing robot built at present meet the initial needs of requirements?; What is the reason?; What are the imperfections of the prototype?; What is the key gap?; and What is the improvement direction? With the guidance of reflective evaluation rubrics, every student was responsible for the judgment results of their peers' projects. After assigning scores for each dimension, students checked the rubrics to make sure that they had given an appropriate score to each dimension. After checking the rubrics, they reviewed the scores again to confirm whether their grading tasks were well done. Checking the rubrics was very important for students in the reflective reviewing task for CT tasks. Students spent a while checking the evaluating rules. They needed guidelines for measuring their peers' performance. With the rubrics, they could know whether their peers' projects met the expected criteria. Before choosing the scores for each dimension, they went through the evaluation rubrics and awarded the appropriate scores to their projects.

To promote CT development, the teacher videotaped a 1-hour lesson. During this period, the students were asked to revise their own projects based on reviewers' feedback. Then, the instructor used the first-round stimulated recall reflections to identify students' thought processes (i.e., behaviors) during the reviewing and revising phases (1 hour). With the replay of the video recordings, these students were allowed to see and explain their actions during the CT development within online inquiry tasks. Therefore, the categories, as well as behavior patterns were found. Finally, with the unfolding of some behavior patterns during CT development, the second-round stimulated recall reflections to involve students to describe their actions as they were watching video recordings of their online tasks. Thus, they were more likely to interpret or reflect on some specific behavior patterns in the consecutive stimulated recall reflections.

Figure 3. Open source software design interface



Using Corona SDK open source mobile phone software

Figure 4. Open source hardware build scenario



# 4. Method

### 4.1. Participants

We employed a quasi-experimental research design on the Fire Extinguishing AI Robot task of a programming course. The study participants (N = 97) were 42 female (43.3%) and 55 male (56.7%) junior high school students in southern China. The students' average age was 15.25 years, ranging between 13 and 16 years. These students were selected because they had all previously participated in a mindtool teaching project. We randomly divided the 97 students into experimental and control groups according to the experimental design, with 48 students assigned to the experimental group adopting RVMS, and 49 assigned to the control group adopting the VMS. All the students were taught by the same teacher who had taught information technology courses for nearly 10 years with enriched experience teaching programming.

### 4.2. Experiment process

Figure 5 shows the experimental procedure of the study, which was conducted for 10 weeks. Before the learning activities, all students were trained to familiarize themselves with the basic structure of the mindtool (e.g., concept mapping tools) needed in the later learning. Students then completed the 30-minute pre-tests and prequestionnaires, which aimed to exclude the effects of students' achievement and perception.

The learning materials were the same during the learning activities, including Arduino open-source hardware, the handbook of fire extinguishing AI robots, programming tools, and the system. Both groups of students used the Arduino device to accomplish the same task of fire extinguishing AI robots, and applied the programming knowledge they learned with the help of concept maps.

The difference between the two groups was that the students in the experimental group learned with the reflection-guided visualized mindtool strategy, while those in the control group adopted the visualized mindtool strategy. For instance, to identify cognitive gaps, students in the experimental group were asked to complete a concept map with the aid of several reflective activities. Through the reflective prompt activities, the concept maps were optimized several times to deepen understanding of programming knowledge to solve complex problems. Unlike the experimental group, control group students understood the CT concepts and decomposed problems without the assistance of reflection. That is, the teacher would directly tell them the correct answer when they encountered problems or had any doubts. Afterwards, the post-tests and post-questionnaires were administered for 30 minutes to examine whether their CT performance and disposition, learning motivation, and meta-cognitive awareness had improved.



### 4.3. Instrument

To collect data regarding CT cultivation, the instruments used for this study are as follows:

CT performance consisted of a pre- and post-test, to examine the effects of the proposed strategy. The tests consisted of two multiple-choice items, two open-ended questions, and one programming question (100 points). An example multiple-choice item is, "What are the intelligent functions of AI that help us complete our work?" An example open-ended question is, "Suppose you are going to design a smart fire alarm for your school. Please write down your ideas and the design proposals." The programming question is, "Please make an intelligent fire extinguishing robot and draw a flowchart of the program."

The pre-and-post questionnaires based on Tsai et al. (2021) were used to determine the level of CT disposition. This questionnaire could measure participants' perspectives on how they tried to think about and use skills related to CT. The original version of the CT disposition questionnaire was in English. Some words were modified to ensure that all items were clearly expressed in Chinese according to the students' opinions. Then, the adopted questionnaire was given to each participant of the two groups before and after teaching interventions. The questionnaire consisted of 25 items with a 5-point Likert scale for five sub-dimensions: abstraction,

algorithm, evaluation, decomposition, and generalization, scored from strongly disagree to strongly agree. The Cronbach's  $\alpha$  of the pre-test and post-test were 0.87 and 0.89, respectively. Indicative items for each sub-dimension are as following:

- Abstraction: I think I will try to think about how the program problems and the results are presented.
- Algorithm: I think I will try to develop detailed steps to solve the programming problem.
- Evaluation: I think I will try to find the right solution to the program problem.
- Decomposition: I think I will try to think about the possibility of a programming problem being decomposed.
- Generalization: I think I will try to determine if there are similarities between different programs.

The meta-cognitive awareness questionnaire was modified from the measuring tool proposed by Lin et al. (2022b), consisting of five items with a 5-point Likert-type rating scheme ranging from 1 (strongly disagree) to 5 (strongly agree). The Cronbach's alpha value of the questionnaire was 0.92. An example item is: I can discover the relationship between the critical issues in the Fire Extinguishing AI Robot task of a programming course that cause the program to fail.

Adapted from the instrument developed by Tapingkae et al. (2020), the learning motivation questionnaire consisted of eight items (e.g., When participating in online inquiry-based CT development in the Fire Extinguishing AI Robot task of a programming course, I always find that the learning is very interesting; Cronbach's alpha = 0.88). Students rated these items on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree).

# 4.4. Data collection and data analysis for the behavior analysis

### 4.4.1. The use of video recording and stimulated recall reflections

To observe students' learning behaviors during the online inquiry-based CT development activities, we recorded each student's behaviors throughout the process via video. We collected the video data in normal AI class situations. The teacher videotaped a 1-hour lesson. Students were asked to review their programming peers' initial projects and a 1-hour lesson during which they revised their projects based on reviewers' feedback. In addition, we used the video to identify students' thought processes (i.e., behaviors) during both the reviewing and revising phases (1 hour). With the replay of the video recordings, these students were allowed to see and explain their actions during the CT development within online inquiry tasks. Therefore, the categories as well as patterns of their behaviors were found for both groups.

### 4.4.2. Coding process

The video data were captured using FASTCAPTURE software installed on the students' computers. The students' behaviors were then video recorded through the CT development activity. Then, we replayed the video files for coding the reviewing and revising phases (180 minutes). LSA was conducted to explore the students' learning behavior patterns in the CT development activities by using the GSEQ software. To analyze the CT development behaviors, we developed an initial coding scheme by synthesizing the conceptual framework of CT from Tsai et al. (2021) and Ehsan et al. (2021). Then we carried out a pilot test by reviewing the recorded video to capture the primary behavior in the CT-oriented programming learning process and constructed the final coding table, as shown in Table 1.

		<i>Table 1.</i> The coding table of learning behaviors
Code	Phase/content	Description
A	Abstracting the gap	Thinking about a problem from a whole point of view to find the gaps rather
		than looking at the details
В	Examining the rubrics	Reading the online inquiry CT development rubrics.
С	Searching the Internet	Browsing the Web and reading the information on the Internet.
D	Decomposing	Breaking down a difficult CT problem into more manageable sub-problems
E	Algorithming	Writing codes with algorithms for a problem
F	Generalizing	Recognizing the specific type of practicing solutions and applying them to
		similar problems.
G	Evaluating	Finding the status of the best solution and resources
Н	Re-algorithming	Rewriting codes with better algorithms for a problem

Two researchers coded the video data based on the coding scheme. Both researchers have undergone comprehensive training on the operational definitions of the behavior codes. Considering that each student could averagely spend at least 20 seconds on the same behavior, the researchers underwent a real-time 20 second-to-20 second data coding. The lead coder was the first researcher (the principal investigator of this study) was the lead coder. During the coding process, the second researcher met regularly with the first researcher in this study to discuss coding disagreements and assess the inter-rater kappa criterion of 0.86.

### 5. Result

### 5.1. CT performance

After conducting the learning activity, we performed an analysis of covariance (ANCOVA) on the CT performance results to test the relationships between the two groups' post-test results. Before the ANCOVA, the Levene's test of determining homogeneity of variance was not violated (F = 0.39, p = .09 > .05), and the homogeneity of regression slopes was confirmed (F = 1.35, p = .15 > .05). Therefore, ANCOVA was conducted. Results in Table 2 showed that the CT performance of the experimental group students was significantly better than that of the control group, thus responding to RQ1.1. Moreover, the  $\eta^2$  value was 0.09, indicating that the finding had a medium effect size (Cohen, 1988).

Table 2. The ANCOVA result of CT performance

Group	Ν	Mean	SD	Adjusted Mean	Std. error	F	$\eta^2$
Experimental group	48	90.27	9.10	91.37	1.97	$4.87^{*}$	0.09
Control group	49	85.76	12.95	84.38	1.91		
$N_{oto} * n < 05$							

*Note.* p < .05.

### 5.2. CT disposition

Before conducting ANCOVA to analyze students' CT disposition, the Levene's test of homogeneity of variances was applied to examine whether variances across samples were equal. The result of this test was not significant (p = .12 > .05), suggesting that the difference between the variances for all groups was not significant. Also, the result (F = 2.36, p > .05) indicated that the homogeneity of regression coefficients was not violated. Therefore, ANCOVA was performed.

According to the results (F = 3.51, p < .05), the difference between the two groups was statistically significant (see Table 3). The CT disposition score of the experimental group was higher than that of the control group, which responded to RQ1.2. Furthermore, the  $\eta^2$  of the proposed method is 0.09, indicating a medium effect size.

Table 5. The ANCOVA result of CT disposition								
Group	Ν	Mean	SD	Adjusted Mean	Adjusted SD	F	$\eta^2$	
Experimental group	48	4.35	0.78	4.35	1.82	3.51*	0.09	
Control group	49	3.82	0.89	3.82	1.82			
Note $*n < 05$								

*Note.* p < .05.

### 5.3. Meta-cognitive awareness

Before the ANCOVA, the homogeneity of variance assumptions and homogeneity of regression coefficients were tested to examine the effect of the proposed strategy on students' meta-cognitive awareness, controlling for the pre-questionnaire scores. Levene's test for equality of variances was not significant (F = 5.87, p > .05). Hence, the homogeneity of variance assumption was not violated. Also, the result (F = 3.79, p > .05) indicated that the assumption of homogeneity of regression coefficients was not violated. Therefore, the ANCOVA was conducted.

The adjusted means and standard error were 4.21 and 0.89 for the experimental group, and 3.68 and 0.94 for the control group (see Table 4). The ANCOVA results indicated that the meta-cognitive awareness scores of the two groups showed a significant difference (F = 8.71, p < .05). As a response to RQ2.1, the meta-cognitive awareness score of the experimental group was statistically higher than that of the control group. Furthermore, the  $\eta^2$  of the proposed approach was 0.68, indicating a large effect size.

Table 4. The ANCOVA result of meta-cognitive awar	eness
---	-------

Group	N	Mean	SD	Adjusted Mean	Adjusted SD	F	$\eta^2$
Experimental group	48	4.21	0.67	4.21	0.89	$8.71^{*}$	0.68
Control group	49	3.68	0.89	3.68	0.94		

*Note.* \**p* < .05.

