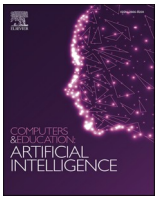
Computers and Education: Artificial Intelligence 3 (2022) 100104

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Adaptive formative assessment system based on computerized adaptive testing and the learning memory cycle for personalized learning

Albert C.M. Yang [a,*], Brendan Flanagan [b], Hiroaki Ogata [b]

[a] Graduate School of Informatics, Kyoto University, Japan, 606-8501, Yoshida-honmachi, Sakyo-ku, Kyoto, Japan
[b] Academic Center for Computing and Media Studies, Kyoto University, Japan, 606-8501, Yoshida-honmachi, Sakyo-ku, Kyoto, Japan

ARTICLE INFO

Keywords:
Personalized learning
Adaptive learning
Formative assessment
Computerized adaptive testing
Learning memory cycle

ABSTRACT

Computerized adaptive testing (CAT) can effectively facilitate student assessment by dynamically selecting questions on the basis of learner knowledge and item difficulty. However, most CAT models are designed for one-time evaluation rather than improving learning through formative assessment. Since students cannot remember everything, encouraging them to repeatedly evaluate their knowledge state and identify their weaknesses is critical when developing an adaptive formative assessment system in real educational contexts. This study aims to achieve this goal by proposing an adaptive formative assessment system based on CAT and the learning memory cycle to enable the repeated evaluation of students' knowledge. The CAT model measures student knowledge and item difficulty, and the learning memory cycle component of the system accounts for students' retention of information learned from each item. The proposed system was compared with an adaptive assessment system based on CAT only and a traditional nonadaptive assessment system. A 7-week experiment was conducted among students in a university programming course. The experimental results indicated that the students who used the proposed assessment system outperformed the students who used the other two systems in terms of learning performance and engagement in practice tests and reading materials. The present study provides insights for researchers who wish to develop formative assessment systems that can adaptively generate practice tests.

1. Introduction

In traditional learning environments, teachers employ one-on-many teaching approaches because of limitations related to human resources, time, and learning tools. However, advances in artificial intelligence (AI) technology have enabled researchers to develop applications that simulate teachers' knowledge and experience to provide personalized support to students (Pai, Kuo, Liao, & Liu, 2021; Xiao & Yi, 2021). The field of artificial intelligence in education (AIED) has demonstrated cases of integrating advanced technology with educational theory to achieve pedagogical impacts (Roll & Wylie, 2016). For example, intelligent tutoring systems have been used for material delivery, feedback provision, and progress monitoring (Bayne, 2015). Al can be used to predict learning performance, engagement, attrition, and retention, enabling instructors to detect at-risk learners early and provide timely intervention. AI also supports modeling learners' behavioral preferences, profiles, interests, and knowledge states, in which the instructors or the AI-enhanced systems can make recommendations (Chen et al., 2021, 2022). The AI-supported eLearning field emphasizes the importance of creating adaptive and personalized learning environments and learning support according to students' learning profiles, including adaptive assessment (Tang et al., 2021). AI-supported technology has been applied in formative assessment for enhancing student learning (Elmahdi et al., 2018). Computer-based assessments not only enable students to identify gaps between their current and desired knowledge but also help teachers improve their teaching methods and monitor students' progress (Tomasik et al., 2018). In addition, immediate and constructive feedback from computer-based assessment tools provides students with timely and personalized assistance, thereby increasing their engagement and facilitating personalized learning (Elmahdi et al., 2018). Rodrigues and Oliveira (2014) discovered that formative assessment tools changed students' study habits by requiring them to start studying earlier to pass practice tests, thereby helping them feel more confident and perform better on the final examination. Although

* Corresponding author.
E-mail addresses: yang.ming.35e@st.kyoto-u.ac.jp (A.C.M. Yang), flanagan.brendanjohn.4n@kyoto-u.ac.jp (B. Flanagan), ogata.hiroaki.3e@kyoto-u.ac.jp (H. Ogata).

https://doi.org/10.1016/j.caeai.2022.100104
Received 25 April 2022; Received in revised form 25 October 2022; Accepted 28 October 2022
Available online 29 October 2022

formative assessments and practice tests have been demonstrated to improve student learning, most students require some encouragement and guidance to apply these techniques. In real educational settings, most students still use relatively ineffective methods, such as relearning material or cramming before final examinations (Karpicke et al., 2009). Dunlosky et al. (2013) surveyed students' attitudes toward different learning techniques, such as attending class regularly, highlighting, and completing practice tests. The survey results revealed that although some students applied effective strategies such as practice testing, others still used less effective approaches. The same result was replicated in a large-scale study by Gurung et al. (2012), in which they determined that the use of practice tests was positively correlated with learning performance in class. One reason for this result is that using practice tests to evaluate knowledge status requires students to exert more effort than does restudying materials or cramming before examinations because practice tests require students to start studying earlier. Nevertheless, repeatedly taking the same formative assessment may lead to boredom among students. Therefore, developing an adaptive and personalized system for formative assessment to encourage and guide students to take practice tests is essential.

Various adaptive formative assessment systems have been developed using computerized adaptive testing (CAT) approaches, such as item response theory (IRT), Elo rating algorithm, or Bayesian knowledge tracing (BKT). These systems usually focused on accurately estimating learners' latent ability and recommending items with appropriate difficulty. Most existing systems, however, did not consider the effect of forgetting. As learners cannot remember everything they learned, their knowledge state may change as the learning process continues. The recommended items may not match learners' latest knowledge state without repeatedly estimating their latent ability. In addition, if prerequisite knowledge is forgotten and not reviewed, learners may find difficulties in learning the subsequent materials, which can decrease their motivation to use the system. Hence, learners' memory length should be considered when designing the adaptive formative assessment system. The current study aimed to address this problem by proposing a system that combines the CAT and learning memory cycle models. The system generates personalized quizzes that contain both unattempted items with appropriate difficulty and attempted items that should be reviewed. The present study provides insights for researchers who wish to develop formative assessment systems that can adaptively generate practice tests by addressing the following research questions:

Q: Can a formative assessment system based on computerized adaptive testing and learning memory cycle increase students' engagement in taking practice tests and learning?

Q: Can a formative assessment system based on computerized adaptive testing and learning memory cycle improve students' learning performance?

