Automatic Generation of Cloze Items for Repeated Testing to Improve Reading Comprehension

Albert C. M. Yang[1*], Irene Y. L. Chen[2], Brendan Flanagan[3] and Hiroaki Ogata[3]
[1]Graduate School of Informatics, Kyoto University, Japan // [2]Department of Accounting, National Changhua University of Education, Taiwan // [3]Academic Center for Computing and Media Studies, Kyoto University, Japan // yang.ming.35e@st.kyoto-u.ac.jp // irene@cc.ncue.edu.tw // flanagan.brendanjohn.4n@kyoto-u.ac.jp // hiroaki.ogata@gmail.com
[*]Corresponding author

ABSTRACT: Reviewing learned knowledge is critical in the learning process. Testing the learning content instead of restudying, which is known as the testing effect, has been demonstrated to be an effective review strategy. However, education research recommends that instructors generate practice tests, but this burdens teachers and may also hinder teaching quality. To resolve this issue, the current study applied a modern artificial intelligence technique (BERT) to automate the generation of tests and evaluate the testing effect through e-books in a university lecture (N = 74). The last 5 minutes of each course session were utilized to review the taught content by having students either answer cloze item questions or restudy the summary of the core concepts covered in the lecture. A reading comprehension pretest was conducted before the experiment to ensure that the differences in prior knowledge were nonsignificant between groups, and a posttest was performed to examine the effectiveness of testing. In addition, we evaluated students' reading skills and reading engagement through their ability to identify key concepts and their interaction with e-books, respectively. A positive effect was observed for students who engaged in cloze item practice before the end of each class. The results indicated that the repeated testing group exhibited significantly better reading skills and engaged more with e-books than the restudying group did. More importantly, compared with only restudying the key concepts, answering the cloze items questions significantly improved students' reading comprehension. Our results suggest that machine-generated cloze testing may benefit learning in higher education.

Keywords: Modern AI, Repeated testing, Testing effect, Test-enhanced learning

1. Introduction

Artificial intelligence (AI) refers to "computers that mimic cognitive functions that humans associate with the human mind, such as learning and problem-solving" (Russell & Norvig, 2005, p. 2). With the increasing development of information technologies, AI has been extensively applied in the area of education. For example, Junco and Clem (2015) applied a hierarchical linear regression model to predict the course GPA of students on the basis of their reading engagement. Süzen et al. (2020) combined data mining techniques and clustering to automatically grade short-answer assignments and provide feedback to students. Recently, modern AI has been generally referred to as deep neural network (DNN)–based approaches (Yosinski et al., 2014), and these have been applied in academic fields. Zhang et al. (2019) applied a long short-term memory neural network to build a model that could learn word sequence information, thus enabling it to automatically grade semi-open-ended questions. Furthermore, Okubo et al. (2017) proposed a recurrent neural network (RNN) model to predict the course grade of students using the log data collected by learning management systems. Their results indicated that RNN outperformed other regression models in the prediction tasks.

Repeated testing has been demonstrated to be effective for improving both short-term and long-term memory (Wiklund-Hörnqvist et al., 2014). Although the majority of the positive effects of repeated testing have been identified in laboratory settings (Karpicke, 2017; Rowland, 2014), researchers and practitioners have recently started implementing testing in educational contexts. Greving and Richter (2018) had college students review lecture content 10 minutes before the end of each class and determined that students who reviewed the content by answering short-answer questions performed better than those who answered multiple-choice questions or restudied the summaries of the lecture content in a later retention test. However, the testing questions in a majority of previous studies were created by humans, and creating a practice test for all learning materials is resource intensive. This is typically the case in colleges because many instructors choose to organize their materials on their own instead of using existing textbooks. To address this issue, Mouri et al. (2019) utilized the digital textbook logs of students to automatically generate a personalized quiz for the purpose of reviewing. Olney et al. (2017) applied natural language processing (NLP) techniques to generate cloze item practice tests, and they found the effectiveness of machine-generated and human-generated tests to be comparable. In the domain of modern AI, researchers have begun applying modern AI–based techniques to automatically generate

questions using sentences from texts. Du et al. (2017) introduced an attention-based, sequence-to-sequence model for this task, and the results suggested that their model significantly outperformed the state-of-the-art rule-based system. Moreover, Chan and Fan proposed (2019) a recurrent BERT-based model to perform the task of short-answer question generation. Their model resolved the shortcomings of directly using BERT for text generation. However, the majority of previous studies that applied modern AI techniques were focusing on short-answer question generation. Drawing from those studies, we developed a BERT-based system to automatically generate cloze items for practice and examined whether cloze item practice generated by modern AI techniques produces testing effect and whether it has a positive impact on reading comprehension. Furthermore, we collected students' reading logs to evaluate their reading skills and reading engagement. Our hypothesis was that students' reading skills and reading engagement are improved through repeated testing.