### 5.4. Learning motivation

In terms of the learning motivation scores, Levene's test for equality of variances was F = 3.56 (p = .12 > .05), indicating no significant difference between the two groups' learning motivation. The homogeneity of regression coefficients was examined to understand whether there was an interaction between the covariate and independent variables (F = 6.21, p > .05). It was observed that there was no interaction between the pre- and post-tests, indicating that the regression coefficient within the group did not reach a significant level. As the homogeneity assumption was satisfied, the ANCOVA could be performed.

The ANCOVA result is shown in Table 5. The results of the questions on learning motivation showed that the experimental group's learning motivation test score was significantly higher than that of the control group (F = 8.04, p < .05) with a large effect size ( $\eta^2 = 0.76$ ). The adjusted mean scores of the experimental and control group were M = 4.32 and M = 3.90, respectively. Based on the results, it is concluded that the learners who used RVMS had better learning motivation compared to those using VMS, responding to RQ2.2.

T 11 C	- TO1	NICONI	4 1.	C 1		
Tables	The L		$\Delta$ recult	of I	learning	motivation
Tuble J.	I IIC I		A ICSUIT	UI I	icarining.	mouvation

Groups	Ν	Mean	SD	Adjusted Mean	Adjusted SD	F	$\eta^2$	
Experimental group	48	4.32	0.81	4.32	0.98	$8.04^{*}$	0.76	
Control group	49	3.90	0.88	3.90	0.64			
$N_{ref} \approx 0.5$								

```
Note. p < .05.
```

#### 5.5. Comparisons of behaviors of the two groups of students

In response to RQ3, this study examined the students' behaviors patterns of the experimental and control groups with LSA in their CT learning. In Figure 6, it can be seen that the two groups were similar in six sets of significant sequences, including  $B \rightarrow B$ ,  $C \rightarrow C$ ,  $E \rightarrow E$ ,  $A \rightarrow D$ ,  $B \rightarrow D$ , and  $G \rightarrow A$ . In other words, all students, no matter which group they belonged to, demonstrated continuity in abstracting the gap ( $B \rightarrow B$ ), searching for the information on the Internet ( $C \rightarrow C$ ), and algorithming ( $E \rightarrow E$ ). Besides, they all showed three unidirectional sequences from Abstracting the gap to Decomposing ( $A \rightarrow D$ ), Examining the rubrics to Decomposing ( $B \rightarrow D$ ), and Evaluating to Abstracting the gap ( $G \rightarrow A$ ).



*Note.* A: Abstracting the gap. B: Examining the rubrics. C: Searching the Internet. D: Decomposing. E: Algorithming. F: Generalizing. G: Evaluating. H: Re-algorithming.

However, in the experimental group, more sets of significant sequences were involved in students' behavior patterns, revealing that the experimental group demonstrated special sequences, including  $C \rightarrow D$ ,  $F \rightarrow G$ ,  $F \rightarrow H$ ,  $H \rightarrow G$ ,  $G \rightarrow D$ , and  $A \rightarrow A$ . During the CT development activities, experimental group students attempted to decompose after searching the Internet ( $C \rightarrow D$ ), while those in the control group did not. Besides, the behavior of abstracting the gap (A) of the experimental group students was repeated continuously, which meant that they tended to think deeply and fully abstract their gaps. Furthermore, after generalizing, students turned to realgorithm then went to evaluation ( $F \rightarrow H$ ,  $H \rightarrow G$ ) or turned to evaluate directly ( $F \rightarrow G$ ). They then tried to decompose ( $G \rightarrow D$ ); that is, the special sets of sequences present in the experimental group are the key to the difference in performance of the two groups.

## 6. Discussion

This study proposed a reflection-guided visualized mindtool strategy to improve students' CT performance, CT disposition, learning motivation, and meta-cognitive awareness. The study further explored the effect of reflection integrated into mindtools on students' behaviors during CT development.

Concerning CT performance and CT disposition, the experimental group had significantly better learning achievements. The improvement in CT performance and disposition implied that the proposed strategy could promote the development of students' CT by effectively incorporating reflection into the use of mindtools during the process of programming for CT. With mindtools, students could regulate CT concepts and make connections between each concept easily and visually. Additionally, the findings suggested the likely effects of reflection on continually directing students to review the process of CT problem-solving and improving their solutions to successfully solve the complex problem of CT. This finding, however, differed from the conclusion of past research (Eshuis et al., 2021), which reported that the combination of reflective prompts and concept maps could not promote students' learning achievement. It might be that students could not actively reflect on the provided information, and their reflection was too superficial only with reflective prompts. The present study adopted various reflective approaches rather than just reflective prompts, including reflective evaluation rubrics and stimulated recall reflections. With the guidance of reflective prompts, students could think more deeply and make the appropriate connection of concepts during the CT online inquiry activity. With the evaluation rubrics, students could measure their peers' performance and know whether their peers' and their own projects met the expected criteria by scoring the CT tasks. In the consecutive stimulated recall reflections, students were allowed to interpret or reflect on some specific behavior patterns during the CT development within online inquiry tasks to improve their performance in the next stage.

In terms of behavioral transition diagrams, both the experimental and control group learning with the mindtool strategy generated the behavior of decomposing after abstracting the gap  $(A \rightarrow D)$  or examining the rubrics  $(B \rightarrow D)$ . This means that CT development with the mindtool strategy allowed students to plan and examine a precise and sophisticated structure of concepts and further decompose all possible main or specific scenes nodeby-node to construct frames to complete the online inquiry-based CT tasks. These findings are consistent with the results of Zhao et al. (2022), which also showed the positive effects of mindtools on the cultivation of CT. However, students using RVMS generated more specific behavior sequences than those learning with VMS. In terms of the repeated sequence  $(A \rightarrow A)$ , it was found that, with the guidance of reflective evaluation rubrics, students in the experimental group could clearly understand the goal of the task and then accurately discover the limitations of their ideas. Moreover, students were more likely to generalize the knowledge they had learned, redesign their algorithms for the tasks, evaluate the feasibility of their proposed schemes, and then better decompose the task  $(F \rightarrow H \rightarrow G \rightarrow D)$ ,  $F \rightarrow G \rightarrow D$ ). These behaviors implied that students could transfer the knowledge to find better solutions after reflecting on their own CT problem-solving process.

The findings based on the different behavior patterns of the two groups in this study indicated that the proposed strategy, which incorporates reflection into the whole process of CT problem-solving by carrying out various reflective activities (i.e., reflective prompts, reflective evaluation rubrics, and stimulated recall reflections), could engage students in deeper reflection and allow them to be more deeply involved in the CT problem-solving process. This may be why students learning with RVMS showed better CT learning outcomes, meta-cognitive awareness, and motivation than those learning with VMS. The study further indicated the value of reflective behaviors in CT cognitive and meta-cognitive processes, including problem and action gap identification, information searching and re-algorithming, and generalizing new ideas and alternative generation. The findings revealed that successful CT development usually depended on individuals' reflection to engage proactively in CT development challenges and persist in attempts to meet the challenges they encountered. Hence, the valuable determinants regarding the reflection-guided mindtool for conducting online inquiry-based CT development in a

programming or AI course are advocated. For better CT outcomes, the cultivation of CT should pay more attention to engaging students in certain behaviors, including problem identification, information collection, idea generation, hypotheses making, generalizing, re-algorithming, and evaluation. This study provides practical recommendations including an effective reflection-guided visualized mindtool strategy for improving students' CT learning performance and behaviors from cognition to meta-cognition. This study extends the existing pedagogy in CT learning by highlighting the importance of visualized mindtools to promote students' conceptual knowledge, self-generated thinking, and complex problem-solving for cognitive development in CT learning, and the role of the reflection-guided strategy to actively engage students in the meta-cognitive process with the guidance of reflective prompts, reflective evaluations, and stimulated recall reflections.

# 7. Conclusion

The present study has provided a thorough look at and understanding of a reflection-guided visualized mindtool strategy aiming to advance students' CT outcomes (i.e., CT performance, CT disposition, learning motivation, and meta-cognitive awareness). Meanwhile, the proposed strategy has great potential to activate students' indepth reflection (i.e., monitoring and regulating their cognitive activities and practices) during online CT development by incorporating reflection with mindtools into the whole process of CT problem-solving with various reflective methods. The results can guide teachers in optimizing their CT pedagogy. For example, teachers can use visualized mindtools, such as concept maps, to guide students' thinking and task decomposition. Besides, teachers can provide proper approaches or tools to arouse students' reflection before, during, and after CT problem-solving to identify gaps in their thinking and behaviors and improve their CT-related problemsolving abilities. In addition, it is also suggested that teachers pay more attention to cultivating students' certain behaviors before implementing CT development activities based on the behavior analysis results. It was observed that the experimental group generated more specific behaviors, including abstracting the gaps, decomposing tasks, generalizing, re-algorithming, and evaluating, which are critical to learning CT (Grover et al., 2016). Thus, instead of solely teaching syntax and how to code during programming courses, teachers should focus more on guiding students to find ways to solve CT problems. The findings of this study have the potential to promote these valuable behaviors.

Although the experimental results showed that RVMS is effective for CT cultivation, some limitations to this study should be noted. First, this study just recorded students' behaviors during the online inquiry-based CT development activity to observe behavioral changes, which may have ignored some potential behavior critical to CT in face-to-face learning or after class. Second, the present study only identified the effectiveness of the proposed strategy in terms of CT concepts and CT practices; other aspects of CT were not considered. Further studies are needed to investigate the effect of the proposed strategy on other aspects of CT, like CT skills. Lastly, the coding process of behavior analysis was manual, which takes time and may inevitably result in human error. The development of an automated recording and coding system may be considered.

# Acknowledgement

This work was supported by the National Natural Science Foundation of China [grant number 62007010]; the Key Project of National Natural Science Foundation of China under grant number 62237001; the National Key R&D Program of China [grant number 2022YFC3303605]; the Science and Technology Projects in Guangzhou [grant number 202102021217]; the Special Funds of Climbing Program regarding the Cultivation of Guangdong College Students' Scientific and Technological Innovation [grant number pdjh2023a0139]; 2023 annual Guangzhou Youth and the Communist Youth League project of "Guangzhou Youths' Participation in Rural Revitalization Research: The I-SEED 'Internet Plus' Cloud Public Welfare to Empower Rural Education Revitalization" [grant number 2023TSW13], Teaching Quality Project of South China Normal University: Professional Development of Artificial Intelligence Teachers under "New Normal" Background [grant number 192]; College Student Innovation and Entrepreneurship Training Program [grant number 202328010]; and South China Normal University "Challenge Cup" Golden Seed Cultivation Project [grant number 2023XKA09].

### References

Buitrago Flórez, F., Casallas, R., Hernández, M., Reyes, A., Restrepo, S., & Danies, G. (2017). Changing a generation's way of thinking: Teaching computational thinking through programming. *Review of Educational Research*, 87(4), 834–860.

Cavilla, D. (2017). The Effects of student reflection on academic performance and motivation. SAGE Open, 7(3). https://doi.org/10.1177/2158244017733790

Chang, C.-C., & Hwang, G.-J. (2022). A Structured reflection-based graphic organizer approach for professional training: A Technology-supported AQSR approach. *Computers & Education*, 183, 104502. https://doi.org/10.1016/j.compedu.2022.104502

Chang, C.-C., Hwang, G.-J., & Tu, Y.-F. (2022). Concept mapping in technology-supported K-12 education: A Systematic review of selected SSCI publications from 2001 to 2020. *Journal of Educational Computing Research*, *60*(7), 1637-1662.