2. Literature review

2.1. Computerized adaptive testing for formative assessment

A typical approach used for computerized adaptive testing (CAT) is item response theory (IRT), which uses statistical models that consider the latent traits of students and items to calculate the probability of a correct response to an item (Jia & Le, 2020; Meiser et al., 2019; Reckase, 1997). In this approach, the questions on an adaptive test are selected on the basis of a student's responses to previous items. This approach aims to maximize the information gained from each item and, in turn, to shorten the test length and testing time (Wang et al., 2020). However, when IRT-based methods are used, the item pool of a test must be calibrated before the test is administered to determine the item parameters to be used for calculating the probability of a correct response (Edelen & Reeve, 2007). This may increase the complexity of test creation and maintenance and deter educators from deploying IRT-based tests for formative assessment in practical settings. Similar to IRT, the Elo rating system is another approach for adaptively selecting items on the basis of student ability and item difficulty. Both methods use the same mechanism to predict the probability of a correct answer to an item (Pelánek, 2016); however, in the Elo rating system, item pool calibration is not required before testing. It can dynamically adjust student and item parameters during the test; thus, it is more convenient for use in practical settings. Other approaches for modeling students' knowledge acquisition, such as knowledge tracing (Zhou et al., 2021) and performance factor analysis (Liu et al., 2021), have been developed; however, these methods are impractical because they employ complex models and require additional information for parameter calibration.

Although the aforementioned CAT-based techniques have been used for creating formative assessments, most of them have been used for student evaluation rather than for learning (McCallum & Milner, 2021). CAT systems that employ such techniques evaluate and record students' knowledge through formative assessments and assume that their knowledge will not change. However, because students' knowledge changes over time, repeated reassessment is necessary (Choi & McClenen, 2020). Using these CAT systems for repeated practice may not be appropriate because they do not account for students forgetting information over time. As the parameter estimates for students and items stabilize, the system may start selecting duplicate items, which may make students feel bored.

2.2. Adaptive assessment based on memory

When designing CAT models for formative assessment, students' memorization abilities must be considered. Memorization is the ability to remember learned information. A learner's retention of information will gradually decline if they do not review the information for a long time. However, although reviewing is helpful for learners to extend their retention of learned information, each learner has a distinct ability to retain the same knowledge. Thereafter, the review cycle must be tailored to each learner to maximize their retention of learned information. One advantage of selecting questions on the basis of memory is that it considers several practices, including interleaving, spacing, and repetition, that can help a learner improve their retention of learned information, as summarized by van Kesteren and Meeter (2020). An adaptive formative assessment system that considers learners' memory retention was proposed by Chen and Chung (2008), which achieved promising results. The researchers applied IRT and memory cycle models to develop a mobile system for personalized English vocabulary learning. They discovered that the system significantly improved students' English vocabulary abilities and enhanced their interest in learning. In this paper, we propose an adaptive formative assessment system that considers student ability, memory retention, and item difficulty to enable the repeated evaluation of students' knowledge. We compared the proposed system with an adaptive assessment system based on CAT only and a traditional nonadaptive assessment system by conducting an experiment in a university programming course. Because of its convenience and strong estimation performance, the Elo rating system is used in the proposed system to evaluate student ability and item difficulty parameters; the system calculates memory retention by using the memory cycle model proposed by Chen and Chung (2008) because it accounts for students forgetting information over time.

3. Adaptive assessment system

3.1. Assessment system

The assessment system presented in this paper was developed by Kyoto University and is able to automatically generate fill-in-the-blank questions for given learning materials. It consists of three modules: a text preprocessing module, a text summarization module, and a question generation module. First, the instructor of a course uploads a PDF of the learning materials to the system. The text preprocessing module

converts the PDF into a plain text file and performs standard text preprocessing, including removing stop words, punctuation, and numbers; lowercasing the text; and lemmatizing the content. Thereafter, the text summarization module applies the Bidirectional Encoder Representations from Transformers (BERT) deep learning model (Devlin et al., 2018) to extract important sentences from the preprocessed text. Finally, the question generation module selects keywords from the extracted sentences using TextRank, a natural-language processing algorithm (Mihalcea & Tarau, 2004), and masks the keywords to generate fill-in-the-blank questions. The detailed process of how the system generates questions has been described by Yang et al. (2021). Before the generated questions are presented to students, they are reviewed by the instructor to ensure that they cover the target knowledge. The instructor can modify the question bank by removing existing questions or adding new questions. Although the current version of the assessment system can automatically generate fill-in-the-blank questions, more questions can be generated through the addition of short-answer questions to the question bank by the instructor. The system generates a practice test that includes both the fill-in-the-blank questions automatically generated by the system and any short-answer questions created by the instructor for each topic, and students can use the system to complete the practice tests to review the course material. Figs. 1 and 2 present screenshots of the assessment system in use. After their login into the system, students can choose the e-book they would like to review and decide whether to answer fill-in-the-blank questions (Fig. 1) or short-answer questions (Fig. 2). After a student submits an answer, immediate feedback is provided.

3.2. Elo rating system

The Elo rating system was originally proposed by Elo (1978) and was used for rating chess players. In this context, each player is assigned a rating, and the rating is updated according to the result of each match. If a strong player beats a weak player, the update is small; if the opposite outcome occurs, the update is large. The system has been widely applied in fields that require contest matching, such as online games and sports (Hvattum & Arntzen, 2010). In the educational context, the Elo rating system can be used to evaluate learner ability and item difficulty in a similar manner: each answer attempt is considered a "match" between the learner and the item. The Elo rating system has several advantages: it is a simple system that requires few parameters, and it can be easily implemented in practical settings; it can be easily employed in online environments; and it can achieve performance comparable to that of more complex systems. Evaluation and application of the Elo rating system in education have been widely studied. Wauters et al. (2012) applied an extension of the Elo rating system to estimate item difficulty in a CAT system and compared the system's estimates with those provided by an IRT model. Klinkenberg et al. (2011) used an extension of the Elo rating system in an online system for adaptive practice in a mathematics course. Papousek et al. (2014) and Nižnan et al. (2015) have applied the Elo rating system for the estimation of learner's prior knowledge based on their performance in adaptive practice. In this study, we incorporated the basic Elo rating system into the proposed online formative assessment system to estimate student ability and item difficulty and to generate adaptive practice exercises accordingly. The principle of the Elo rating system is as follows: each student s has skill parameter θ_s, and each item i has difficulty parameter d_i. The correctness of the response of student s to item i is denoted as correct_{si} {0,1}, and the probability of student s answering item i correctly can be represented as a logistic function of the difference between skill and difficulty:

$$P(\text{correct}_{si}=1)=\frac{1}{(1+e^{-(\theta_s-d_i)})} \tag{1}$$

After each answer, the skill and difficulty are updated as follows:

$$\theta_s := \theta_s + K \bullet (correct_{si} - P(correct_{si}=1))$$

$$d_i := d_i + K \bullet (P(correct_{si}=1) - correct_{si}) \tag{2}$$

The initial values of θ_s and d_i are set to 0. Constant K is the uncertainty parameter that determines the influence of each attempt. A small K value may cause the estimation to converge too slowly, whereas a large K value may result in an unstable estimation because the system may place too much weight on the student's last few attempts. A common approach to ensuring high performance is replacing constant K with an "uncertainty function"; that is, the scale of updates for a new learner or a new item should be large because the estimation is still uncertain and because each update brings more information into the system. As the sample size increases, the scale of updates should decrease. In this study, we applied the simple uncertainty function used by Papousek et al. (2014) and Nižnan et al. (2015), which is described as $U(n) = a/(1+bn)$, where n is the number of attempts and a and b are the hyperparameters adjusted for data. We set $a = 1$ and $b = 0.05$ for our dataset.