2. Literature review

2.1. Reading skills

Reading skills refer to the ability to understand and recall reading content (Memory, 1983). High-skill readers tend to apply different strategies to extract relevant information from the target content and to better recall learned knowledge during the review stage. This phenomenon occurs more frequently in college because college textbooks often contain longer and more difficult sentences; for many students, such reading demands considerable attention to fully understand the content. Therefore, students with high-level reading skill are expected to perform better than those with weaker skills. Furthermore, a previous study demonstrated that high-skill readers are more likely to comprehend learning content than low-skill readers are (Lorch & Pugzles-Lorch, 1985); thus, low-skill readers seem to face difficulties in identifying the most relevant information in the texts that they read. In support of this claim, other researchers have observed that high-skill and low-skill readers differ in terms of the concepts they perceive to be important in a text (Winograd, 1984). Furthermore, Coiro (2011) indicated that differences in prior knowledge can even be compensated for by adolescents with high reading skills when they are learning with others with prior knowledge.

Text marking is a common and effective reading skill. By highlighting or underlining the most relevant information in a text, students can separate the identified valuable information from other irrelevant content and can easily recall key information during later review. Research has indicated that students who used the highlighting feature in digital textbooks achieved better academic outcomes (Junco & Clem, 2015). However, without considering the content of marked text, students might overuse this skill by simply marking more text. Bell and Limber (2009) indicated that text-marking skills represent students' ability to identify and isolate key information and found that low-skill readers tend to highlight more than high-skill readers do because of their inability to identify relevant concepts. That is, the highlight frequency and reading skills are positively correlated only up to a certain extent—when students are unable to distinguish between critical and trivial textbook content, they may overuse the highlighting strategy. Therefore, the measurement of students' text-marking ability in this study was measured by the content of text they marked, instead of the number of highlights they added. Furthermore, Yue et al. (2015) proposed that the effectiveness of highlighting can be optimized when students are trained on how to use this skill. Therefore, we want to investigate whether students' reading skills can be enhanced by taking practice tests since the questions in tests are the key concepts in materials. We measured students' text-marking ability in e-books to evaluate their reading skills in this study.

2.2. Reading engagement

Reading has been shown to directly correlate with course outcomes (Daniel & Woody, 2013). Landrum, Gurung, and Spann (2012) observed that students' self-reported percentage of completed readings in textbooks strongly related to their quiz scores and final grades. Junco and Clem (2015) collected students' engagement index to predict their course outcomes. They found that the time spent on reading was the most significant factor in their prediction model. In addition, reading engagement was found to vary for different texts, with more advanced lectures requiring more reading time (Fitzpatrick & McConnell, 2009). Studies have highlighted that although many students may not read a complete text, they do engage with the interactive features in digital textbooks, and such engagement improves their learning outcomes (Berry et al., 2010; Dennis, 2011; Fouh et al., 2014). Dennis (2011) discovered that the number of annotations was positively related to learning outcomes, whereas the number of pages students read was not, which seems to contradict the finding of Junco and Clem (2015). This suggests that it might not be enough to measure students' reading engagement solely by reading time or the number of pages read; instead, annotation tools, including notes or highlights, allow student to interact with the

text and, in turn, reflect the effort they make during reading, should be considered as well. Therefore, textbook analytics can be applied to measure reading engagement with e-books, and this indicator can be employed to predict students' learning outcomes (Bossaller & Kammer, 2014). In support of this claim, research has demonstrated that students who read more or interact more with their textbooks perform better in class (Dawson, McWilliam & Tan, 2008; DeBerard, Speilmans, & Julka, 2004; Woody et al., 2010). In sum, improving students' reading engagement not only motivates them to interact with the text but also improves their learning performance. Testing has been shown to improve students' learning engagement as they need to spend more time on reading textbooks and readjust their learning strategies in order to answer the questions (Soderstrom & Bjork, 2014). In this study, we measured students' reading engagement by both reading time and the number of annotation tools they used and hypothesized that students' reading engagement with digital textbooks increases after appearing for cloze test practice.