Chang, K. E., Sung, Y. T., & Chen, S. F. (2001). Learning through computer-based concept mapping with scaffolding aid: Learning through computer-based concept mapping. *Journal of Computer Assisted Learning*, *17*(1), 21–33.

Chen, C.-H., Liu, T.-K., & Huang, K. (2021). Scaffolding vocational high school students' computational thinking with cognitive and meta-cognitive prompts in learning about programmable logic controllers. *Journal of Research on Technology in Education*, 1–18. https://doi.org/10.1080/15391523.2021.1983894

Chuang, C.-W., Hwang, G.-J., & Tsai, W.-J. (2018). A Peer tutoring-based concept mapping approach to improving students' learning achievements and attitudes for a social studies course. *International Journal of Online Pedagogy and Course Design*, 8(1), 1–12. https://doi.org/10.4018/IJOPCD.2018010101

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Colbert, C. Y., Graham, L., West, C., White, B. A., Arroliga, A. C., Myers, J. D., Ogden, P. E., Archer, J., Mohammad, Z. T. A., & Clark, J. (2015). Teaching meta-cognitive skills: Helping your physician trainees in the quest to 'Know What They Don't Know'. *The American Journal of Medicine*, *128*(3), 318–324.

Denner, J., Campe, S., & Werner, L. (2019). Does computer game design and programming benefit children? A Metasynthesis of research. ACM Transactions on Computing Education, 19(3), 1–35. https://doi.org/10.1145/3277565

Denning, P. J. (2017). Remaining trouble spots with computational thinking. Communications of the ACM, 60(6), 33–39.

Dewey, J. (1933). *How we think: A Restatement of the relation of reflective thinking to the educative process*. D.C. Heath and Company.

Ehsan, H., Rehmat, A. P., & Cardella, M. E. (2021). Computational thinking embedded in engineering design: Capturing computational thinking of children in an informal engineering design activity. *International Journal of Technology and Design Education*, 31(3), 441–464.

Eshuis, E. H., Vrugte, J., Anjewierden, A., & Jong, T. (2021). Expert examples and prompted reflection in learning with concept maps. *Journal of Computer Assisted Learning*, 38(2), 350–365.

Ezeamuzie, N. O. (2022). Abstractive-based programming approach to computational thinking: Discover, extract, create, and assemble. *Journal of Educational Computing Research, https://doi.org/10.1177/07356331221134423* 

Ezeamuzie, N. O., & Leung, J. S. C. (2021). Computational thinking through an empirical lens: A Systematic review of literature. *Journal of Educational Computing Research*, 60(2), 481-511.

Fang, J.-W., Shao, D., Hwang, G.-J., & Chang, S.-C. (2022). From critique to computational thinking: A Peer-assessment-supported problem identification, flow definition, coding, and testing approach for computer programming instruction. *Journal of Educational Computing Research*, 60(5), 1301–1324.

Ghanizadeh, A. (2017). The Interplay between reflective thinking, critical thinking, self-monitoring, and academic achievement in higher education. *Higher Education*, 74, 101–114. https://doi.org/10.1007/s10734-016-0031-y

Grover, S., Pea, R., & Cooper, S. (2016). Factors influencing computer science learning in middle school. In *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (pp. 552-557). Association for Computing Machinery. https://doi.org/10.1145/2839509.2844564

He, Z., Wu, X., Wang, Q., & Huang, C. (2021). Developing eighth-grade students' computational thinking with critical reflection. *Sustainability*, *13*(20), 11192. https://doi.org/10.3390/su132011192

Hsu, T.-C., Chang, S.-C., & Hung, Y.-T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education*, *126*, 296–310. https://doi.org/10.1016/j.compedu.2018.07.004

Jonassen, D. H., & Carr, C. S. (2020). Mindtools: Affording multiple knowledge representations for learning. In *Computers as cognitive tools, volume two: No more walls* (pp. 165-196). Routledge. https://doi.org/10.1201/9781315045337-8

Jonassen, D. H., Carr, C., & Yueh, H.-P. (1998). Computers as mindtools for engaging learners in critical thinking. *TechTrends*, 43(2), 24–32.

Jong, M. S. Y., Geng, J., Chai, C. S., & Lin, P. Y. (2020). Development and predictive validity of the computational thinking disposition questionnaire. *Sustainability*, *12*(11), 4459. https://doi.org/10.3390/su12114459

Krieglstein, F., Schneider, S., Beege, M., & Rey, G. D. (2022). How the design and complexity of concept maps influence cognitive learning processes. *Educational Technology Research and Development*, 70(1), 99–118.

Lei, H., Chiu, M. M., Li, F., Wang, X., & Geng, Y.-J. (2020). Computational thinking and academic achievement: A Metaanalysis among students. *Children and Youth Services Review*, *118*, 105439. https://doi.org/10.1016/j.childyouth.2020.105439

Lin, Y.-T., Yeh, M. K.-C., & Tan, S.-R. (2022a). Teaching programming by revealing thinking process: Watching experts' live coding videos with reflection annotations. *IEEE Transactions on Education*, 65(4), 617–627.

Lin, X.-F., Hwang, G.-J., Wang, J., Zhou, Y., Li, W., Liu, J., & Liang, Z.-M. (2022b). Effects of a contextualised reflective mechanism-based augmented reality learning model on students' scientific inquiry learning performances, behavioural patterns, and higher order thinking. *Interactive Learning Environments*, 1-21. https://doi.org/10.1080/10494820.2022.2057546

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. https://doi.org/10.1016/j.chb.2014.09.012

McAleese, R. (1998). The Knowledge arena as an extension to the concept map: Reflection in action. *Interactive Learning Environments*, 6(3), 251–272.

Medina, M. S., Castleberry, A. N., & Persky, A. M. (2017). Strategies for improving learner metacognition in health professional education. *American Journal of Pharmaceutical Education*, 81(4), 78. https://doi.org/10.5688/ajpe81478

Merino-Armero, J. M., González-Calero, J. A., & Cózar-Gutiérrez, R. (2022). Computational thinking in K-12 education. An Insight through meta-analysis. *Journal of Research on Technology in Education*, *54*(3), 410-437.

Mouza, C., Pan, Y.-C., Yang, H., & Pollock, L. (2020). A Multiyear investigation of student computational thinking concepts, practices, and perspectives in an after-school computing program. *Journal of Educational Computing Research*, 58(5), 1029–1056.

Novak, J. D., & Gowin, D. B. (1984). Learning how to learn. Cambridge University Press.

Oliver, K. (2008). A Comparison of Web-based concept mapping tasks for alternative assessment in distance teacher education. *Journal of Computing in Teacher Education*, 24(3), 95-103.

Omer, U., Farooq, M. S., & Abid, A. (2020). Cognitive learning analytics using assessment data and concept map: A Framework-based approach for sustainability of programming courses. *Sustainability*, *12*(17), 6990. https://doi.org/10.3390/su12176990

Radović, S., Firssova, O., Hummel, H. G., & Vermeulen, M. (2021). Improving academic performance: Strengthening the relation between theory and practice through prompted reflection. *Active Learning in Higher Education*. https://doi.org/10.1177/14697874211014411

Rodgers, C. (2002). Defining reflection: Another look at John Dewey and reflective thinking. *Teachers College Record*, 104(4), 842–866.

Scherer, R., Siddiq, F., & Sánchez Viveros, B. (2019). The Cognitive benefits of learning computer programming: A Metaanalysis of transfer effects. *Journal of Educational Psychology*, 111(5), 764–792.

Schön, D. A. (1987). *Educating the reflective practitioner: Toward a new design for teaching and learning in the professions*. Wiley.

Schön, D. A. (2017). The Reflective practitioner. Routledge.

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142–158. https://doi.org/10.1016/j.edurev.2017.09.003

Sun, D., Ouyang, F., Li, Y., & Chen, H. (2021). Three contrasting pairs' collaborative programming processes in China's secondary education. *Journal of Educational Computing Research*, 59(4), 740–762.

Tapingkae, P., Panjaburee, P., Hwang, G.-J., & Srisawasdi, N. (2020). Effects of a formative assessment-based contextual gaming approach on students' digital citizenship behaviours, learning motivations, and perceptions. *Computers & Education*, *159*, 103998. https://doi.org/10.1016/j.compedu.2020.103998

Tsai, M.-J., Liang, J.-C., & Hsu, C.-Y. (2021). The Computational thinking scale for computer literacy education. *Journal of Educational Computing Research*, 59(4), 579–602.

Wang, M., Cheng, B., Chen, J., Mercer, N., & Kirschner, P. A. (2017). The Use of web-based collaborative concept mapping to support group learning and interaction in an online environment. *The Internet and Higher Education*, *34*, 28–40. https://doi.org/10.1016/j.iheduc.2017.04.003

Wing, J. M. (2006). Computational thinking. Communications of the ACM, 49(3), 33-35.

Wong, R. M., Sundararajan, N., Adesope, O. O., & Nishida, K. R. A. (2021). Static and interactive concept maps for chemistry learning. *Educational Psychology*, *41*(2), 206–223.

Wu, B., Hu, Y., Ruis, A. R., & Wang, M. (2019). Analysing computational thinking in collaborative programming: A Quantitative ethnography approach. *Journal of Computer Assisted Learning*, *35*(3), 421–434.

Yue, M., Zhang, M., Zhang, C., & Jin, C. (2017). The Effectiveness of concept mapping on development of critical thinking in nursing education: A Systematic review and meta-analysis. *Nurse Education Today*, *52*, 87–94.

Zhang, J.-H., Meng, B., Zou, L.-C., Zhu, Y., & Hwang, G.-J. (2021). Progressive flowchart development scaffolding to improve university students' computational thinking and programming self-efficacy. *Interactive Learning Environments*, 1–18. https://doi.org/10.1080/10494820.2021.1943687

Zhao, L., Liu, X., Wang, C., & Su, Y.-S. (2022). Effect of different mind mapping approaches on primary school students' computational thinking skills during visual programming learning. *Computers & Education*, 181, 104445. https://doi.org/10.1016/j.compedu.2022.104445

# A Framework for Applying Sequential Data Analytics to Design Personalized Digital Game-Based Learning for Computing Education

# Zhichun Liu<sup>1\*</sup> and Jewoong Moon<sup>2</sup>

<sup>1</sup>Human Communication, Development, and Information Sciences, The University of Hong Kong, Hong Kong SAR, China // <sup>2</sup>Department of Department of Educational Leadership, Policy, & Technology Studies, University of Alabama, Tuscaloosa, AL, USA // liulukas91@gmail.com // jmoon19@ua.edu

\*Corresponding author

**ABSTRACT:** In this study, we have proposed and implemented a sequential data analytics (SDA)-driven methodological framework to design adaptivity for digital game-based learning (DGBL). The goal of this framework is to facilitate children's personalized learning experiences for K–5 computing education. Although DGBL experiences can be beneficial, young children need personalized learning support because they are likely to experience cognitive challenges in computational thinking (CT) development and learning transfer. We implemented the educational game Penguin Go to test our methodological framework to detect children's optimal learning interaction patterns. Specifically, using SDA, we identified children's diverse gameplay patterns and inferred their learning states related to CT. To better understand children's gameplay performance and CT development in context, we used qualitative data as triangulation. We discuss adaptivity design based on the children's gameplay challenges indicated by their gameplay sequence patterns. This study shows that SDA can inform what in-game support is necessary to foster student learning and when to deliver such support in gameplay. The study findings suggest design guidelines regarding the integration of the proposed SDA framework.