3.3. Memory cycles

We used the approach proposed by Chen and Chung (2008), which considers item difficulty, learner skill, and response results for updating a learner's memory cycle. The scheme for updating the memory cycle of each learner can be formulated as follows:

$$MC_{t+1}^{si} = MC_t^{si} + \frac{\theta_s}{b_i} \bullet F_{ot_{si}} \quad \text{Correct response}$$

$$MC_{t+1}^{si} = MC_t^{si} - \frac{b_i}{\theta_s} \bullet F_{xt_{si}} \quad \text{Incorrect response} \tag{3}$$

where MC_{t+1}^{si} is the updated memory cycle of student s for item i, MC_t^{si} is the original memory cycle of student s for item i, θ_s represents the ability

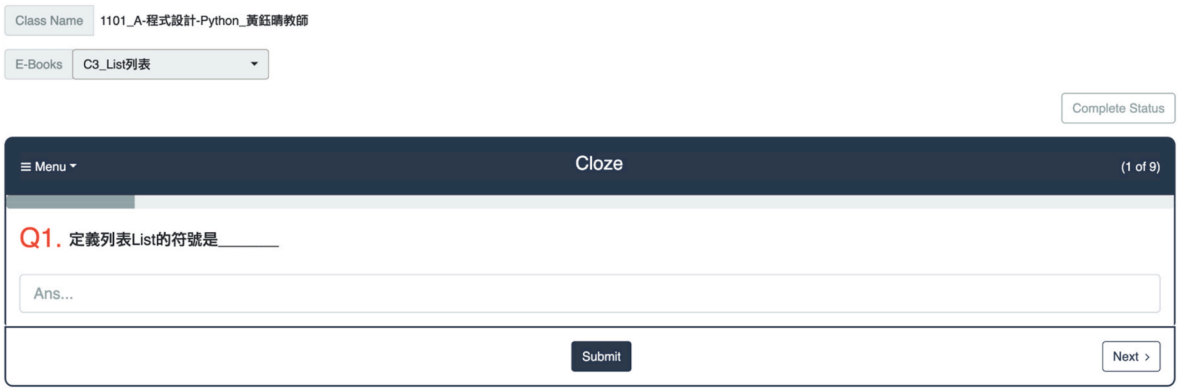


Fig. 1. Fill-in-the-blank question.

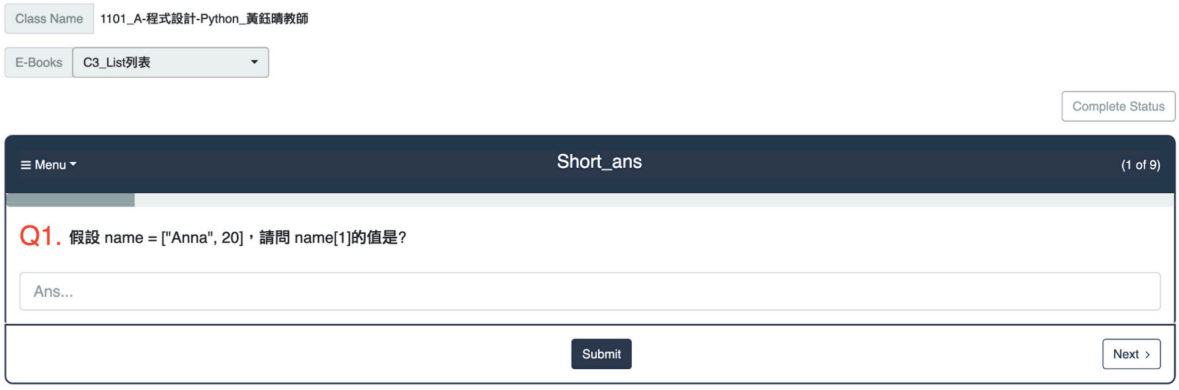


Fig. 2. Short-answer question.

of student s, b_i represents the difficulty of item i, ot_{si} represents the number of times student s has correctly answered item i, xt_{si} represents the number of times of student s has incorrectly answered item i, and F denotes the Fibonacci sequence. The Fibonacci sequence is a series of numbers in which each number is the sum of the previous two numbers, and it can be formulated as follows:

$$F_n = 1 \text{ n} = 1,2$$

$$F_n = F_{n-2} + F_{n-1} \quad \text{otherwise} \tag{4}$$

According to the forgetting curve described by Ebbinghaus (2013), people tend to forget information that they have known for a long time more slowly than they forget newly learned information. Therefore, the gradually increasing numbers of the Fibonacci sequence can be used to update a learner's memory cycle. The proposed scheme updates a learner's memory cycle on the basis of learner ability, item difficulty, and the numbers of times the learner has correctly and incorrectly responded to a given item. The memory cycle will be extended if learner ability is high, item difficulty is low, and the number of correct responses to that item is high and will be shortened if learner ability is low, item difficulty is high, and the number of incorrect responses to that item is high. We set the minimum value of the memory cycle to 0 because it cannot have a negative value.

3.4. Adaptive assessment system based on Elo rating algorithm and learning memory cycle model

The current study proposed an adaptive assessment system based on the Elo rating algorithm and learning memory cycle model that can help students select items to review their knowledge. Compared to existing adaptive assessment systems that recommend items based on latent ability, the proposed system recommends items based on the memory cycle, which can reduce the appearance of repeated items. After students submit their responses, the system would update their latent ability and the memory cycle based on the results in real time. Items with a shorter memory cycle will be prioritized for the recommendation. Since items that were never responded correctly to and items that have not been reattempted for a while will have a shorter memory cycle, it ensures that students can review the concepts they are not yet familiar with and that they might have forgotten. The memory cycle will be extended for the recommended items answered correctly by the student. These items will not be recommended next time if other items with a shorter memory cycle exist. Conversely, the student's memory cycle of items answered incorrectly will be shortened, and the items will be recommended again. This approach enables the system to ensure students' latent abilities are updated and help them select items that need to be reviewed, encouraging them to use the system. In addition, this mechanism facilitates students to employ several critical practices that can help them improve their retention of learned information, including interleaving, spacing, and repetition.

3.5. Adaptive assessment process

Three types of assessment systems were used in this study: (1) the proposed adaptive assessment system based on the Elo rating algorithm and learning memory cycle, (2) an adaptive assessment system based on the Elo rating algorithm alone, and (3) a conventional assessment system. The detailed assessment processes of the three systems are described as follows:

1) Proposed adaptive assessment system based on the Elo rating algorithm and learning memory cycle:

Students use the adaptive assessment system to generate personalized practice tests based on their knowledge level, item difficulty, and their retention of the information learnt from each item. After a student completes an attempt, the system updates the estimated student knowledge and item difficulty by using the Elo rating algorithm and updates the student's retention of the information learnt from each item by using the memory cycle algorithm. Thereafter, for the next adaptive test, the system selects 25 items about the previously learned concepts according to the retention length. The item with a shorter retention length will be prioritized.