2.3. Repeated testing

Traditionally, testing is used to assess students' knowledge and assign grades. However, its employment to facilitate learning is an application of testing that has been largely neglected by educationalists (Butler & Roediger, 2007). Empirical studies have emphasized that compared with traditional restudying of learning materials, taking repeated tests greatly improves students' performance in later recall tests (Butler & Roediger, 2007; McDaniel et al., 2007). One explanation for this effect is that repeated testing forces student to reencode the information they have learned, whereas restudying requires them to only reproduce the encoding of the learned knowledge (Karpicke & Roediger, 2008). The superiority of repeated testing over restudying learning material is known as the testing effect (Roediger & Karpicke, 2006a). Compared with simply rereading the learning material, students subjected to quizzes after reading a chapter of a textbook or upon completion of a course exhibited improved long-term retention of knowledge. This phenomenon is known as the direct testing effect. The indirect testing effect refers to the use of improved strategies or increased motivation to study in anticipation of taking a test. Soderstrom and Bjork (2014)'s results revealed that practice testing motivated participants to readjust their monitoring process and therefore enhanced their learning engagement. Recently, studies on testing effects have gradually shifted from laboratory settings to real classrooms. Bobby et al. (2018) reported that the testing effect of a closed book examination combined with feedback was effective in improving the learning performances for medical students studying biochemistry. Schwieren et al. (2017) conducted a meta-analysis of testing effect and identified a significant overall effect size of $d = 0.56$, highlighting that testing was beneficial to the learning outcomes of psychology students. The number of tests a student can take during the practice phase is a key aspect of the testing effect. Repeated testing has been demonstrated to improve retention as opposed to a singular test (Karpicke & Roediger, 2008). Moreover, the effects of repeated testing are more pronounced when tests are administered over time (Karpicke & Roediger, 2007). Another crucial aspect is the provision of feedback. Feedback enhances the benefit of testing through the correction of errors and confirmation of correct answers (Butler & Roediger, 2008). Studies have demonstrated that feedback can dramatically amplify the knowledge retention achieved through repeated testing (Butler et al., 2008). Generally, testing effects are larger for more difficult tests because they require more cognitive effort for information retrieval (Kang et al., 2007). However, raising the difficulty level of tests may lead to increased unsuccessful retrieval. According to one study, retrieval must be successful to reap the benefits of repeated testing (Rowland, 2014). Therefore, feedback can be useful for overcoming the limited effect of unsuccessful retrieval by correcting incorrect responses (Rowland, 2014). Wiklund-Hörnqvist et al. (2014) demonstrated that compared with short- and long-term restudying, repeated testing with feedback significantly promoted learning. Furthermore, they emphasized the importance of educationalists adopting teaching methods that involve repeated testing. With the advancement of information technology, researchers have started applying AI in repeated testing by automatically generating practice tests. For example, Olney et al. (2017) applied NLP techniques to automatically generate cloze items and found machine-generated items to be as effective as human-generated ones for enhancing reading comprehension. In this study, we hypothesized that the direct testing effect will promote student retention of learned knowledge and therefore achieve better scores in the reading comprehension posttest. In addition, we hypothesized that students' reading engagement and reading skills will be enhanced by readjusting their reading behaviors after practice testing. We leveraged modern AI techniques to automatically generate cloze item practice for repeated testing and addressed the following research questions:

(1) Can students improve their reading skills with machine-generated cloze item practice?
(2) Can students improve their reading engagement with machine-generated cloze item practice?
(3) Can students improve their reading comprehension with machine-generated cloze item practice?

3. Methods

3.1. Research context

A 4-week experiment was conducted in two mandatory courses for undergraduate students from the accounting department at a university in Taiwan. These courses could be taken as elective courses by students from other departments as well. Both classes were taught by the same instructor using the same materials. A total of 74 students enrolled in this experiment. Both courses employed BookRoll, an e-book reading system (Flanagan & Ogata, 2017) developed by Kyoto University; instructors can upload materials, and students can use the e-book reader to read the content and interact with the text using the provided tools, such as notes and highlights. The actions performed by students are stored in the database for later analysis. The e-book reading actions available in BookRoll have been described in detail by Ogata et al. (2015) and Flanagan and Ogata (2018). Participants took a reading comprehension pretest and posttest during the first and the final week of the experiment that evaluated whether the use of cloze item practice promoted their reading proficiency. The reading comprehension pretest and posttest each comprised 28 multiple-choice questions that had been randomly extracted from a test bank with 50 questions related to the accounting field. The test bank had been created by two instructors at the department with accounting experience.

3.2. Procedure

During the experiment, one class was assigned to be the experimental group and the other constituted the control group. In the first week of the experiment, students were asked to complete a reading comprehension pretest. The instructor uploaded the materials a week before each class. Students were required to review the materials and mark the sentences or words that they thought were important. Their marking scores were calculated according to the content they had marked, which was considered to be a reflection of their reading skills. Moreover, the actions students performed during their reading were examined to assess their reading engagement. The measurement of reading skills and reading engagement is explained in the following section. The instructor briefly discussed the content of the materials shared and answered students' questions during the class. Students in the experimental group were required to take a close test practice at the end of each class, whereas the control group students restudied the key concepts in the learning materials summarized by our system. To investigate whether different review methods affect learning, the questions in cloze item practice for experimental group and the key concepts for control group consisted of the same sentences extracted from the materials, except that one or two words in each sentence were masked for the questions, whereas the original sentences were presented in key concepts. The experimental group students could take the test and practice (the number of correct answers was not counted in their final course grade) repeatedly. The experimental group students were encouraged to test themselves after class, and the control group were encouraged to restudy the key concepts as well. Finally, both groups took a reading comprehension posttest in the last week of the experiment, and the results were used to evaluate the effectiveness of cloze item practice. The questions in the reading comprehension test were different from the questions in cloze item practice and key concepts presented to students.

3.3. Automatic cloze item generation

We applied the advanced neural network technique BERT and the machine learning model TextRank to generate cloze items in this study. BERT is a pretrained model that was developed by Google for NLP. During the pretraining phase, BERT develops bidirectional representations from a plain text corpus by taking into account the context of each occurrence of a given word. Thus, unlike other word-embedding models such as Word2vec or GloVe that create a single-word embedding for each word, BERT generates a contextualized embedding representation that varies depending on the sentence. As a result, the pretrained model can be fine-tuned by simply introducing an additional layer to create a specific model for various tasks such as question answering and language inference. TextRank is an unsupervised machine learning algorithm based on PageRank, which is often used for keyword extraction and text summarization. TextRank constructs a graph denoting the relationships between the words in a text and ranks the items in the graph. This method allows TextRank to generate summaries without a training corpus or labeling and makes it appropriate for application in various language tasks. In this study, the open-source transformers packages developed by Hugging Face and TextRank4ZH were adopted to implement BERT and TextRank, respectively.