Keywords: Digital game-based learning, Computational thinking, Sequential data analytics, Adaptivity, Personalized learning

# **1. Introduction**

A major goal of recent computing education is to enhance children's computational thinking (CT). CT is a way of thinking that involves representing solutions via computational practices (Grover & Pea, 2013). Research has shown a concern that young children are likely to face cognitive challenges in developing CT due to its complexity (Lye & Koh, 2014). CT-related learning tasks are likely to overwhelm children and then undermine motivation and learning engagement (Zhao & Shute, 2019). It hence necessitates engaging and effective ways to support CT development for young children. Correspondingly, recent research has called for digital game-based learning (DGBL) as a means that promotes children's problem-solving and hands-on experiences — resulting in the development of concrete cognitive footings for abstract knowledge (Zhao & Shute, 2019). Previous works have demonstrated purposeful DGBL design that facilitates children's CT skills development through playful learning (e.g., Asbell-Clarke et al., 2020; Bers, 2020; Israel-Fishelson & Hershkovitz, 2020). Children as players, are guided to explore a variety of game missions where CT skills are necessary. Through playing, students are expected to initiate hypotheses and then come up with creative solutions derived from appropriate CT skills and concepts through multiple rounds of game trials. Despite the emergence of DGBL in computing education, skepticism exists on whether and how DGBL supports students with different knowledge levels and backgrounds.

Despite increasing DGBL research on computing education, there is a lack fo studies that discussed how DGBL supports children's personalized learning experiences (Hooshyar et al., 2021). Whereas DGBL enhances engagement and motivation, research reports that young children may undergo cognitive distractions and ingame frustration easily (Lye & Koh, 2014; Bers, 2020). To guide children's attention and mindful gameplay in DGBL, it is essential to help them keep engaged and focused in gameplay through personalized support. However, there is little systematic guide for designing the content of the support, the timing of support delivery, and the format of the support (Liu et al., 2020). Since DGBL with evidence-centered design (ECD) supposes observable game actions that represent children's learning states, it is essential to seek ways to grasp and analyze the nature of children's in-game behaviors aligned with CT. Whereas researchers used various data analytics to investigate learners' in-game behaviors in DGBL, existing data-driven approaches are limited in identifying children's needs under the gameplay nature (Moon & Liu, 2019). In this study, we propose, implement, and test sequential data analytics (SDA)-driven methodological framework to investigate young children's (K-5) gameplay patterns in the educational game Penguin Go. Furthermore, we discuss how this SDA-driven approach helps to conduct data-driven decisions for developing adaptive DGBL for young children.

# 2. Literature review

### 2.1. Computing education and GBL

The field of computing education highlights CT, which is an analytical ability to decompose complicated problems, identify their patterns, and execute tailored solutions by computational means (Lye & Koh, 2014). Shute et al. (2017) identified the main competencies of CT as follows: (1) decomposition; (2) algorithm thinking; (3) abstraction; (4) debugging; (5) iteration; (6) generalization. However, due to children's inexperience entering computing education, they tend to undergo cognitive challenges that may result in low engagement and high frustration. Therefore, it is essential to provide children with motivating environments to boost their learning engagement.

A current CT movement has focused on enabling all learners to engage in computing education (Weintrop et al., 2016). There are two pivotal design rationales of DGBL in computing education. First, a major assumption of DGBL in computing education is implicit learning (Rowe et al., 2021) from everyday play behavior that does not explicitly appear. A game is a good platform that allows learners to demonstrate a particular pattern of behavior through play. Individuals' gaming actions and their consequences in game tasks are linked with the implicit CT learning. In this sense, many researchers sought to create a game mechanic that purposefully fosters learners' CT-related behaviors from play. Second, another lens of DGBL for computing education is constructionism. Weintrop et al. (2016) stated three design principles of an educational game: (1) personally meaningful artifact design, (2) exploration and discovery in play, and (3) engaging with powerful ideas to be advanced. They underscored that a game needs to present challenges that allow learners to initiate and test their conceptions from simple to complex. While building a pile of codes with iterations, learners can build and elaborate design rationales and internalize their programming logics through a series of game tasks. Game challenges and failure experiences help them to detect misconceptions, analyze consequences, and debug execution codes during multiple rounds of play. In this sense, DGBL has been useful to introduce computing education to young children. The key to incorporating DGBL into computing education is to make computer programming practices more engaging to young learners (Hsu et al., 2018). Previous research indicated that DGBL benefits learners' CT development by enhancing their engagement via gameplay (Israel-Fishelson & Hershkovitz, 2020; Turchi et al., 2019; Zhao & Shute, 2019). Moreover, during gameplay, learners can build and test their problem-solving solutions (Grover et al., 2017). Such problem-solving processes during gameplay seamlessly facilitate learners' iterations of hypothesis testing and solution executions, which in turn contribute to their development of CT skills. Asbell-Clarke et al. (2020), for example, created and implemented Zoombinis, a 2D learning game teaching CT to young children. Using data-driven automatic detectors of student gameplay (i.e., classification algorithms), they reported that children who demonstrated evidence of active problem solving in the game (e.g., change one variable while holding others constant) were more proficient in CT skills compared to those who were still learning the game mechanics (e.g., repeatedly using the same but ineffective solution in one puzzle).

### 2.2. Challenges in children's gameplay and adaptive game design

Although DGBL engages young learners in computing education, research has suggested that young children are likely to face cognitive challenges in CT-related problem-solving in gameplay. Young children tend to demonstrate inefficient solution implementations and unsystematic debugging (e.g., trial-and-error) caused by random, non-strategic, or sometimes unproductively wheel-spinning. Such inefficient solutions often involve step-by-step execution, testing with random combinations, or debugging without meaningful subgoals (Fessakis et al., 2013; Liu et al., 2017). Although iterative trial-and-error may help to solve game problems, such patterns do not always lead to meaningful learning (Owen et al., 2019). Multiple trials and errors without further improvement rather give rise to frustration and disengagement. This behavior pattern is largely attributed to children's limited cognitive and meta-cognitive resources. In a highly interactive environment such as DGBL, children are exposed to high cognitive load (Azevedo & Aleven, 2013; Morrison et al., 2015), which poses challenges for higher-order CT skills—such as loop and conditional statement development (Ching et al., 2018).

In addition, research has reported learning transfer as a significant issue after the gameplay: Children seem to enjoy and excel within the game, but they did not perform well on the knowledge test outside of the game (Arena & Schwartz, 2014; Mason et al., 2011). When children are asked to perform the learned skills in a different context (often referred as far transfer), they need to first understand the similarity between the original learning context and then apply the learned cognitive processes into a new context (Taatgen, 2013). Both steps require a large amount of cognitive and meta-cognitive resources; hence, it is less likely that they can perform well on transfer tasks after simply playing games (Liu & Jeong, 2022). In a highly interactive environment such as

games, children should pay mindful attention requiring cognitive and meta-cognitive resources under diverse gaming trajectories (Ke & Abras, 2013). Therefore, it is essential for DGBL researchers to identify the cognitive or meta-cognitive needs and design personalized support to help children to acquire transferrable skills through games.

Children's cognitive challenges augment the importance of personalized learning experiences. Personalized learning is a learning design that adjusts either learning modules and instructional strategies tailored to children's learning states or interests (Walkington, 2013). To perform personalized learning, identifying children's learning trajectories and dynamic problem-solving processes in advance is crucial (Lin et al., 2013). In DGBL, to systematically support children's personalized learning, emerging research has incorporated adaptivity in games (Vanbecelaere et al., 2020). Here, adaptivity refers to the systematic and dynamic delivery of game-based instructional activities through ongoing and in-situ learner analyses (Liu et al., 2020). Furthermore, to determine either level or format of adaptive learning support best suited to individuals, DGBL systems need to collect and analyze learner profiles and present appropriate support to them. A recent study by Hooshyar et al. (2021) showed how to provide personalized CT learning experiences via gameplay. They introduced AutoThinking, which is a 2D agent-based computer programming game. This game allowed players to use a collection of icons to control a game character's movement in a maze environment. They adopted Bayesian networks algorithm to decide the adaptivity level of students' gameplay. A game system automatically assessed players' CT skills and presented different types of game character movement patterns (i.e., random, provocative, aggressive, and lenient). Despite a promising view of adaptivity implementation in DGBL for computing education, limited research has demonstrated how to orchestrate systematic and data-driven decision-making with adaptive DGBL design. Specifically, few studies discussed how to implement data analytics to drive the design of adaptivity in DGBL.

### 2.3. Evidence-centered design and data analytics

For learner analysis and corresponding adaptive support in DGBL, research has suggested implementing stealth assessment. Stealth assessment is designed to collect students' competency states in an unobtrusive way (Moore & Shute, 2017). Evidence-centered design (ECD) provides rationales for the implementation of stealth assessments (Shute & Kim, 2014). ECD is a framework with which to design learning assessments to measure students' knowledge, skills, and attitudes. To detect student learning states through stealth assessment, research used various data analytics that model learners' competency (e.g., Akram et al., 2018; Min et al., 2019). However, existing competency models typically focus on evaluating the entire learning history, but they are limited in collecting and analyzing in-situ data indicating individuals' learning trajectories in real time. In research of DGBL, previous predictive modeling approaches tend to compute cumulative performance levels instead of their chronological development of gameplay learning experiences. For instance, previous featured DGBL studies with ECD frameworks (Shute & Moore, 2017; Ke & Shute, 2015; Levy, 2019) used Bayesian networks to compute the conditional probability to operate the adaptivity during gameplay. To determine game adaptivity levels, they discretized a granular level of game log data by accumulations. However, this approach has limited success in understanding learners' behavior from a chronological perspective and projecting individuals' gameplay sequences that function as a proxy of their way of thinking during gameplay.

To better capture student learning trajectories in gameplay, emerging research has introduced SDA in DGBL (Moon & Liu, 2019; Tlili et al., 2021). Given a pronounced concern of existing prediction models above, SDA is advantageous to better capturing and delineating learners' temporal and salient sequences of gameplay behaviors representing individuals' "learning paths." Because students' gameplay patterns are likely to expose their knowledge paths in learning tasks, SDA enables researchers to better understand whether and how students face learning challenges in gameplay. Gameplay patterns indicate children's understanding of given game rules and clues. If a child goes to wrong paths and actions related to a game task, it indicates students' game challenges. Under this analytics assumption, DGBL research increasingly tends to use SDA to measure students' patterns of self-regulated learning (Kinnebrew et al., 2015) and scientific reasoning (Taub et al., 2018). Given that SDA is particularly useful to visualize individuals' way of thinking amid a collection of gameplay event data, it is useful to be implemented in DGBL for computing education. Since analytics in computing education requires researchers to identify students' stepwise compilation of blocks to execute their codes with success, SDA can be useful to gather relevant evidence effectively.

## 2.4. Research gap

Despite aforementioned challenges, limited research has implemented data analytics to better capture, model, and understand children's learning states during related to CT development. Existing data analytic approaches in DGBL rarely analyzed how students learn and what challenges occur aligned with game contexts. Corresponding to such problems, this study proposes and implement an SDA-driven framework to provide evidence of designing personalized learning experiences of CT in DGBL. Aligned with this study's goal, we propose research questions as follows.

(1) What are the emerging gameplay patterns among children who played Penguin Go?

(2) What are the differences in gameplay patterns between children in different game conditions (i.e., with or without additional cognitive support)?

(3) What are the design implications of the highlighted gameplay patterns in terms of promoting personalized learning experiences and the development of transferrable CT skills?

To answer the research questions, we have implemented three steps: (1) implementing an educational game (Penguin Go) for CT development; (2) building SDA-driven assessment framework DGBL for adaptivity design; and (3) implementing a case study to explore the relationships among children's gameplay patterns, CT skill development, and learning transfer as the evidence for adaptivity design.