2) Adaptive assessment system based on the Elo rating algorithm alone:

Students use the adaptive assessment system to generate personalized practice tests based on their knowledge level and item difficulty. After a student completes an attempt, the system updates the estimated student knowledge and item difficulty by using the Elo rating algorithm. Thereafter, for the next adaptive test, the system selects 25 items about the previously learned concepts according to the probability of the student answering the items correctly. The system prioritizes items that the student has a lower chance of answering correctly because the goal of the practice test is to help students review information.

3) Conventional assessment system:

Students use the conventional assessment system for evaluating their knowledge after class and can decide which practice tests they would like to complete. The items in each test do not change after an attempt.

4. Experimental design

4.1. Participants

This study employed a quasiexperimental design. The study participants were three classes of first-year university students from the Department of Computer Science at a university in Taiwan. All three classes were enrolled in a course titled "Introductory Programming

Language," which covers the fundamental concepts of Python and how to write code in Python. None of the participants had previously taken similar courses or had background knowledge of Python, and all three classes were led by the same instructor. One class of 37 students was designated as experimental group A, which used the proposed adaptive assessment system based on the Elo rating algorithm and learning memory cycle; another class of 37 students was designated as experimental group B, which used the adaptive assessment system based on the Elo rating algorithm alone; and one class of 34 students was designated as the control group, which used the conventional assessment system.

4.2. Experimental procedure

All three groups took a 20-min pretest before the experiment. Thereafter, they participated in learning activities for 7 weeks. Six concepts were taught in the first 6 weeks: data types and variables, operators, lists, if/else conditional statements, for loops, and while loops. Each week, the instructor taught the classes a new concept by using BookRoll, an e-book reading system designed by Kyoto University (Flanagan & Ogata, 2018; Ogata et al., 2015). BookRoll allows instructors to upload slides of learning materials for students to access. Students can perform various actions such as highlighting text, posting memos, and adding bookmarks when reading e-books. In this study, the students learnt new concepts in the classroom using BookRoll, reviewed previous materials using BookRoll, and evaluated their knowledge using their designated assessment system after class. The question banks contained six question sets, one for each concept. Each question set consisted of fill-in-the-blank and short-answer questions. The numbers of questions for concepts 1 up to 6 were 27, 24, 18, 18, 10, and 6, respectively. The two experimental groups using the adaptive assessment systems had access to all the questions in the banks; that is, they were able to attempt questions that were not recommended by their designated systems. The last week of the course was designated as a review week, during which the students could review previously learned concepts, and no new material was taught in this week. At the end of the experiment, all the students took a 40-min posttest and a 20-min computer-based quiz. The full experimental process is presented in Fig. 3.

4.3. Evaluation

The pretest and posttest used in this study was designed by the instructor, who is a Professor in the Department of Computer Science. The students took a pretest before the experiment to evaluate their prior knowledge of programming languages. The pretest comprised 12 factual questions regarding fundamental concepts of programming languages. At the end of the experiment, the students took a posttest, which comprised 20 factual questions regarding the concepts taught during the course. The posttest mainly evaluated the students' basic knowledge of programming languages, and their test scores will be used as their learning performance. The instructor also designed a computer-based quiz that contained 5 coding exercises to measure the students programming skills. Each student's actions when using the assessment system and reading e-books (hereafter "assessment behaviors" and "reading behaviors," respectively) were logged to measure their learning engagement. The actions that the students could perform when using the assessment and e-book reading systems are presented in Table 1. Each student's total number of assessment behaviors, number of questions attempted, and average attempts per question were used to assess their assessment engagement; their total number of reading behaviors and the frequency with which they opened e-books were used to assess their reading engagement.

5. Results

5.1. Analysis of assessment engagement and reading engagement

The learning logs of the students using the different assessment systems were analyzed. Table 2 presents the average frequency with which the students in each group performed NEXT, PREV, and SUBMIT actions. Experimental group A had the highest frequencies for all behaviors. We conducted one-way analysis of variance (ANOVA) to examine the differences in the total numbers of assessment behaviors. Because the students could make unlimited attempts on the practice tests, some of the students may have performed a significantly higher number of actions than did others. We substituted the frequencies of

Table 1
Formative assessment and reading behaviors.

Activity	Behavior	Description
Formative assessment	NEXT	Go to next question
	PREV	Go to previous question
	SUBMIT	Submit answer
E-book reading	OPEN	Open e-book
	NEXT	Go to next page
	PREV	Go to previous page
	ADD_MARKER	Use marker to highlight text
	ADD_MEMO	Post a memo on the page
	ADD_BOOKMARK	Add a bookmark to the page

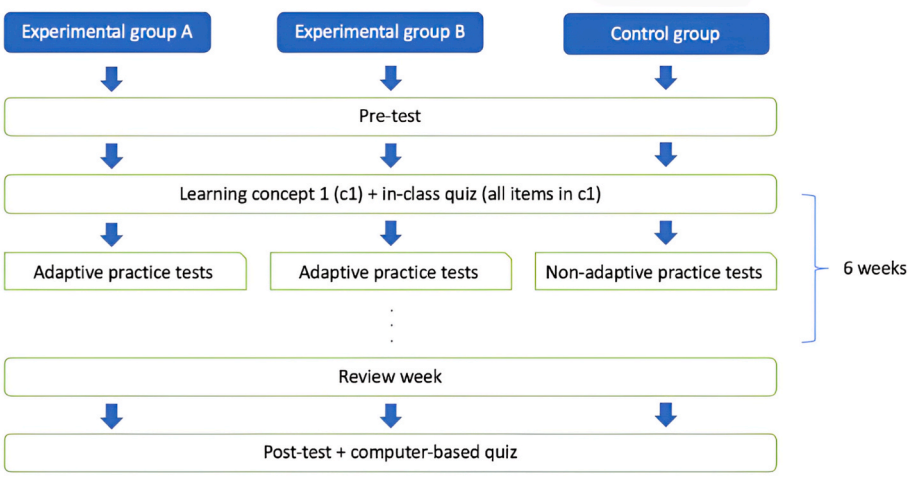


Fig. 3. Experimental process.

Table 2
Mean frequencies of assessment behaviors.