In our study, the generation of cloze items involved two steps: key sentence extraction and keyword extraction. First, we split the text into sentences and applied BERT to generate the embedding of the full text and the

embedding of each sentence. The cosine distance between the embedding of the text and the embedding of each sentence was calculated, and the sentences that were close to the text in the vector space were selected as the core concepts. The selected sentences were provided to the control group students to review. Second, TextRank was applied to extract keywords from each selected sentence. Subsequently, words with the highest weight were masked as cloze items for the experimental group. Figure 1 shows a snapshot of cloze item practice. When students enter the module, they need to choose the e-Book they want to review. The class name, e-Book, and student ID will be displayed. Students are aware of the total number of questions and the number of questions they have completed. When students click a mask, an input field will be displayed. Then, students need to enter and submit their answer. They are not required to answer the questions in order. For example, they can jump between the pages to answer the questions they are familiar with first, or skip the questions that they already know the answer. After completing the practice testing, they close the module to leave the system. All students' behaviors during testing will be recorded in the database for future analysis.

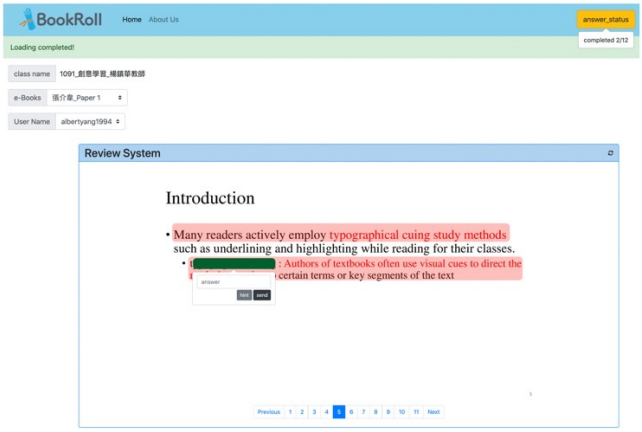

Figure 1. A snapshot of activities in testing module

3.4. Measurement of reading skills and reading engagement

According to Bell and Limber (2009), text-marking skills indicate a reader's ability to identify relevant information in a text, and only high-skill readers are able to achieve this task. Hence, we utilized the sentences generated by BERT during the creation of cloze items as essential information in the text, after which we calculated the similarity between those sentences and the content marked by students using Bilingual Evaluation Understudy (BLEU; Papineni et al., 2002). The BLEU score was subsequently employed as the marking score to represent reading skills. The score ranged from 0.0 to 1.0 and was calculated every week. A higher score denoted better reading skills. We used students' reading actions on BookRoll to assess their reading engagement. The actions included their reading time (25%), the number of highlights made (25%), the number of memos posted (25%), and the number of bookmarks added (25%). All feature values were standardized, and the score of their reading engagement was calculated by the sum of the weighted feature values. For example, if student A's standardized score of reading time is 70, standardized score at making highlights is 80, standardized score of posting memos is 60, and the standardized score of adding bookmarks is 60, the reading engagement score of student A is 70 * 0.25 + 80 * 0.25 + 60 * 0.25 + 60 * 0.25 = 67.5. Therefore, more actions on BookRoll indicated higher reading engagement.

4. Results

4.1. Analysis of Reading Skills and Reading Engagement

An independent t test was performed to evaluate the influence of cloze item practice on reading skills. The results of the Levene test were not significant ($F = 0.28$, $p = .59$), indicating that variance homogeneity existed between the groups. As presented in Table 1, the experimental group exhibited a significantly higher marking score than the control group did ($t = 2.70$, $p < .01$). The mean and standard deviations for the experimental group were 66.34 and 11.21, respectively, and those for the control group were 59.52 and 10.51, respectively. These results suggested that students' reading skills improved after the administration of cloze item practice.

Table 1. Independent *t*-test result of the marking scores of two groups

Dimension	Group	*N*	Mean	*SD*	*t*
Marking score	Experimental group	36	66.34	11.21	2.70**
	Control group	38	59.52	10.51	

Note. $^{**}p < .01$.

Subsequently, we measured the differences in reading engagement between the groups. The Levene test for determining the homogeneity of variance showed no violations ($F = 0.00$, $p = .92$), indicating that the assumption was tenable and that the independent *t* test could be used to interpret the relationship between the application of cloze item practice and reading engagement. Table 2 shows that the experimental group exhibited a significantly higher reading engagement than the control group did ($t = 2.34$, $p < .05$). The mean and standard deviations of the experimental group and control group were 75.77 and 11.59 and 69.05 and 13.00, respectively. This indicated that students demonstrated more reading engagement with their e-books after the use of cloze item practice. Furthermore, the independent *t* test was performed again to compare the reading time of two groups outside the class. The Leven test results indicated the homogeneity of variance existed in two groups ($F = 0.02$, $p = .88$). The independent *t* test results showed that experimental group had a significantly higher reading time outside the class than the control group had ($t = 2.28$, $p < .05$; Table 2), meaning that students spent more time on reading after class in order to pass the practice testing. The mean and standard deviations for the experimental group were 455.19 and 370.55, respectively, and those for the control group were 250.00 and 401.79, respectively.