# 3. Method

### 3.1. Penguin Go and computational thinking skills

Penguin Go is an educational game teaching block-based programming language for both elementary and middle school students' CT development developed by the research team (Liu & Jeong, 2022; Zhao & Shute, 2019). This game provides various game tasks to children in the context of the breeding behaviors of emperor penguins. The game's goal is to move the penguin to the destination (i.e., the footprint) using different combinations of code blocks (Figure 1). The game has 18 levels in total. Players need to plan the path of the penguin strategically based on the level terrain. For example, the penguin can waddle on snow (i.e., the light blue blocks) but will slip on the ice (i.e., the deep blue blocks) and has to travel with a toboggan. Table 1 demonstrates the relationships between CT competencies and the concepts covered in the game.



Figure 1. Level "Which Way?" in Penguin Go and a possible solution

~~ .			-	
CT competencies	Sequence	Conditional	Loop	Description
	structure	structure	structure	
Decomposition	Х	X	Х	Identify the goal of each level, the potential pathways, constraints, and patterns in a solution.
Algorithm thinking	Х	Х		Translate the solution into a sequence of blocks that guide the penguin through the maze.
Abstraction		X	Х	Use as few blocks as possible in the solution. Successful implementation of the conditional structure and loop structure can increase the abstraction level of the solution.
Debugging and iteration	Х	Х	Х	Identify the problems and improve the solutions iteratively if the coding blocks do not work as desired

|--|

### 3.2. SDA-driven assessment framework of DGBL for adaptivity design

Previous research using *Penguin Go* suggested that children tend to undergo difficulty developing abstract thinking (Zhao & Shute, 2019). Abstract thinking is one of the hard-to-achieve but a core CT competency for K– 5 children (Lye & Koh, 2014; Wing, 2008; Zhang & Nouri, 2019). In this study, we aim to design a personalized support mechanism that promotes children's transferrable CT across various contexts. Empirical evidence has also shown, however, that mandatory instructional activities might reduce autonomy, which hinders motivation and engagement (Clark et al., 2011; Zhao & Shute, 2019). Therefore, personalized learning supports should be delivered to the children during their in-game problem solving. With personalized learning supports, children are more likely to engage in gameplay instead of receiving instructions passively.

We propose an SDA-driven framework to assess young children's gameplay that evidence of designing adaptivity in DGBL. Here, we aim at identifying meaningful gameplay patterns related to children's either CT development and game challenges. We then focus on exploring how to inform the design of adaptivity based on gameplay results extracted from SDA, putting forth the methodological framework to guide the adaptivity design integrated with SDA.

Figure 2 presents our methodological framework. This framework consists of three major phases based on both the ECD approach (Mislevy et al., 2003) and the four-process adaptive cycle (Shute & Zapata-Rivera, 2012): (1) evidence identification; (2) evidence accumulation; and (3) activity selection. In comparison to the existing adaptive cycle, the proposed framework specifies what kinds of data the system capture in DGBL (e.g., frequent play patterns). Whereas the architecture of the original adaptive cycle poses a general adaptivity design, the proposed model better contextualizes data collection and analyses aligned with SDA. For example, in evidence identification, this framework particularly collects data that orderly arranges a chain of multiple behavior states. Such a collection of behavior states represents students' gameplay patterns that imply decision-making processes. If a sequence of specific game actions is frequent, it is defined as an emerging pattern of gameplay. Whereas existing frameworks tend to emphasize the macro level of adaptivity design and implementation, the proposed framework particularly aims at capturing the in-situ data containing children's gameplay patterns in the adaptive system cycle.

The framework depicts how best to guide children's personalized learning and design adaptivity in DGBL. *Evidence identification* refers to the phase of collecting children's behavioral data through computer logs and/or qualitatively annotated behavior codes and using SDA techniques to identify frequently occurring behaviors or emerging sequence patterns. The identified evidence describes children's gaming sequences and serves as the empirical evidence for the later phases. The purpose of *evidence accumulation* is to interpret existing input data (*evidence identification*) via external measures because identifying the noticeable pattern may not necessarily be self-explanatory. In this phase, we can understand the identified emerging patterns and behaviors in context. For example, we can determine whether a substantial behavioral difference between high performers and low performers is present. As a result, evidence can be accumulated to infer children's competency and identify the potential challenges children are facing, which, in turn, inform the design of the task models. The *activity selection* phase adjusts the instructional activity based on the evidence identified and accumulated (i.e., adaptivity). The goal of this adjustment is to match the appropriate support to children and elicit further

behaviors that feedback to the evidence identification phase. Researchers need to select which learner variables to estimate (e.g., cognitive competency, problem-solving states, affective states), when to intervene, and which instructional content or support to present.



*Figure 2.* Schematic representation of the proposed conceptual framework

#### 3.3. Study procedure

We conducted a case study with an experimental design at two large K-8 schools with a diverse student population in the southeast of the United States. The population was selected because (a) the game was designed for elementary school students, and (b) computational thinking and programming learning opportunities have often been reserved for more advantageous groups (Lachney et al., 2021). The goal of this case study is to understand children's gameplay data and discuss what learning supports are appropriate based on the collected data under the proposed methodological framework. In total, 85 students enrolled in the study, and six students dropped out because of various reasons, including lack of interest or not finishing the posttest. The sample consisted of 79 children (43 self-reported to be female and 27 self-reported to be male; ages ranged from 9 to 11 years old with a median of 10). About half of the sample was from underrepresented ethnic groups (i.e., 22 Black or African American students, 7 Hispanic students, and 2 American Indian or Alaska Native students). We randomly assigned all participants to one of two conditions prior to the experiment; control or treatment. The control group (n = 39) only accessed the *Game Mechanism Support* (GMS) during the gameplay voluntarily. Besides the GMS, the treatment group (n = 40) voluntarily interacted with additional cognitive support in the form of information prompts and partial worked examples (i.e., Concept-Specific Support and Level-Specific Tips, Table 2) in addition to experiencing GMS. We used this treatment design to validate the efficacy of cognitive supports on children's CT development. Here, aligned with the scope of this study, we focus on reflecting the design implications from the experiment not investigating the treatment effect. The study participants joined five 50-minute class sessions and yielded a total of 135 minutes of gameplay. We assessed children's CT development at the pretest, near transfer, and far transfer levels.

Table 2. Supports in Penguin Go

Support	Description
Game Mechanism Support	Static explanations and examples of the programming concepts
Concept-Specific Support	Interactive prompt that introduces the new block
Level-Specific Tips	A partial worked example that (1) encourages the use of a minimum number of
	blocks, (2) presents the target block, and (3) presents other blocks that nest inside
	the loop

Note. The game mechanism support can be accessed by both groups voluntarily. Only treatment group could access Concept-Specific Support and level-specific Tips.

### **3.4. Instruments**

### 3.4.1. CT tests

We developed and implemented three tests to assess children's CT skills. All tests were designed based on the Computational Thinking Test (CTt; Román-González et al., 2017). The pretest was a simplified version of the CTt (17 items). Based on the pretest, we also developed the near transfer test (NTt) that presents the problems in the context of *Penguin Go* while sharing the identical solutions of CTt. Finally, the far transfer test (FTt) mirrored the pretest in terms of the solutions but presented the problems beyond navigating through mazes. All three tests were isomorphic to each other regarding the CT competencies and concepts involved (Figure 3).





### (c) FTt

### 3.4.2. Gameplay data

We collected gameplay logs to identify children's game interactions. All game interactions are logged. Gameplay logs included the data of (a) starting/ending the level; (b) creating/deleting a new block in the solution; (c) changing an existing block; (d) running coding blocks; (e) resetting the position of the penguin; and (f) accessing support. The log data also contained the game ID, action, level, code, and timestamp (an example is presented in Table 3). For data analysis, we removed the time gap between study sessions and aggregated each individual child's gameplay as one unit of analysis. Table 4 shows the descriptive data of each behavior. However, the raw descriptive data only did not indicate how children solve problems in Penguin Go. Therefore, we implemented SDA for further analyses.

	Table 3. Sample gameplay data					
User ID	Verb	Object	Level	Timestamp		
tsms009	start	level	0-5	18:26:15		
tsms009	create	blocks	0-5	18:26:40		
tsms009	create	blocks	0-5	18:27:05		
tsms009	run	blocks	0-5	18:27:07		
tsms009	change	blocks	0-5	18:27:21		
tsms009	reset	blocks	0-5	18:27:37		
tsms009	access	support	0-5	18:30:02		
tsms009	run	blocks	0-5	18:31:29		
tsms009	end	level	0-5	18:31:37		

Table 4. Descriptive game interaction data							
	Trea	tment	Cor	ntrol	Т	Total	
	Mean	SD	Mean	SD	Mean	SD	
Start level	19.40	5.986	14.87	4.354	17.16	5.687	
End level	15.15	4.481	14.64	4.094	14.90	4.275	
Create blocks	182.75	72.875	184.87	60.270	183.80	66.530	
Change blocks	27.83	14.595	28.77	14.377	28.29	14.403	
Delete blocks	32.25	16.295	34.23	11.966	33.23	14.266	
Reset blocks	50.98	23.818	56.95	22.797	53.92	23.364	
Run blocks	66.10	25.129	71.49	23.124	68.76	24.158	
Access support	17.53	15.563	5.03	5.747	11.35	13.295	
Total	411.98	140.544	410.85	108.160	411.42	124.804	

### 3.5. Sequential data analytics

As a technique of SDA, we conducted sequential pattern mining (SPM) with a *cSPADE* algorithm to understand children's gameplay patterns (Zaki, 2001). The purpose of sequential pattern mining here was to identify emerging gameplay patterns that most likely to occur. Each sequence refers to the gameplay data of one level completed by one student, and the chain of multiple sequences pattern consisted of several gameplay events that orderly occurred. We preset the sequence gap to be 2 (i.e.,  $max\_gap = 2$ , where the next event in the identified pattern should appear within two steps of the prior event but are not necessarily consecutive). The minimum support of a sequence was preset to be .5 (i.e.,  $min\_sup = .5$ ; only displaying the frequent sequence patterns that occur over 50% of the time across all children's gameplay). If the support of a particular sequence was detected to be .6, it indicates that 60% of children's gameplay demonstrates such sequence.

#### 3.6. Qualitative observations and field notes

In addition to the quantitative data collection (i.e., group comparison of CT tests and sequential pattern mining), we also conducted qualitative data analysis through behavior observations from facilitators. Four facilitators managed the gameplay sessions and then took notes on children's in-game problem solving and gameplay challenges. Specifically, the observation and field notes focused on (a) the gameplay experiences, (b) problem solving approaches, (c) attitudinal reactions, and (d) study logistics. At the end of each session, the facilitators debriefed their observations. We compiled and analyzed all the qualitative data through multiple rounds of open coding. The analysis focused on identifying children's particular gameplay behaviors and notable problem-solving patterns during the experiment. The qualitative data is used as secondary data to provide triangulation and contextual information to the quantitative findings.

# 4. Results

In the following sections, we present our study findings in accordance with our research questions and the proposed conceptual framework (i.e., evidence identification, evidence accumulation, and activity selection).