Group	NEXT	PREV	SUBMIT	Total
(a) Experimental group A	211.81	59.56	276.59	547.97
(b) Experimental group B	166.54	37.32	214.25	417.39
(c) Control group	150.37	31.58	190.78	372.74

total assessment behaviors with the logarithms of the frequencies to account for the skewed data. Levene's test indicated that the assumption of homogeneity of variance among the average logarithms of the frequencies was satisfied ($F = 2.95$; $p = 0.055 > 0.05$). As indicated in Table 3, the intergroup difference ($F = 9.66$; $p = 0.00 < 0.05$) in the logarithm of the frequency of total assessment behaviors was significant. The geometric mean frequency of experimental group A (geometric mean (GM) = 511.60; geometric standard deviation (GSD) = 1.49) was significantly higher than those of experimental group B (GM = 379.58; GSD = 1.60) and the control group (GM = 359.05; GSD = 1.32). This finding indicates that the students that applied the adaptive assessment system based on the Elo rating algorithm and learning memory cycle were more motivated to take practice tests than were the students who used the nonadaptive system and the adaptive assessment system based on the Elo rating algorithm alone.

We then explored the number of questions attempted and the average attempts per question for each group. As indicated in Table 4, the number of questions attempted by the groups was similar, but experimental group A had the highest number of average attempts per question. Intergroup differences in both indicators were evaluated through one-way ANOVA. The logarithm of the average number of attempts per question was used to minimize the effect of skewed data. Levene's test indicated no significant differences in the numbers of questions attempted ($F = 2.50$; $p = 0.08 > 0.05$) and in the logarithm of the average attempts per question ($F = 1.57$; $p = 0.20 > 0.05$), indicating that the assumption of homogeneity of variance was satisfied for both indicators. ANOVA revealed no significant intergroup differences in the number of questions attempted ($F = 0.18$; $p = 0.83 > 0.05$), but the difference in the logarithm of the average attempts per question among the groups was significant ($F = 19.64$; $p = 0.00 < 0.05$). As presented in Table 5, the logarithm of average attempts per question of experimental group A (GM = 2.89; GSD = 1.38) was significantly higher than those of experimental group B (GM = 2.07; GSD = 1.39) and the control group (GM = 1.96; GSD = 1.27). Therefore, although the adaptive assessment system based on the Elo rating algorithm and learning memory cycle did not recommend more questions than did the adaptive assessment system based on the Elo rating algorithm alone or the conventional assessment system, it encouraged the students to repeatedly practice questions that they had attempted before.

To explore whether the assessment system used affected reading engagement, the differences in the reading engagement of the groups were examined through one-way ANOVA. Levene's test for the logarithm of the frequency of reading behaviors ($F = 0.67$; $p = 0.51 > 0.05$) and logarithm of the frequency of opening an e-book ($F = 0.00$; $p = 0.99 > 0.05$) revealed that the data for both variables satisfied the assumption of homogeneity of variance. Significant differences in the logarithms of the frequencies of reading behaviors ($F = 7.25$; $p = 0.00 < 0.05$) and opening e-books ($F = 21.85$; $p = 0.00 < 0.05$) were identified. As shown in Table 6, the logarithm of the frequency of reading behaviors of experimental group A (GM = 3648.00; GSD = 1.40) was significantly higher than those of experimental group B (GM = 2815.21; GSD = 1.36) and the control group (GM = 2934.17; GSD = 1.37). Furthermore, experimental group A (GM = 84.34; GSD = 1.45) opened e-books significantly more frequently than did experimental group B (GM = 51.81; GSD = 1.42) and the control group (GM = 53.35; GSD = 1.43), indicating that the proposed system can effectively increase the frequency with which students study or review using e-books (relative to conventional assessment systems), whereas adaptive assessment systems based on the Elo rating algorithm alone cannot increase the frequency. The findings also indicate that the proposed adaptive system is conducive to reading engagement.

The students' average assessment behaviors and average reading behaviors for each concept were evaluated to compare the effects of different assessment systems on assessment and reading engagement. As illustrated in Fig. 4, experimental group A had the highest average numbers of assessment and reading behaviors for every concept. However, the differences were most pronounced for the first three concepts, indicating that the proposed system motivated the students to review the earlier material. Experimental group B and the control group had similar assessment and reading engagement patterns, indicating that the

Table 3
Results of one-way ANOVA: logarithms of the frequencies of total assessment behaviors.

Group	*N*	GM	GSD	F	Post Hoc (*t*-test)
(a) Experimental group A	37	511.60	1.49	9.66***	(a) > (b) (a) > (c)
(b) Experimental group B	37	379.58	1.60		
(c) Control group	34	359.05	1.32		

Table 4
Number of attempted questions and average attempts per question of each group.

Group	Number of questions attempted	Average attempts per question
(a) Experimental group A	94.31	3.04
(b) Experimental group B	95.63	2.19
(c) Control group	94.05	2.02

Table 5
Results of one-way ANOVA: logarithms of average attempts per question.

Group	*N*	GM	GSD	F	Post Hoc (*t*-test)
(a) Experimental group A	37	2.89	1.38	19.64***	(a) > (b) (a) > (c)
(b) Experimental group B	37	2.07	1.39		
(c) Control group	34	1.96	1.27		

Table 6
Results of one-way ANOVA: frequencies of reading behaviors and of open e-books.

Indicator	Group	*N*	GM	GSD	F	Post Hoc (*t*-test)
Logarithm of the frequency of reading behaviors	(a) Experimental group A	37	3648.00	1.40	7.25**	(a) > (b) (a) > (c)
	(b) Experimental group B	37	2815.21	1.36		
	(c) Control group	34	2934.17	1.37		
Logarithm of the frequency of opening an e-book	(a) Experimental group A	37	84.34	1.45	21.85***	(a) > (b) (a) > (c)
	(b) Experimental group B	37	51.81	1.42		
	(c) Control group	34	53.35	1.43		

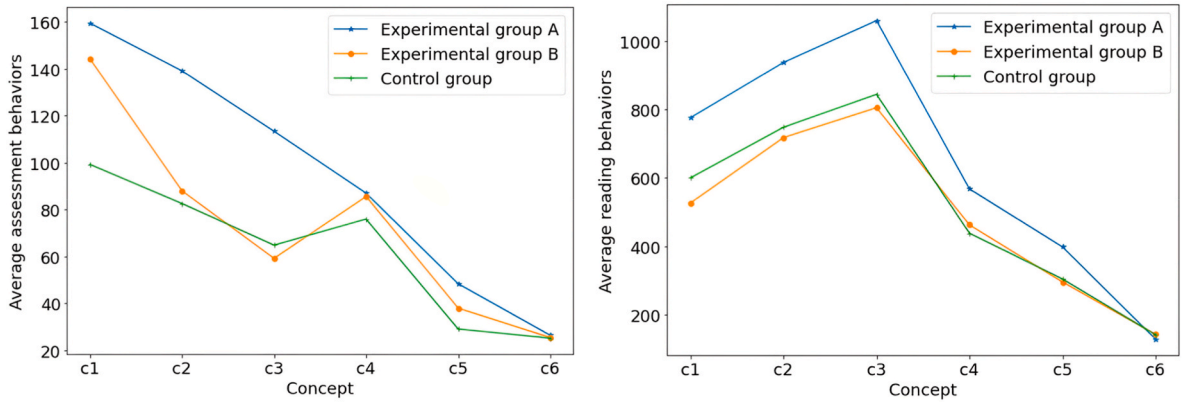


Fig. 4. Average assessment and reading behaviors per concept.

adaptive assessment system based on the Elo rating algorithm alone was unable to increase students' assessment and reading engagement relative to the conventional assessment system.