Table 2. Independent *t*-test results of the reading engagement and the reading time outside the class of both groups

Dimension	Group	*N*	Mean	*SD*	*t*
Reading engagement	Experimental group	36	75.77	11.59	2.34*
	Control group	38	69.05	13.00	
Reading time outside class (minutes)	Experimental group	36	455.19	370.55	2.28*
	Control group	38	250.00	401.79	

Note. $^{*}p < .05$.

4.2. Analysis of reading comprehension

After obtaining the pretest and posttest results concerning reading comprehension, we analyzed the mean and standard deviation of the data and used the Python package Pingouin to conduct a one-way analysis of covariance (ANCOVA), where the covariate was the pretest score, the independent variable was the use of cloze item practice, and the dependent variable was the posttest score. The mean and standard deviations of the posttest scores of both groups are presented in Table 3. The pretest and posttest each comprised 28 multiple-choice questions. A total of 28 points could be scored on each test. The *t* test outcome of the pretest was $t = 1.31$, $p = 0.19$. This indicated that no significant discrepancy existed between the prior knowledge of both groups.

Table 3. Pretest and posttest scores for reading comprehension under different review conditions

	Control group		Experimental group	
Factors	Mean	*SD*	Mean	*SD*
Pretest score				
Reading comprehension	23.68	1.49	24.16	1.65
Posttest score				
Reading comprehension	23.86	1.29	25.50	0.79

One-way ANCOVA was performed to verify whether the between-group differences in the reading comprehension results of the pretest and posttest were statistically significant. Regression coefficients revealed no significant interaction between the covariates and independent variables ($F = 0.78$, $p = .68$); hence, the regression coefficients within the groups did not violate the assumption of homogeneity. Likewise, the results of the Levene test were not significant ($F = 3.25$, $p = .07$). This indicated that homogeneity of variance existed between the groups and that one-way ANCOVA could be conducted to explore any significant differences in the reading comprehension posttest scores of the two groups. The mean of the posttest scores between students in the experimental group (Mean = 25.50, $SD = 0.79$, Adjusted mean = 25.45) and control group (Mean = 23.86, $SD = 1.29$, Adjusted mean = 23.90) was significantly different ($F = 38.83$, $p < .001$, $\eta^2 = 0.34$; Table 4). This finding suggested that students who repeatedly tested themselves showed largely improved reading comprehension compared with those who restudied the materials. Moreover, an independent *t* test was employed to measure the

within-subject difference in the posttest scores. Students in the experimental group exhibited significantly improved performance (t = 4.35, p < .001), whereas students who restudied the materials failed to exhibit a significant improvement in their posttest scores compared with their pretest scores (t = −0.86, p = 0.39; Figure 2). Figure 3 presents that both low-skill readers and high-skill readers of the experimental group achieved a better performance in their posttest.

Table 4. Posttest scores for reading comprehension under different review conditions

Source of variance	*SS*	*df*	*F*	η^2
Covariates	4.85	1	4.33*	0.03
Intergroup	43.47	1	38.83***	0.34
Residual	79.48	71		

Note. $^{*}p < .01$; $^{***}p < .001$.

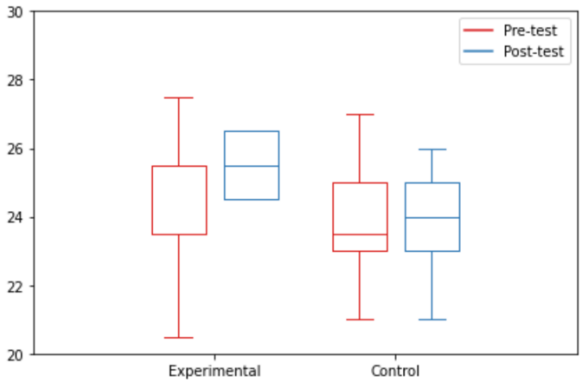

Figure 2. Within-subject differences in the pretest and posttest scores of the experimental group (repeated testing) and the control group (restudying)

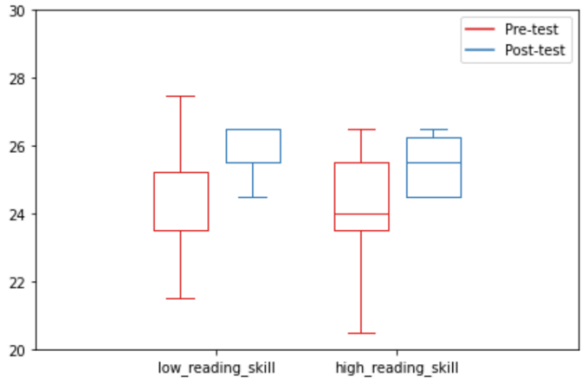

Figure 3. Difference in the pretest and posttest scores of the low-skill readers and high-skill readers in experimental group