### 4.1. RQ1: Sequence pattern emerged (evidence identification)

We first modeled all the children's in-game behaviors across all levels with sequential pattern mining. The probability of behavioral transition is shown in Figure 4. We identified 28 sequence patterns containing five unique behaviors based on the threshold (i.e., *min\_sup* = .5 and *max\_gap* = 2). Among the identified patterns, the most frequent behavior was *Create Blocks*, which appeared in 26 sequence patterns. *Run Blocks* appeared in 16 patterns, and *Reset Blocks* was present in 13 patterns. The least frequent behavior patterns were *Delete Blocks* and *Change Blocks*, which appeared in only five of the patterns and one of the patterns, respectively. *Access Support* did not appear in any of the patterns. This result suggests that children relied more on solution implementation (i.e., *Create Blocks* and *Run Blocks*) rather than refining solutions (i.e., *Reset, Delete*, and *Change Blocks*). The average support value for the identified sequential patterns was .67. We examined the top 10 gameplay sequences with the highest support values to identify emerging gameplay patterns among all children (Table 5). The support values ranged from .65 to .97.

Table 5. Most frequent sequence patterns identified across of
---

Rank	Sequence	Support	Category
1	${\text{create blocks}} \rightarrow {\text{run blocks}}$	0.971	SI
2	${\text{create blocks}} \rightarrow {\text{create blocks}}$	0.943	CI
3	${\text{create blocks}} \rightarrow {\text{create blocks}} \rightarrow {\text{run blocks}}$	0.909	SI
4	${\text{create blocks}} \rightarrow {\text{create blocks}} \rightarrow {\text{create blocks}}$	0.841	CI
5	${\text{create blocks}} \rightarrow {\text{create blocks}} \rightarrow {\text{create blocks}} \rightarrow {\text{run blocks}}$	0.800	SI
6	${\text{create blocks}} \rightarrow {\text{create blocks}} \rightarrow {\text{create blocks}} \rightarrow {\text{create blocks}}$	0.756	CI
7	{create blocks} $\rightarrow$ {create blocks} $\rightarrow$ {create blocks} $\rightarrow$ {create blocks} $\rightarrow$ {run	0.703	SI
	blocks}		
8	$\{\text{reset blocks}\} \rightarrow \{\text{run blocks}\}$	0.690	SE
9	{create blocks} $\rightarrow$ {create blocks} $\rightarrow$ {create blocks} $\rightarrow$ {create blocks} $\rightarrow$ {create blocks}	0.661	CI
	blocks}		
10	${\text{create blocks}} \rightarrow {\text{reset blocks}}$	0.656	SE

*Note.* See Table 6 for details about solution implementation with execution (SI), consecutive solution implementation (CI), and solution evaluation (SE).

We classified gameplay patterns into three categories: (a) solution implementation with execution (SI, Pattern 1, 3, 5, 7), (b) consecutive solution implementation (CI, Pattern 2, 4, 6, 9), and (c) solution evaluation (SE, Pattern 8 and 10). SI patterns start with block creation and end with running the blocks, and CI patterns only consist of consecutive block creation. SE patterns involve *Reset Blocks* compared to SI and CI. *Reset Blocks* refers to resetting penguin position in the game, which does not appear until the blocks begin to run. *Reset Blocks* happens only when someone would like to interrupt the execution of the algorithm. Table 6 summarizes the characteristics and implications of each sequence pattern.

Table 6.	Categories	of sequen	ce patterns
----------	------------	-----------	-------------

Category	Pattern description	Implications
Solution	Start with a series of	Implements and executes a solution with a clear algorithm
implementation	Create Blocks and end	in mind. The frequent occurrence of the SI behavior
with execution (SI)	with Run Blocks.	indicates the trial-and-error problem-solving heuristic,
		which is often inefficient.
Consecutive solution	Only contains	Does not have a clear plan of the algorithm, which could
implementation (CI)	consecutive Create	indicate unsystematic exploration or sometimes random
	Blocks with no Run	block creation.
	Blocks.	
Solution evaluation	Contains Reset Blocks in	Interrupts the solution execution. Involves prediction of
(SE)	combination with Run	where the penguin is moving and the evaluation of the
	Blocks or Reset Blocks.	solution. Often associate with debugging.

Based on the descriptive results of gameplay sequences and each game behavior, we also infer children's problem-solving patterns. First, we suggest that the children tended to undergo inefficient problem-solving heuristics—such as (a) the frequent occurrence of CI patterns because many levels (e.g., loop levels) can be solved with just a few blocks and frequent block creation could indicate hesitation and trial-and-error, (b) multiple trials on one level (e.g., 4.83 runs per level completion), and (c) infrequent change of blocks (e.g., 1.65 changes per level start). Second, we found less frequent prediction- and evaluation-related gameplay patterns, indicating children's lack of systematic problem solving. Third, the absence of accessing learning support in the

gameplay patterns suggests that children used few learning supports and appeared less mindful in problem solving. Such findings highlight that children should have experienced more personalized supports, guiding their in-game problem solving. Overall, these findings help a DGBL system to tentatively identify the noticeable gameplay patterns that can be used for evidence accumulation.

#### 4.2 RQ2: Understanding interaction pattern in situ (evidence accumulation)

#### 4.2.1. Performance data

We first examined the performance difference between the two experimental conditions (Figure 4). The regression analysis results suggested that when controlling the pretest, both groups performed equally well on near transfer (t(76) = -.62, p = .54) and the control group outperformed the treatment group at the far transfer level (t(76) = -2.69, p = .009).





#### 4.2.2. Behavioral data

We then investigated the difference between the two conditions regarding the sequence patterns. The same threshold ( $min\_sup = .5$ ,  $max\_gap = 2$ ) was used to keep consistent with the previous analysis. Table 7 shows a summary of the top 10 frequent gameplay patterns we identified.

Based on the classification, both conditions demonstrated similar patterns in terms of the most frequent behaviors. More than 70% of children's gameplay demonstrated similar SI and CI behavioral patterns in the treatment and control group based on the support value. In addition, SE patterns were relatively less frequent, and the support access was minimal. However, the children in the control condition demonstrated more frequent SE patterns than those in the treatment condition.

Children's sequence patterns demonstrate a high-level summary of their gameplay. As a result, we can infer that the similarity in general behavior patterns between the two groups could potentially explain why children in both two conditions performed equally well at the near transfer. However, the difference in engagement of SE could possibly contribute to the performance difference at the far transfer level.

Table 7. Most frequent sequence patterns identified by condition

	Treatment group			Control group		
#	Sequence	Support	Category	Sequence	Support	Category
1	{create blocks}→{run	0.961	SI	{create blocks}→{run	0.981	SI
	blocks}			blocks}		
2	{create blocks}→{create	0.923	CI	{create blocks}→{create	0.964	CI
	blocks}			blocks}		
3	$\{\text{create blocks}\} \rightarrow \{\text{create }$	0.881	SI	$\{\text{create blocks}\} \rightarrow \{\text{create blocks}\}$	0.940	SI
	blocks}→{run blocks}			blocks}→{run blocks}		
4	{create blocks}→{create	0.822	CI	{create blocks}→{create	0.862	CI

	$blocks\} \rightarrow \{create blocks\}$			$blocks$ $\rightarrow$ {create $blocks$ }		
5	{create blocks}→{create	0.774	SI	$\{\text{create blocks}\} \rightarrow \{\text{create}\}$	0.827	SI
	blocks}→{create			blocks}→{create		
	blocks} $\rightarrow$ {run blocks}			blocks} $\rightarrow$ {run blocks}		
6	$\{\text{create blocks}\} \rightarrow \{\text{create blocks}\}$	0.729	CI	{create blocks} $\rightarrow$ {create	0.786	CI
	blocks}→{create			blocks}→{create		
	blocks} $\rightarrow$ {create blocks}			$blocks$ $\rightarrow$ {create blocks}		
7	$\{\text{create blocks}\} \rightarrow \{\text{create}\}$	0.666	SI	{create blocks} $\rightarrow$ {create	0.743	SI
	blocks}→{create			blocks}→{create		
	blocks}→{create			blocks}→{create		
	blocks} $\rightarrow$ {run blocks}			blocks} $\rightarrow$ {run blocks}		
8	{reset blocks} $\rightarrow$ {run blocks}	0.665	SE	{reset blocks}→{run	0.717	SE
				blocks}		
9	{create blocks}→{create	0.639	CI	{create blocks}→{reset	0.689	SE
	blocks}→{create			blocks}		
	blocks}→{create					
	blocks}→{create blocks}					
10	{create blocks}→{reset	0.626	SE	{run blocks}→{reset	0.668	SE
	blocks}			blocks}		

*Note.* See Table 6 for details about solution implementation with execution (SI), consecutive solution implementation (CI), and solution evaluation (SE).

### 4.2.3. Qualitative data: Data triangulation

To further understand the difference in children's performance and gameplay patterns, we then triangulated SPM results with behavior observations from facilitators' field notes and debriefing results. The qualitative data included primarily four categories: (a) the gameplay experiences (e.g., number of levels played, challenges students had, notable game interactions such as accessing learning resources), (b) problem solving approaches (e.g., trial-and-error, pause-and-think, disengagement), (c) attitudinal reactions (e.g., excitement, confusion, boredom), and (d) study logistics (e.g., technological issues). In this study, we aim to use qualitative as the secondary data to ensure the consistency and trustworthiness of the quantitative findings. Specifically, we identified three notable themes through the qualitative data regarding children's gameplay (i.e., RQ1 and RQ2). First, the field notes in behavior observations reported that children relied on inefficient problem-solving approaches such as trial-and-error. Facilitators observed that some children were frequently moving back and forth between creating blocks and running blocks and built a solution incrementally. One facilitator noted that some children did not spend time reading the pre-level prompts when a new block was introduced.

Second, children were less engaged in problem decomposition and debugging in the gameplay. The children appeared impatient because they tended to construct a solution and immediately delete blocks back after the penguin failed to move to the destination. Given that children's solutions comprise simple sequence structures, this result suggests that the children did not demonstrate a high level of abstraction during the in-game problem solving. They tended to choose simple solutions, which involve fewer cognitive resources.

Finally, the behavior observation also indicated that children did not access the learning support very often. Some children in the treatment group even used the in-level tips as cognitive shortcuts to plan simple solutions. The tips ended up being a "cheat sheet" to them and did not guide them to plan or evaluate their solutions.

These findings further explain the patterns in the context of CT development and transfer. The results address the potential challenges to children's gameplay and learning. They indicate which helps to inform the activity selection phase in designing adaptivity for DGBL. The triangulation from the qualitative data provides further support to the previous SDA findings, which are the basis of the design adaptivity.

### 4.3. RQ3: Design implications of personalization (activity selection)

### 4.3.1. Using SDA to understand in-game problem solving

One of the challenges of the current version of *Penguin Go* is that children demonstrated inefficient problemsolving heuristics and did not interact with the cognitive supports under the voluntary condition. Based on the game challenges, we found evidence of designing adaptivity from a competency-driven approach, emphasizing children's problem solving. With SDA implemented, the game can (a) infer children's general problem-solving competency (i.e., game performance history and pattern recognized); (b) monitor the noticeable sequence patterns; and (c) infer the stage of in-game problem solving.

### 4.3.2. Adaptive game challenges

Adaptive game challenges can guide children to focus on target skill acquisition and abstraction on knowledge. Based on the previous analyses, we concluded that children tended to demonstrate mostly SI and CI rather than CI, which can be inefficient. If such gameplay patterns emerge continuously, this continuous occurrence of the patterns indicated that children do not mindfully engage in problem solving particularly related to abstract thinking. Therefore, imposing constraints on the number of blocks (e.g., Zhao & Shute, 2019) can guide children to mindfully plan their solutions because of the limited resources. Moreover, based on the student gameplay proficiency (e.g., level completion time), constraints can be adjusted accordingly. In the context of the current study, one indicator that we can use is the support value of CI patterns being consistently higher than 90% across multiple levels, given that the population demonstrate such pattern more than 90% of the time on average. However, this baseline might vary across different populations with different proficiency levels.