5.2. Analysis of programming learning achievement

The programming learning achievement of the three groups was evaluated through one-way analysis of covariance (ANCOVA), wherein the covariate was the pretest score, the independent variable was the assessment system used, and the dependent variable was the posttest score. The means and standard deviations of the groups' pretest scores are presented in Table 7. Levene's test revealed homogeneity of variance among the pretest scores of the three groups ($F = 0.39$; $p = 0.67 > 0.05$). One-way ANOVA revealed no significant intergroup differences in pretest scores, indicating comparable levels of prior programming knowledge of the three groups.

Levene's test showed that the assumption of homogeneity of variance among the posttest scores of the groups was satisfied ($F = 2.31$; $p = 0.10 > 0.05$). The test results of the interaction effects between the covariate and the independent variable indicated that the assumption of homogeneity of the regression coefficients within the groups was met ($F = 1.45$; $p = 0.23 > 0.05$). The ANCOVA results revealed a significant intergroup difference ($F = 5.09$; $p = 0.007 < 0.05$) after adjustment of the effect of pretest scores, indicating that the students who used distinct formative assessment systems achieved significantly different posttest scores (Table 8). The adjusted mean posttest score of experimental group A (adjusted mean = 91.33; SD = 7.53) was significantly higher than those of experimental group B (adjusted mean = 85.95; SD = 10.03) and the control group (adjusted mean = 83.94; SD = 13.25), indicating that the students who used the adaptive assessment system based on the Elo rating algorithm and learning memory cycle outperformed those who used the adaptive assessment system based on the Elo rating algorithm alone and those who used the conventional assessment system. Therefore, the proposed system was the most effective in improving the students' learning performance.

We also explored the effects of the different adaptive assessment systems on higher cognitive skills through one-way ANOVA. Levene's test revealed no significant intergroup differences in variance ($F = 0.17$; $p = 0.83 > 0.05$), indicating that the assumption of homogeneity was satisfied. The results of ANOVA of the students' coding exercise scores are presented in Table 9. The coding exercise scores of the groups differed significantly ($F = 3.20$; $p = 0.04 < 0.05$). In the post hoc test, the average score of experimental group A (mean = 76.59; SD = 17.41) was significantly higher than that of the control group (mean = 65.47; SD = 18.89), whereas the mean scores of experimental group A and experimental group B (mean = 71.35; SD = 19.17) did not differ significantly. The findings showed that the students using the adaptive assessment system based on the Elo rating algorithm and learning memory cycle significantly outperformed the students using the conventional assessment system in the coding exercise, but the students using the adaptive assessment system based on the Elo rating algorithm alone did not achieve a significantly higher score than the control group did. This finding indicated that compared with the conventional assessment system, the proposed adaptive assessment system was able to more effectively improve the students' higher cognitive skills, whereas the adaptive assessment system that did not account for memory retention was unable to achieve the same result.

Table 7
Descriptive statistics of pretest scores.

Group	*N*	Mean	SD	F
(a) Experimental group A	37	43.53	20.01	0.20
(b) Experimental group B	37	42.99	20.16	
(c) Control group	34	40.73	18.25	

5.3. Analysis of memory cycles

Finally, we performed a within-group comparison of the data of experimental group A to analyze the relationship between student memory cycles, programing learning achievement, and assessment and reading engagement. The students were divided into long-retention and short-retention groups according to their total memory cycle for all items; the students whose total memory cycles were in the top 20% and bottom 20% were assigned to the long-retention and short-retention groups, respectively. The total memory cycles, posttest scores, coding exercise scores, and frequencies of assessment and reading behaviors of the long-retention and short-retention groups are presented in Tables 10 and 11, respectively. The long-retention group had higher average posttest and coding exercise scores than did the short-retention group (92.85 and 68.57, respectively, vs. 78.57 and 60.00, respectively). The average number of assessment and reading behaviors of the long-retention group was also higher than that of the short-retention group (788.42 and 4440.85 behaviors, respectively, vs. 451.71 and 2967.71 behaviors, respectively). These findings indicated that the students with longer memory cycles were more engaged in practicing and reviewing learning materials and therefore achieved higher scores on the programming posttest and the coding exercise.

6. Discussion

The paper proposed an adaptive assessment system integrating a new update scheme and compared it with previously developed CAT-based and nonadaptive systems. We tested the different assessment systems in the context of a university programming course and explored whether the proposed system is suitable for adaptive formative assessment in

Table 8
Results of ANCOVA: posttest scores.

Group	*N*	Mean	SD	Adjusted mean	F	η^2	Post Hoc (*t*-test)
(a) Experimental group A	37	92.48	7.53	91.33	5.09**	0.08	(a) > (b) (a) > (c)
(b) Experimental group B	37	87.02	10.03	85.95			
(c) Control group	34	84.70	13.25	83.94			

Table 9
Results of one-way ANOVA: coding exercise scores.

Group	*N*	Mean	SD	F	Post Hoc (*t*-test)
(a) Experimental group A	37	76.59	17.41	3.20*	(a) > (c)
(b) Experimental group B	37	71.35	19.17		
(c) Control group	34	65.47	18.89		

Table 10
Total memory cycles, posttest and coding exercise scores, and average assessment and reading behaviors of students with total memory cycles in the top 20%.

Learner No.	Total (days)	Posttest score	Coding exercise score	Frequency of assessment behaviors	Frequency of reading behaviors
1	1347250	95	60	656	4894
2	198728	90	60	791	5384
3	7264	95	60	952	4940
4	751	90	80	677	4292
5	552	100	60	816	5170
6	413	95	60	699	4926
7	391	85	100	928	1480
Average		92.85	68.57	788.42	4440.85

Table 11
Total memory cycles, posttest and coding exercise scores, and average assessment and reading behaviors of students with total memory cycles in the bottom 20%.

Learner No.	Total (days)	Posttest score	Coding exercise score	Frequency of assessment behaviors	Frequency of reading behaviors
8	144	75	80	418	4859
9	133	80	60	302	3077
10	128	55	80	536	2147
11	124	80	60	589	2156
12	122	90	60	580	2940
13	119	70	20	394	2943
14	107	100	60	343	2652
Average		78.57	60.00	451.71	2967.71

practical settings.

RQ: Can a formative assessment system based on computerized adaptive testing and learning memory cycle increase students' engagement in taking practice tests and learning?