5. Discussion

5.1. Differences in reading skills and reading engagement

5.1.1. Research Question 1: Can students improve their reading skills with machine-generated cloze item practice?

The first question addressed by this study was whether repeated testing had a positive effect on students' reading skills. We assumed that if the students were directly shown and tested on the key concepts of a text, they would better understand the core material of the class and their reading skills could be improved. Our results revealed that students who took the review test achieved a significantly better marking score than those who restudied the materials, indicating that they were superior at finding the key concepts in a text. Nist and Simpson (1988) and Yue et al. (2015) have contended that the effectiveness of marking or underlining can be optimized when students are trained on how to use these skills. Because the test questions generated by our system included key sentences and keywords from the materials, students could compare the sentences in the list of questions with

those that they had highlighted. This process indirectly showed students the correct method to mark the key concepts. The restudy group also reviewed the important concepts in the text; however, repeated testing was found to better promote retention than restudying did (Butler & Roediger, 2007; McDaniel et al., 2007). Therefore, the experimental group demonstrated better reading skills. According to Bell and Limber (2009), high-skill readers are superior at identifying the important information in a text compared with their counterparts. By repeatedly taking the after-class practice tests, students learned how to correctly highlight the important concepts, which, in turn, improved their marking scores and reading skills.

5.1.2. Research question 2: Can students improve their reading engagement through machine-generated cloze item practice?

We examined whether students who used the reviewing system demonstrated different levels of reading engagement than those who restudied the materials. The results indicated that the experimental group students showed higher reading engagement than the control group did at a statistically significant level, meaning that they spent more time reading the e-books outside the class. To be able to answer the questions in the practice test, the experimental group students needed to review the e-books before taking the test. Therefore, they likely had more reading time than the control group students did and used interactive tools to facilitate their review process. After the test, they adjusted their reading skills on the basis of the results, suggesting that repeated testing motivated them to interact with the materials. Repeated testing can be used as a tool by students to evaluate their reading skills and revise it according to the results. The more tests students take, the more effort they put into learning. These effects are known as the indirect effects of testing (Roediger & Karpicke, 2006a). The increased duration of learning and improved reading skills after taking tests facilitate students' reading engagement and reading comprehension (Olney et al., 2017; Larsen et al., 2009).

5.2. Improved reading comprehension

5.2.1. Research question 3: Can students improve their reading comprehension with machine-generated cloze item practice?

We explored whether testing promoted students' reading comprehension. The study results revealed that students who took the practice test demonstrated significant improvement in reading comprehension compared with those who restudied the materials. This finding was consistent with the benefit of testing effect highlighted by Wiklund-Hörnqvist et al. (2014), who had also conducted an experiment in which feedback was provided. Meanwhile, the questions in reading comprehension posttest were different from the questions in cloze item practice for experimental group and the key concepts for control group, meaning that the improvement in reading comprehension were not caused by having more opportunities to practice the questions, as two groups were reviewing the same knowledge. Instead, it is the review methods that contributed to the difference in learning performance. Knowledge of key concepts is critical for students to comprehend a course (Kintsch et al., 1998). Studies have shown that learning the meaning of keywords improves reading comprehension (McDaniel & Pressley, 1989). However, different students normally exhibit various levels of reading skills in educational contexts—high-skill readers read more and learn more key concepts than low-skill readers do (Mol & Bus, 2011). In the present study, we addressed this reality by automatically generating practice tests that included key concepts using NLP techniques; we expected to reduce the gap in knowledge concerning key concepts between readers with different reading skills. The results highlighted that students who took the practice test demonstrated improved reading comprehension, regardless of their reading skills. This indicated that students with low reading skills could understand essential information even if they had failed to identify it before the test.

Provision of feedback is another factor that can improve reading comprehension. Kornell et al. (2011) stated that practice without feedback leads to a bifurcated item distribution in which only those items that are successfully retrieved are highly accessible by memory, whereas items that are not retrieved do not result in the testing effect. When students are provided with feedback, their memory strength becomes high enough to exceed a certain threshold; upon this threshold being crossed, the information becomes recallable. This promotes memory retention and prevents erroneous learning. Rowland (2014) indicated that no testing effect can be observed in the absence of feedback and that the retrievable rate is ≤50% in a laboratory setting. In the current study, the mask was removed from the cloze items for correct responses. Furthermore, students were allowed to see a hint if they could not answer correctly, which made each item recallable during every attempt. Our results indicated that the combination of repeated cloze item practice and the provision of feedback engendered the testing effect of enhancing students' reading comprehension.

6. Conclusion

Repeated testing has been shown to be an effective strategy for promoting memory retention and learning motivation. In this study, we employed cloze item practice that was automatically generated by BERT to explore two indirect testing effects, namely improvement in reading skills and reading engagement, and one direct testing effect, namely enhancement of reading comprehension. The results indicated that repeated testing significantly improved students' ability to extract key concepts from a text and motivated them to actively read the e-Books before and after the test, respectively. More importantly, their retention of learning content was also enhanced.