### 4.3.3. Adaptive cognitive supports

While constraints provide personalized challenges, adaptive cognitive supports provide personalized support. For example, when CI patterns frequently occurred within one level (an indication of unsystematic problem-solving), the game delivered cognitive supports that helped children understand the content knowledge. When repetitive SI emerged, cognitive supports—such as worked examples—were delivered to help children refine solutions. SDA can help to identify these gameplay patterns by setting the minimum support value: if the algorithm detects a frequent pattern (e.g.,  $min\_sup > .5$ ), the game will trigger the relevant support.

### 4.3.4. Adaptive meta-cognitive supports

Children's unsystematic problem solving was related not only to inefficient uses of cognitive resources but also to the limited access to meta-cognitive resources (Azevedo & Aleven, 2013). Such unsystematic problem-solving pattern is supported by children's infrequent SE pattern, and even the control group outperformed the treatment group at the far transfer level. SDA is viable to identify what type of meta-cognitive support should be presented and when to intervene within the game level. For instance, once the cumulative gameplay sequences of a child indicate the infrequent SE pattern during gameplay, a game needs to deliver meta-cognitive supports (e.g., analysis prompts, evaluation guides, or reflection activities) upon individuals' diverse paths. Furthermore, children's gameplay action transitions (e.g., consecutive block creation, resetting, or deleting blocks) indicate various problem solving stages (e.g., wheel-spinning or solution refinement). Based on the identified gameplay pattern results, we can then match the appropriate meta-cognitive supports to the individuals' play to support systematic gameplay related to CT development.

# 5. Discussion

This study implemented SDA into DGBL—performing an assessment to inform evidence of adaptivity design to promote young children's CT development. Based on our analysis findings, we discuss how each phase of the proposed framework helped to design children's personalized DGBL learning experiences by adaptivity design.

### 5.1. Using SDA to facilitate the evidence identification

SDA benefits researchers in collecting and identify the evidence of children's gameplay behaviors for designbased research in a game environment. The results of gameplay patterns in this study demonstrated young children's challenges overall when the supports were not tailored to individuals' diverse learning trajectories. Specifically, the children experienced difficulty in building a correct solution throughout in-game tasks without personalized support. Such patterns also represent students' challenges, including inefficient gaming performance and low understandings of CT during gameplay. These results are aligned with previous research that young children had difficulty mastering the concept of loops and conditional statements to build a complete solution (Ching et al., 2018). Children's such challenges augment the significance of guidance in experiential and interactive learning environments—considering young learners' cognitive capability (Ke et al., 2019; Kirschner et al., 2006; Mayer, 2004). In other words, the data in the *evidence identification* phase shows preliminary evidence of when and how to provide adaptive supports to guide children's problem-solving and promote solution design based on their current learning states.

When it comes to designing adaptivity for DGBL, SDA distilled students' gameplay data (a chain of sequences) and then examined children's frequent play paths as quantitative and contextual evidence. Given that an adaptive game system collects, assembles the evidence, and makes empirically data-driven decisions, at this stage, SDA illustrates what kinds of gameplay pattern data emerged and estimate children's states of game successes and challenges by estimating the frequency of certain gameplay pattern data. The information is essential to build different predictive supervised or semi-supervised algorithms for the purpose of learner modeling in designing adaptive DGBL systems (e.g., Almond et al., 2020; Basu et al., 2017; Rowe et al., 2021).

### 5.2. Validating and triangulating the evidence accumulation

The *evidence accumulation* phase in this study helped researchers to ensure the validity of data collected from SDA. For example, the group comparison of the interaction pattern and how the learning transfer performance was related to patterns highlighting the importance of children's SE patterns and the inefficiency of CI patterns in DGBL. Using triangulation, we further corroborated these findings. The children in the treatment group, with additional cognitive supports, tended to misuse the supports. The supports helped students at the near transfer level but not necessarily at the far transfer level. In comparison, the children in the control group, without cognitive supports, were more likely to engage in SE. Such pattern was not related to the near transfer performance, but it possibly contributed to children's transferrable knowledge and skill development evidenced by the study finding that the control group outperformed the treatment group at the far transfer level.

SDA is an exploratory approach that does not make a priori assumption (Sanderson & Fisher, 1994). The evidence identified, therefore, might not fully reflect students' learning needs. Consequently, researchers need to use external measures (e.g., learning measures, performance measures, or observations) to validate the meaning of the collected evidence. This step helps researchers and practitioners to identify emerging learning interaction patterns in context and further understand the learners' needs and challenges. This is consistent with the call of adding expandability to exploratory approaches of educational data mining (Lim et al., 2021; Shibani et al., 2020). With data triangulation, we identified how specific gaming actions and interactions fostered children's CT development at a fine-grain level. Subsequently, based on the study findings, we can suggest more robust instructional design decisions.

### 5.3. Designing personalized learning experiences with activity selection

Based on our understanding of children's learning interactions and challenges from the previous phases, we yielded decisions of the adaptivity design in Penguin Go. Basically, researchers need to answer three questions in response to designing adaptivity in DGBL: which of the learners' variables to adapt, when to intervene, and which instructional content or support to present (Shute & Zapata-River, 2012). With the help of SDA, we systematically approached these questions using data-driven systems grounded throughout students' gameplay. First, we identified children's needs during gameplay. The SDA findings revealed children's inefficient problem solving. SDA enabled researchers to either monitor noticeable play patterns or estimate the levels of competency in problem solving. Subsequently, the collected data from SDA supported the design decisions as to when and how to intervene children's play (e.g., a behavioral trigger based on observed play patterns or a threshold based on the baseline competency level). Finally, we explored children's interactions with the embedded instructional supports-adaptive game challenges, adaptive cognitive supports, and adaptive meta-cognitive supports based on the children's needs we identified. SDA-driven data collection and decision helped researchers to understand children's interactions with given supports and this data informs which types of supports can be useful across individuals' learning profiles. Through this process, we aim to propose a systematic framework to approach instructional design for DGBL environment driven by learning analytics (c.f. Ifenthaler, 2017). This approach also provides a viable way to design adaptive learning experiences through real-time assessments (e.g., Roll et al., 2011; Rowe et al., 2021; Shute et al., 2020).
## 5.4. Theoretical and practical implications

The study contributes to the previous instructional design research by proposing a framework for applying learning analytics techniques such as SDA in the learning design of adaptive DGBL experiences for computing education. DGBL environments engage children in complex and interactive problem solving, which often needs systematic guidance and facilitation (see Kirschner et al., 2006). Practically, the conceptual framework proposed by this study provides instructional designers with a feasible way to utilize learning analytics in supporting instructional design (Ifenthaler, 2017). Based on the conceptual framework, we provided empirical evidence of how to integrate SDA into DGBL and discussed how to approach the design systematically with multiple sources of data. Specifically, the current study presents a case for how to design personalized learning experiences based on identified learners' needs through SDA.

In addition, the empirical data highlighted children's gameplay patterns and challenges in learning. This further advances the field's knowledge of how children learn through playing and the role of problem-solving in DGBL (c.f. Taub et al., 2020). Both quantitative and qualitative data underscore the needs in children's CT learning and provide practical design recommendations (i.e., game challenges, cognitive supports, and metacognitive supports) about how to potentially address the needs through adaptive design.

## 5.5. Limitations and future directions

This study has a few limitations. First, we did not fully execute a personalized game system, including real-time prediction modeling and not testing the usability of the proposed adaptivity design in DGBL. The scope of the current study was to suggest a methodological framework using SDA that informs evidence of adaptivity design in DGBL. Therefore, future research should develop and contextualize a validated prediction model based on SDA data to measure children's either problem-solving phases or CT development states and examine the efficacy of adaptivity triggered by SDA. Second, we did not refine relevancy behavior codes that indicate how gameplay event transitions refer to specific problem-solving phases. Future studies should refine behavior codes to clearly show the different stages of problem solving. For instance, the data of the SDA appeared skewed because one type of event (e.g., *creating blocks*) dominantly emerged. This event occurred through children's gameplay across different contexts (e.g., consecutive block creations and support abuse that switched back and forth between block creation and support access), but we could not label them differently in this study.

## 6. Conclusions

In this study, we have presented our SDA-driven methodological framework that focuses on collecting evidence of adaptivity design in DGBL. Specifically, using the game Penguin Go, we implemented a case study and the study finding demonstrates how the proposed methodological framework and its implementation ran to detect children's game behavior patterns. Through the case study, SDA identified children's key gameplay patterns and highlighted the effect of solution evaluation on developing CT. Finally, this study has presented design implications based on SDA results in DGBL for computing education.

## Acknowledgement

We wish to acknowledge the contributions of Dr. Weinan Zhao for initially creating Penguin Go. We thank our colleagues, Dr. Ginny Smith, Dr. Demetrius Rice, Chih-pu Dai, Curt Fulwider, and Renata Kuba for their tremendous help in participants recruitment and facilitation of the sessions. We would also like to express our appreciation to the anonymous reviewers for their insightful comments and suggestions. Finally, this research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

# References

Akram, B., Min, W., Wiebe, E., Mott, B., Boyer, K. E., & Lester, J. (2018). Improving stealth assessment in game-based learning with LSTM-based analytics. In *Proceeding of the International Conference of Educational Data Mining* (pp. 208-218). https://par.nsf.gov/biblio/10100664.

Arena, D. A., & Schwartz, D. L. (2014). Experience and explanation: Using videogames to prepare students for formal instruction in statistics. *Journal of Science Education and Technology*, 23, 538-548. https://doi.org/10.1007/s10956-013-9483-3

Almond, R. G., Shute, V. J., Tingir, S., & Rahimi, S. (2020). Identifying observable outcomes in game-based assessments. In R. Lissitz & H. Jiao (Eds.), *Innovative psychometric modeling and methods* (pp. 163-192). Information Age Publishing.

Asbell-Clarke, J., Rowe, E., Almeda, V., Edwards, T., Bardar, E., Gasca, S., Baker, R. S., & Scruggs, R. (2020). The Development of students' computational thinking practices in elementary-and middle-school classes using the learning game, Zoombinis. *Computers in Human Behavior*, 106587. https://doi.org/10.1016/j.chb.2020.106587

Azevedo, R., & Aleven, V. (2013). Metacognition and learning technologies: An Overview of current interdisciplinary research. In R. Azevedo & V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 1–16). Springer. https://doi.org/10.1007/978-1-4419-5546-3\_1

Basu, S., Biswas, G., & Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a computational thinkingbased science learning environment. *User Modeling and User-Adapted Interaction*, 27(1), 5-53. https://doi.org/10.1007/s11257-017-9187-0

Bers, M. U. (2020). *Coding as a playground: Programming and computational thinking in the early childhood classroom.* Routledge.