To determine whether adaptive features can motivate students to review learned information by taking practice tests and restudying e-books, we evaluated the students' engagement by measuring their assessment and reading behaviors. Our results provide a positive answer to the first research question. The total number of assessment behaviors and the average attempts per question of the students who used the proposed system were significantly higher than those of the students who used the adaptive assessment system based on the Elo rating system alone and those who used the nonadaptive assessment system. Most existing models for adaptive assessment are based on learners' knowledge and item difficulty (Jia & Le, 2020; Pelánek, 2016). Although these models have been successfully deployed in many assessment systems (Klinkenberg et al., 2011; Nižnan et al., 2015), most of these systems have been used for one-time evaluation. When these systems are used for repeated practice, students' knowledge and the difficulty of specific items stabilize as students repeatedly attempt practice tests. Consequently, the systems may frequently recommend the same items in new practice sessions, which may lead to boredom among students. Therefore, taking more personal information into account to minimize the repetition of questions in each practice session is crucial when designing adaptive assessment systems for review. We achieved this goal by developing a system that considers a student's retention of each item, their knowledge, and item difficulty when selecting questions. This approach enables the system to recommend questions about later concepts that students are not yet familiar with and earlier concepts that they may have forgotten, which prevents students from feeling bored and, in turn, makes them more willing to use the system. We determined that the proposed adaptive assessment system was more effective in promoting the students to engage in reading e-books than were the other two systems tested. Because the assessment system only displays the answer results and does not provide students with correct answers, students must refer to e-books and may even choose to restudy e-books before taking the practice tests or use BookRoll's annotation tools to avoid searching for answers during the tests.

In college courses, students' motivation to review concepts from earlier classes in the course may decrease as the course progresses and the information they have learned accumulates. Therefore, adaptive features that can select the questions to which students are most likely to have forgotten the answers to are necessary, thus motivating students to review earlier concepts. By analyzing the students' total numbers of assessment and reading behaviors for each concept, we discovered that the students who used the proposed adaptive assessment system tended to review earlier topics more frequently than did the students in the other two groups. Students who use nonadaptive assessment systems might feel comfortable taking practice tests covering newly learned topics because the number of questions in each test is limited; however, they may be reluctant to take practice tests regarding earlier concepts because they may not want to review all the previous questions. Although the adaptive assessment system based on knowledge and item difficulty alone can select questions that fit students' knowledge levels, the practice tests generated by the system may contain repeated questions after a student completes several attempts, requiring students to use the nonadaptive system to access additional questions. On the other hand, the proposed system selects questions from earlier concepts that might be forgotten by the students, which saves the time for students looking to review earlier information.

RQ: Can a formative assessment system based on computerized adaptive testing and learning memory cycle improve students' learning performance?

To answer the second research question, we measured the students' learning performance and programming skills after they used the proposed adaptive assessment system. The results indicated that the students who used the proposed system outperformed the students who used the nonadaptive assessment system on the posttest and the coding exercise, which is consistent with the findings of Chen and Chung (2008) that highlighted the importance of memory in the design of personalized and adaptive learning systems. In real educational contexts, students' engagement in learning activities usually does not remain consistent throughout the course because of fatigue or insufficient time. Without considering memory changes, adaptive assessment systems may fail to motivate students to repeatedly reassess their knowledge, thereby

failing to improve their learning performance. The adaptive assessment system in the present study was designed to make the review process more interesting and efficient to enhance students' engagement. In our study, the students who used the proposed system attempted each question more frequently than did those who used the nonadaptive system. Previous studies have reported the benefits of repeated retrieval over single retrieval and of restudying materials (Roediger & Butler, 2011). The proposed system motivated the students to repeatedly take assessments and attempt questions multiple times, which enhanced their learning performance. Another factor that contributes to learning performance is spacing. Spaced learning is a technique that requires certain intervals between repeated reviews of learned information to facilitate the storage of knowledge in long-term memory (Ebbinghaus, 2013). In regular classes, students are often able to review previously learned information while learning new information. A memory trace is formed when a student learns information and is reactivated when a student gains additional new knowledge. By incorporating multiple concepts into review activities, memory traces for both old and new information can be reactivated, thereby helping the student recall the information more easily in the future. These memory traces also help students remember information when learning new lessons (Nakata & Elgort, 2021). The results presented in section 4.3 indicated that the students with longer memory retention of the learned concepts were more engaged in assessment and reading behaviors and had higher posttest and coding exercise scores.

Although adaptive assessment and spacing have been demonstrated to positively affect learning, most relevant studies have evaluated these effects only in fields that require memorization (e.g., language learning; Chen & Chung, 2008). This study evaluated the students' programming skills based on their coding exercise scores. Programming is a higher cognitive skill than rote memorization; students must comprehend specific concepts and learn how to apply them in combination to complete programming exercises. Our results indicated that the proposed system was effective for improving students' programming skills. Although adaptive assessment systems can only help strengthen students' memory retention of fundamental knowledge, such systems may still affect students' performance on coding exercises. For example, if a coding exercise requires students to use operators, if/else conditional statements, and loops simultaneously, students must understand and memorize these concepts to complete the exercise.

7. Implications and future research

The proposed system can be helpful for instructors who wish to employ repeated assessment activities but are concerned about low engagement due to the limited number of questions. The instructors can let students take an assessment before learning new content to ensure they equip the prerequisite knowledge. The estimated latent ability enables the instructors to identify at-risk students and provide personalized suggestions. The item difficulty can help instructors find the concepts that students struggled with and adjust the teaching approaches. For students, the adaptive assessment systems prevent them from having to review all the questions related to earlier concepts, which facilitates the review process and promotes learning engagement. The system also helps students improve retention by employing effective memory practices, such as spacing and interleaving. Our results provide insights for researchers trying to develop adaptive assessment tools for review purposes.

Future studies can consider other personalized information when developing adaptive features for assessment systems. For example, students' creativity, self-regulated learning skills, and self-efficacy can be used to adaptively generate assessments in which question types (e.g., multiple choice, fill-in-the-blank, or short-answer), assessment length, and question difficulty are all tailored to each student. This information can be acquired through questionnaires or inferred by analyzing learning log data through learning analytics approaches (Kosinski et al., 2013). Although the proposed system cannot be used to evaluate higher cognitive skills directly, it can help students memorize key concepts, thereby improving their programming skills. Future researchers can consider developing an adaptive assessment system that can select coding exercises for examining students' programming skills. Such a system may require a more complex adaptive model because coding exercises often require the integration of multiple concepts.

8. Conclusions

The main contribution of the current study is that conventional CAT models may be insufficient for formative assessment in college courses, specifically for encouraging students to review previously learned material through practice tests and for improving students' learning performance. Therefore, students' personal information, such as memory retention, should be considered in the design of future adaptive formative assessment systems.