Several contributions are made by this study. First, the present study applied a modern AI technique to automatically generate tests for repeated testing in a real educational context. A majority of related studies that have examined the testing effect required instructors to prepare the practice test (Butler & Roediger, 2007; McDaniel et al., 2007; Wiklund-Hörnqvist et al., 2014). Although Olney et al. (2017) proposed a model that automatically generated cloze items for practice, they were using traditional machine learning and NLP techniques. The present study applied a DNN model (BERT) to build a system for generating practice tests. The results indicated that the use of cloze item practice along with the provision of feedback yielded a testing effect that positively influenced reading skills, reading engagement, and reading comprehension. Second, although the benefit of the testing effect has been broadly discussed in many studies (Greving & Richter, 2018; Wiklund-Hörnqvist et al., 2014), most have focused only on the improvement in the retention of taught content. Our study, by contrast, explored whether the testing effect is beneficial for not only students' reading comprehension but also their cognitive behaviors (i.e., reading skills and reading engagement) and determined that test-enhanced learning promotes students' ability to identify important information and motivates them to read. Finally, whether the question format influences the effectiveness of testing has been well investigated in prior research (Greving & Richter, 2018), with findings demonstrating that both short-answer type and multiple-choice questions yield the testing effect; however, the efficacy of other question formats has rarely been discussed. One of our objectives in this study was to investigate whether testing with cloze items is also effective for improving learning. True to our hypothesis, students who took the cloze item practice after class demonstrated greatly improved comprehension.

The current study's findings offer insights for instructors and researchers in related fields. Instructors can use these findings as a reference for guiding students in distinguishing relevant information from trivial content by testing the key concepts and enhancing their reading engagement. Furthermore, the summary generated by our model can be applied in other educational contexts. For example, instructors can use the summary to perform a test before a class to understand the average knowledge level of the class. The instructor can also adjust the summary by adding more sentences that they expect their students to learn or by removing some irrelevant sentences from the summary to develop a personalized summary that closely fits the course objectives. Furthermore, the current study suggests that the automatically generated cloze items are effective in enhancing students' comprehension. Future researchers can apply the same model as our study (BERT) or other modern AI techniques to generate different formats of questions, such as short-answer questions, for repeated learning. Moreover, researchers can develop personalized tests for individual students on the basis of their prior knowledge to improve their learning.

The present study has three limitations that warrant mention. First, the materials used in our experiment concerned topics that involve students' memory (accounting). Although repeated testing has been shown to improve memory retention, whether testing is still effective in promoting learning for materials that require logic and computation is unclear. Second, despite the encouragement given to the experimental group students to use the proposed system outside of class, this action was not mandatory. Therefore, we were unable to evaluate whether repeated testing promotes retention better than taking a single test does (Karpicke & Roediger, 2008). Finally, despite the retention test was conducted at the end of the experiment to measure students' comprehension, which we considered as a relatively long period, it is unclear whether the testing effects promoted long-term or short-term retention as students may make extensive use of the system to review right before appearing for a retention test. In this case, we can only argue that the testing effects in this experiment provided short-term retention.

In sum, our study results demonstrate that testing with BERT-generated cloze items is effective in promoting students' reading skills, reading engagement, and reading comprehension at the undergraduate level. More modern AI-driven testing can be applied to educationally relevant materials to facilitate learning. In our future research, students' review behaviors will be analyzed during testing and a personalized test will be generated on

the basis of their learning profile. Furthermore, we expect to try other DNN models for generating other question formats to develop a more comprehensive test.

Acknowledgement

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

References

Bell, K. E., & Limber, J. E. (2009). Reading skill, textbook marking, and course performance. *Literacy research and instruction*, *49*(1), 56-67.

Berry, T., Cook, L., Hill, N., & Stevens, K. (2010). An Exploratory analysis of textbook usage and study habits: Misperceptions and barriers to success. *College Teaching*, *59*(1), 31-39.

Bobby, Z., Nandeesha, H., Thippeswamy, D. N., Archana, N., Prerna, S., & Balasubramanian, A. (2018). 'Test–enhanced learning' by Closed book examination followed by feedback in Biochemistry. *South-East Asian Journal of Medical Education*, *12*(2), 19-24.

Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514-527.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & cognition*, *36*(3), 604-616.

Bossaller, J., & Kammer, J. (2014). Faculty views on eTextbooks: A Narrative study. *College Teaching*, *62*(2), 68-75.

Chan, Y. H., & Fan, Y. C. (2019, November). A Recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering* (pp. 154-162). Hong Kong, China: Association for Computational Linguistics.

Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of literacy research*, *43*(4), 352-392.

Daniel, D. B., & Woody, W. D. (2013). E-textbooks at what cost? Performance and use of electronic v. print texts. *Computers & Education*, 62, 18-23.

Dawson, S. P., McWilliam, E., & Tan, J. P. L. (2008). Teaching smarter: How mining ICT data can inform and improve learning and teaching practice. *Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education* (pp. 221-230). Melbourne, Australia: Deakin University.

DeBerard, M. S., Spielmans, G. I., & Julka, D. L. (2004). Predictors of academic achievement and retention among college freshmen: A Longitudinal study. *College student journal*, *38*(1), 66-81.