Ching, Y. H., Hsu, Y. C., & Baldwin, S. (2018). Developing computational thinking with educational technologies for young learners. *TechTrends*, 62(6), 563-573. https://doi.org/10.1007/s11528-018-0292-7

Clark, D. B., Nelson, B. C., Chang, H. Y., Martinez-Garza, M., Slack, K., & D'Angelo, C. M. (2011). Exploring Newtonian mechanics in a conceptually-integrated digital game: Comparison of learning and affective outcomes for students in Taiwan and the United States. *Computers & Education*, 57(3), 2178–2195. https://doi.org/10.1016/j.compedu.2011.05.007

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. https://doi.org/10.1016/j.compedu.2012.11.016

Grover, S., Basu, S., Bienkowski, M., Eagle, M., Diana, N., & Stamper, J. (2017). A Framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Transactions on Computing Education*, *17*(3), 1-25. https://doi.org/10.1145/3105910

Grover, S., & Pea, R. (2013). Computational thinking in K-12: A Review of the state of the field. *Educational Researcher*, 42(1), 38-43. https://doi.org/10.3102/0013189X12463051

Hooshyar, D., Pedaste, M., Yang, Y., Malva, L., Hwang, G. J., Wang, M., Lim, H., & Delev, D. (2021). From gaming to computational thinking: An Adaptive educational computer game-based learning approach. *Journal of Educational Computing Research*, 59(3), 383-409. https://doi.org/10.1177/0735633120965919

Hsu, T. C., Chang, S. C., & Hung, Y. T. (2018). How to learn and how to teach computational thinking: Suggestions based on a review of the literature. *Computers & Education*, *126*, 296-310. https://doi.org/10.1016/j.compedu.2018.07.004

Ifenthaler, D. (2017). Designing effective digital learning environments: Toward learning analytics design. *Technology, Knowledge and Learning*, 22(3), 401-404. https://doi.org/10.1007/s10758-017-9333-0

Israel-Fishelson, R., & Hershkovitz, A. (2020). Persistence in a game-based learning environment: The Case of elementary school students learning computational thinking. *Journal of Educational Computing Research*, 58(5), 891-918. https://doi.org/10.1177/0735633119887187

Ke, F., & Abras, T. (2013). Games for engaged learning of middle school children with special learning needs. *British Journal of Educational Technology*, 44(2), 225-242. https://doi.org/10.1111/j.1467-8535.2012.01326.x

Ke, F., Shute, V. J., Clark, K. M., & Erlebacher, G. (2019). Designing dynamic support for game-based learning. In *Interdisciplinary Design of Game-based Learning Platforms. Advances in Game-Based Learning* (pp. 119-140). https://doi.org/10.1007/978-3-030-04339-1\_6

Ke, F., & Shute, V. (2015). Design of game-based stealth assessment and learning support. In C. Loh, Y. Sheng, D. Ifenthaler (Eds), *Serious games analytics* (pp. 301-318). https://doi.org/10.1007/978-3-319-05834-4\_13

Kinnebrew, J. S., Segedy, J. R., & Biswas, G. (2015). Integrating model-driven and data-driven techniques for analyzing learning behaviors in open-ended learning environments. *IEEE Transactions on Learning Technologies*, *10*(2), 140-153. https://doi.org/ 10.1109/TLT.2015.2513387

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An Analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist*, *41*(2), 75-86. https://doi.org/10.1207/s15326985ep4102\_1

Lachney, M., Ryoo, J., & Santo, R. (2021). Introduction to the special section on justice-centered computing education, Part 1. *ACM Transactions on Computing Education*, 21(4), 1-15. https://doi.org/10.1145/3477981

Levy, R. (2019). Dynamic Bayesian network modeling of game-based diagnostic assessments. *Multivariate Behavioral Research*, 54(6), 771-794. https://doi.org/10.1080/00273171.2019.1590794

Lin, C. F., Yeh, Y. C., Hung, Y. H., & Chang, R. I. (2013). Data mining for providing a personalized learning path in creativity: An application of decision trees. *Computers & Education*, 68, 199-210. https://doi.org/10.1016/j.compedu.2013.05.009

Lim, L. A., Gasevic, D., Matcha, W., Ahmad Uzir, N. A., & Dawson, S. (2021). Impact of learning analytics feedback on self-regulated learning: Triangulating behavioural logs with students' recall. In M. Scheffel, N. Dowell, S. Joksimovic, & G. Siemens (Eds.), *11<sup>th</sup> international learning analytics and knowledge conference* (pp. 364-374). ACM. https://doi.org/10.1145/3448139.3448174

Liu, Z., & Jeong, A. C. (2022). Connecting learning and playing: The Effects of in-game cognitive supports on the development and transfer of computational thinking skills. *Educational Technology Research and Development*, 70, 1867-1891. https://doi.org/10.1007/s11423-022-10145-5

Liu, Z., Moon, J., Kim, B., & Dai, C. (2020). Integrating adaptivity to educational games: A Combination of bibliometric analysis and meta-analysis review. *Educational Technology Research and Development*, 68(4), 1931-1959. https://doi.org/10.1007/s11423-020-09791-4

Liu, Z., Zhi, R., Hicks, A., & Barnes, T. (2017). Understanding problem solving behavior of 6–8 graders in a debugging game. *Computer Science Education*, 27(1), 1-29. https://doi.org/10.1080/08993408.2017.1308651

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. https://doi.org/10.1016/j.chb.2014.09.012

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The Case for guided methods of instruction. *American Psychologist*, 59(1), 14-19. https://doi.org/10.1037/0003-066X.59.1.14

McGill, M. M., & Decker, A. (2020). Tools, languages, and environments used in primary and secondary computing education. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education* (pp. 103-109). https://doi.org/10.1145/3341525.3387365

Min, W., Frankosky, M., Mott, B. W., Rowe, J., Smith, P. A. M., Wiebe, E., Boyer, K. E., & Lester, J. (2019). DeepStealth: Game-based learning stealth assessment with deep neural networks. *IEEE Transactions on Learning Technologies*, *13*(2), 312-325. https://doi.org/10.1109/TLT.2019.2922356

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3-62. https://doi.org/10.1207/S15366359MEA0101\_02

Moon, J., & Liu, Z. (2019). Rich representations for analyzing learning trajectories: Systematic review on sequential-data analytics in game-based learning research. In A. Tlili., & M. Chang. (Eds.), *Data Analytics Approaches in Educational Games and Gamification Systems* (pp. 27-53). Springer. https://doi.org/10.1007/978-981-32-9335-9\_2

Moore, G. R., & Shute, V. J. (2017). Improving learning through stealth assessment of conscientiousness. In *Handbook on digital learning for K-12 schools* (pp. 355-368). Springer, Cham.

Morrison, J. R., Bol, L., Ross, S. M., & Watson, G. S. (2015). Paraphrasing and prediction with self-explanation as generative strategies for learning science principles in a simulation. *Educational Technology Research and Development*, 63(6), 861–882. https://doi.org/10.1007/s11423-015-9397-2

Owen, V. E., Roy, M. H., Thai, K. P., Burnett, V., Jacobs, D., Keylor, E., & Baker, R. S. (2019). Detecting wheel-spinning and productive persistence in educational games. In C. F. Lynch, A. Merceron, M. Desmarais, & R. Nkambou (Eds.), *Proceedings of the 12th International Conference on Educational Data Mining* (pp. 378-383). International Educational Data Mining Society.

Román-González, M., Pérez-González, J. C., & Jiménez-Fernández, C. (2017). Which cognitive abilities underlie computational thinking? Criterion validity of the Computational Thinking Test. *Computers in Human Behavior*, 72, 678–691. https://doi.org/10.1016/j.chb.2016.08.047

Roll, I., Aleven, V., McLaren, B. M., & Koedinger, K. R. (2011). Improving students' help-seeking skills using metacognitive feedback in an intelligent tutoring system. *Learning and Instruction*, 21(2), 267-280. https://doi.org/10.1016/j.learninstruc.2010.07.004

Rowe, E., Almeda, M. V., Asbell-Clarke, J., Scruggs, R., Baker, R., Bardar, E., & Gasca, S. (2021). Assessing implicit computational thinking in Zoombinis puzzle gameplay. *Computers in Human Behavior, 120*, 106707. https://doi.org/10.1016/j.chb.2021.106707

Sanderson, P. M., & Fisher, C. (1994). Exploratory sequential data analysis: Foundations. *Human–Computer Interaction*, 9(3-4), 251-317. https://doi.org/10.1207/s15327051hci0903%264\_2

Shibani, A., Knight, S., & Shum, S. B. (2020). Educator perspectives on learning analytics in classroom practice. *The Internet and Higher Education*, *46*, 100730. https://doi.org/10.1016/j.iheduc.2020.100730

Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In *Handbook of research on educational communications and technology* (pp. 311-321). Springer.

Shute, V. J., & Moore, G. R. (2017). Consistency and validity in game-based stealth assessment. In H. Jiao & R. W. Lissitz (Eds.), *Technology enhanced innovative assessment: Development, modeling, and scoring from an interdisciplinary perspective* (pp. 31-55). Information Age Publishing, Inc.

Shute, V. J., Rahimi, S., Smith, G., Ke, F., Almond, R., Dai, C-P, Kamikabeya, R., Liu, Z., Yang, X., & Sun, C. (2020). Maximizing learning without sacrificing the fun: Stealth assessment, adaptivity, and learning supports in educational games. *Journal of Computer-Assisted Learning*, *37*(1), 127-141. https://doi.org/10.1111/jcal.12473

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review*, 22, 142-158. https://doi.org/10.1016/j.edurev.2017.09.003

Shute, V. J., & Zapata-Rivera, D. (2012). Adaptive educational systems. In P. Durlach (Ed.), Adaptive technologies for training and education (pp. 7-27). Cambridge University Press.

Taub, M., Azevedo, R., Bradbury, A. E., Millar, G. C., & Lester, J. (2018). Using sequence mining to reveal the efficiency in scientific reasoning during STEM learning with a game-based learning environment. *Learning and Instruction*, 54, 93-103. https://doi.org/10.1016/j.learninstruc.2017.08.005

Taub, M., Sawyer, R., Smith, A., Rowe, J., Azevedo, R., & Lester, J. (2020). The Agency effect: The Impact of student agency on learning, emotions, and problem-solving behaviors in a game-based learning environment. *Computers & Education*, 147, 103781. https://doi.org/10.1016/j.compedu.2019.103781

Tlili, A., Chang, M., Moon, J., Liu, Z., Burgos, Chen, N., & Kinshuk. (2021). Literature review of empirical studies on learning analytics in educational games: From 2010 to 2018. *International Journal of Interactive Multimedia and Artificial Intelligence*, 7(2), 250-261. http://dx.doi.org/10.9781/ijimai.2021.03.003

Turchi, T., Fogli, D., & Malizia, A. (2019). Fostering computational thinking through collaborative game-based learning. *Multimedia Tools and Applications*, 78(10), 13649-13673. https://doi.org/10.1007/s11042-019-7229-9

Vanbecelaere, S., Van den Berghe, K., Cornillie, F., Sasanguie, D., Reynvoet, B., & Depaepe, F. (2020). The Effectiveness of adaptive versus non-adaptive learning with digital educational games. *Journal of Computer Assisted Learning*, *36*(4), 502-513. https://doi.org/10.1111/jcal.12416

Walkington, C. A. (2013). Using adaptive learning technologies to personalize instruction to student interests: The Impact of relevant contexts on performance and learning outcomes. *Journal of Educational Psychology*, *105*(4), 932–945. https://doi.org/10.1037/a0031882

Weintrop, D., Holbert, N., Horn, M. S., & Wilensky, U. (2016). Computational thinking in constructionist video games. *International Journal of Game-Based Learning*, 6(1), 1-17.

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society* A: *Mathematical, Physical and Engineering Sciences, 366*(1881), 3717-3725. https://doi.org/10.1098/rsta.2008.0118

Zaki, M. J. (2001). SPADE: An Efficient algorithm for mining frequent sequences. *Machine learning*, 42(1-2), 31-60. https://doi.org/10.1023/A:1007652502315

Zhang, L., & Nouri, J. (2019). A Systematic review of learning computational thinking through Scratch in K-9. *Computers & Education*, 141, 103607. https://doi.org/10.1016/j.compedu.2019.103607

Zhao, W., & Shute, V. J. (2019). Can playing a video game foster computational thinking skills? *Computers & Education*, 141, 1-13. https://doi.org/10.1016/j.compedu.2019.1