Although valuable findings were obtained from the experimental results, our study has some limitations. First, the short interval between the introduction of the final concept and the posttest resulted in no significant intergroup differences in assessment and reading engagement for the last few concepts. We hypothesized that the difference between learning engagement and learning performance would become more pronounced as the interval between the last concept and the posttest increased. Second, few questions were included in the coding exercise, and the weight of the score of each question on the students' final scores may have been excessive. The differences among the students' final scores may have been too small or too large, which may have affected the analysis results. Additional exercises containing more questions are required to further test the degree to which the proposed system can benefit students' programming skills. Finally, the sample size was relatively small, and the proposed system was deployed only to students in a university programming course. Therefore, exploration of the application of the system in other types of courses with larger sample sizes is required to evaluate the effectiveness of the proposed system.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) 20H01722, JSPS Grant-in-Aid for Scientific Research (S) 16H06304, NEDO JPNP20006 and JPNP18013, and JST SPRING JPMJSP2110.

This work was approved by the Research Ethics Committee of National Taiwan University (case number: 202005ES032).

References

Bayne, S. (2015). Teacherbot: Interventions in automated teaching. *Teaching in Higher Education, 20*(4), 455–467. https://doi.org/10.1080/13562517.2015.1020783

Chen, C. M., & Chung, C. J. (2008). Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education, 51*(2), 624–645.

Chen, X., Zou, D., Cheng, G., & Xie, H. (2021). Artificial intelligence-assisted personalized language learning: Systematic review and co-citation analysis. In *2021 International Conference on Advanced Learning Technologies (ICALT)* (pp. 241–245). IEEE.

Chen, X., Zou, D., Xie, H., Cheng, G., & Liu, C. (2022). Two decades of artificial intelligence in education: Contributors, collaborations, research topics, challenges, and future directions. *Educational Technology & Society, 25*(1), 28–47.

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences, 10*(22), 8196.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805*.

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

Ebbinghaus, H. (2013). Memory: A contribution to experimental psychology. *Annals of Neurosciences, 20*(4), 155.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5–18.

Elmahdi, I., Al-Hattami, A., & Fawzi, H. (2018). Using technology for formative assessment to improve students' learning. *Turkish Online Journal of Educational Technology-TOJET, 17*(2), 182–188.

Elo, A. E. (1978). *The rating of chessplayers, past and present* (Vol. 3). London: Batsford.

Flanagan, B., & Ogata, H. (2018). Learning analytics infrastructure for seamless learning. In *Proceedings of the 8th International Conference on Learning Analytics & Knowledge (LAK18)*. Sydney, Australia: Association for Computing Machinery (ACM). http://hdl.handle.net/2433/233071.

Gurung, R. A., Daniel, D. B., & Landrum, R. E. (2012). A multisite study of learning in introductory psychology courses. *Teaching of Psychology, 39*(3), 170–175.

Hvattum, L. M., & Arntzen, H. (2010). Using Elo ratings for match result prediction in association football. *International Journal of Forecasting, 26*(3), 460–470.

Jia, J., & Le, H. (2020). The design and implementation of a computerized adaptive testing system for school mathematics based on item response theory. In *International Conference on Technology in Education* (pp. 100–111). *Sin*gapore: Springer.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17*(4), 471–479.

van Kesteren, M. T. R., & Meeter, M. (2020). How to optimize knowledge construction in the brain. *Npj Science of Learning, 5*(1), 1–7.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. (2011). Computer adaptive practice of maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*(2), 1813–1824.

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802–5805.

Liu, M., Kitto, K., & Shum, S. B. (2021). Combining factor analysis with writing analytics for the formative assessment of written reflection. *Computers in Human Behavior, 120*, Article 106733.

McCallum, S., & Milner, M. M. (2021). The effectiveness of formative assessment: Student views and staff reflections. *Assessment & Evaluation in Higher Education, 46*(1), 1–16.

Meiser, T., Plieninger, H., & Henninger, M. (2019). IRTree models with ordinal and multidimensional decision nodes for response styles and trait-based rating responses. *British Journal of Mathematical and Statistical Psychology, 72*(3), 501–516.

Mihalcea, R., & Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 404–411). *Barcelona, S*pain: Association for Computational Linguistics.

Nakata, T., & Elgort, I. (2021). Effects of spacing on contextual vocabulary learning: Spacing facilitates the acquisition of explicit, but not tacit, vocabulary knowledge. *Second Language Research, 37*(2), 233–260.

Nižnan, J., Pelánek, R., & Rihák, J. (2015). *Student models for prior knowledge estimation*. International Educational Data Mining Society.

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). China: Asia-Pacific Society for Computers in Education https://kyushu-u.pure.elsevier.com/en/publications/e-book-based-learning-analytics-in-university-education.

Pai, K. C., Kuo, B. C., Liao, C. H., & Liu, Y. M. (2021). An application of Chinese dialogue-based intelligent tutoring system in remedial instruction for mathematics learning. *Educational Psychology, 41*(2), 137–152.

Papousek, J., Pelánek, R., & Stanislav, V. (2014). Adaptive practice of facts in domains with varied prior knowledge. In *Educational Data Mining 2014*.

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education, 98*, 169–179.

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*(1), 25–36.

Rodrigues, F., & Oliveira, P. (2014). A system for formative assessment and monitoring of students' progress. *Computers & Education, 76*, 30–41.

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education, 26*(2), 582–599. https://doi.org/10.1007/s40593-016-0110-3

Tang, K. Y., Chang, C. Y., & Hwang, G. J. (2021). Trends in artificial intelligence-supported e-learning: A systematic review and co-citation network analysis (1998–2019). *Interactive Learning Environments*, 1–19.

Tomasik, M. J., Berger, S., & Moser, U. (2018). On the development of a computer-based tool for formative student assessment: Epistemological, methodological, and practical issues. *Frontiers in Psychology, 2245*.

Wang, W., Song, L., Wang, T., Gao, P., & Xiong, J. (2020). A note on the relationship of the Shannon entropy procedure and the Jensen–Shannon divergence in cognitive diagnostic computerized adaptive testing. *Sage Open, 10*(1), Article 2158244019899046.

Wauters, K., Desmet, P., & Van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education, 58*(4), 1183–1193.

Xiao, M., & Yi, H. (2021). Building an efficient artificial intelligence model for personalized training in colleges and universities. *Computer Applications in Engineering Education, 29*(2), 350–358.

Yang, A. C. M., Chen, I. Y. L., Flanagan, B., & Ogata, H. (2021). Automatic generation of cloze items for repeated testing to improve reading comprehension. *Educational Technology & Society, 24*(3), 147–158. https://www.jstor.org/stable/27032862?seq=1#metadata_info_tab_contents.

Zhou, Y., Li, X., Cao, Y., Zhao, X., Ye, Q., & Lv, J. (2021). *LANA: Towards personalized deep knowledge tracing through distinguishable interactive sequences. arXiv preprint arXiv:2105.06266*.