Dennis, A. (2011). e-Textbooks at Indiana University: A Summary of two years of research. *IRB*, *912000863*(1003001166), 0908000546. Retrieved from https://assets.uits.iu.edu/pdf/eText%20Pilot%20Data%202010-2011.pdf

Du, X., Shao, J., & Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. Retrieved from https://arxiv.org/abs/1705.00106

Fitzpatrick, L., & McConnell, C. (2009). Student reading strategies and textbook use: An Inquiry into economics and accounting courses. *Research in Higher Education Journal*, *3*, 1-10.

Flanagan, B., & Ogata, H. (2017). Integration of learning analytics research and production systems while protecting privacy. In *Proceedings of the 25th International Conference on Computers in Education* (pp. 333-338). New Zealand: Asia-Pacific Society for Computers in Education.

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).

Fouh, E., Breakiron, D. A., Hamouda, S., Farghally, M. F., & Shaffer, C. A. (2014). Exploring students learning behavior with an interactive etextbook in computer science courses. *Computers in Human Behavior*, *41*, 478-485.

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*, 2412.

Junco, R., & Clem, C. (2015). Predicting course outcomes with digital textbook usage data. *The Internet and Higher Education*, 27, 54-63.

Karpicke, J. D., & Roediger III, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 704 719. doi:10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2008). The Critical importance of retrieval for learning. *science*, *319*(5865), 966-968.

Karpicke, J. D. (2017). Retrieval-based learning: a decade of progress. In J. T. Wixted (Ed.), *Cognitive Psychology of Memory, of Learning and Memory: A Comprehensive Reference* (Vol. 2, pp. 487–514). Oxford, UK: Academic Press. doi:10.1016/B978- 0-12-809324-5.21055-9

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558.

Kintsch, W. (1998). *Comprehension: A Paradigm for cognition.* Cambridge, UK: Cambridge University Press.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A Distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85-97.

Landrum, R. E., Gurung, R. A., & Spann, N. (2012). Assessments of textbook usage and the relationship to student course performance. *College Teaching*, *60*(1), 17-24.

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A Randomised controlled trial. *Medical education*, *43*(12), 1174-1181.

Lorch, R. F., & Lorch, E. P. (1985). Topic structure representation and text recall. *Journal of Educational Psychology*, *77*(2), 137-148. doi:10.1037/0022-0663.77.2.137

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494-513.

McDaniel, M. A., & Pressley, M. (1989). Keyword and context instruction of new vocabulary meanings: Effects on text comprehension and memory. *Journal of Educational Psychology*, *81*(2), 204-213. doi:10.1037/0022-0663.81.2.204

Memory, D. M. (1983). Main idea prequestions as adjunct aids with good and low-average middle grade readers. *Journal of Reading Behavior*, *15*(2), 37-48.

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A Meta-analysis of print exposure from infancy to early adulthood. *Psychological bulletin*, *137*(2), 267–296. doi:10.1037/a0021890.

Mouri, K., Uosaki, N., Hasnine, M., Shimada, A., Yin, C., Kaneko, K., & Ogata, H. (2019). An Automatic quiz generation system utilizing digital textbook logs. *Interactive Learning Environments*, 1-14. doi:10.1080/10494820.2019.1620291

Nist, S. L., & Simpson, M. L. (1988). The effectiveness and efficiency of training college students to annotate and underline text. *National Reading Conference Yearbook, 37*, 251–257.

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

Okubo, F., Yamashita, T., Shimada, A., & Ogata, H. (2017). A Neural network approach for students' performance prediction. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 598-599). New York, NY: Association for Computing Machinery.

Olney, A. M., Pavlik, P. I., & Maass, J. K. (2017). Improving reading comprehension with automatically generated cloze item practice. In *Artificial intelligence in education* (pp. 262–273). Cham: Springer International Publishing. doi:10.1007/978-3-319-61425-0_22

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A Method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311318). Philadelphia, PA: Association for Computational Linguistics.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, *17*(3), 249-255.

Rowland, C. A. (2014). The Effect of testing versus restudy on retention: Meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463. doi:10.1037/a0037559

Russell, S., & Norvig, P. (2005). AI a modern approach. *Learning*, *2*(3), 4.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The Testing effect in the psychology classroom: A Meta-analytic perspective. *Psychology Learning & Teaching*, *16*(2), 179-196.

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*, 99-115.

Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feedback using text mining methods. *Procedia Computer Science*, *169*, 726-743.

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian journal of psychology*, *55*(1), 10-16.

Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly*, *19*(4), 404-425.

Woody, W. D., Daniel, D. B., & Baker, C. A. (2010). E-books or textbooks: Students prefer textbooks. *Computers & Education*, *55*(3), 945-948.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems* (Vol. 27, pp. 3320-3328). Retrieved from https://arxiv.org/abs/1411.1792

Zhang, L., Huang, Y., Yang, X., Yu, S., & Zhuang, F. (2019). An automatic short-answer grading model for semi-open-ended questions. *Interactive Learning Environments*, 1-14. doi:10.1080/10494820.2019.1648300