From Human Grading to Machine Grading: Automatic Diagnosis of e-Book Text Marking Skills in Precision Education

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: Precision education is a new challenge in leveraging artificial intelligence, machine learning, and learning analytics to enhance teaching quality and learning performance. To facilitate precision education, text marking skills can be used to determine students' learning process. Text marking is an essential learning skill in reading. In this study, we proposed a model that leverages the state-of-the-art text summarization technique, Bidirectional Encoder Representations from Transformers (BERT), to calculate the marking score for 130 graduate students enrolled in an accounting course. Then, we applied learning analytics to analyze the correlation between their marking scores and learning performance. We measured students' self-regulated learning (SRL) and clustered them into four groups based on their marking scores and marking frequencies to examine whether differences in reading skills and text marking influence students' learning performance and awareness of self-regulation. Consistent with past research, our results did not indicate a strong relationship between marking scores and learning performance. However, high-skill readers who use more marking strategies perform better in learning performance, task strategies, and time management than high-skill readers who use fewer marking strategies. Furthermore, high-skill readers who actively employ marking strategies also achieve superior scores of environment structure, and task strategies in SRL than low-skill readers who are inactive in marking. The findings of this research provide evidence supporting the importance of monitoring and training students' text marking skill and facilitating precision education.

Keywords: Text summarization, Marker grading, Self-regulated learning, Precision education, Text marking

1. Introduction

Precision education (Yang, 2019) is a challenge in applying artificial intelligence and machine learning techniques as well as learning analytics to enhance teaching quality and learning performance. Its goal is to identify at-risk students as early as possible and provide timely assistance through diagnosis, prediction, treatment, and prevention—specifically, it aims to identify students' learning patterns and behaviors during the learning process and predict their learning outcomes. The instructor can then provide personalized feedback or intervention to those at-risk students to prevent them from failing the course.

With advancements in information and communication technology, learning actions can be logged by a learning management system, such as Moodle (Ogata et al., 2017). Learning analytics can be used to analyze these logs—for example, the time students stay on a certain page or the markings or notes they make. Nian et al. (2019) and Yin et al. (2019) analyzed how often students used the e-reader functions such as NEXT, PREV, and MARKER to determine whether these behaviors are related to learning performance. They found that students who used the marker function tended to achieve superior learning performance. Al-khazraji (2019) observed that learners who used markers to highlight the sentences they thought were important exhibited considerably improved learning effectiveness. Yufan et al. (2020) proposed that in addition to analyzing the marking frequency, the area of marking may also be related to learning performance. However, if the content of the text being marked is not considered, marking can be overused or misused, resulting in decreased learning performance.

Measuring the content of markings reveals student's comprehension of the course. Normally, students mark the sentences or words they perceive as important during their learning process. The increasing use of e-learning systems has made it easier for instructors to observe students' learning behavior and provide relevant advice. Traditionally, the assessment process has typically been performed by instructors or teachers. However, this process is unsuitable when teaching resources are limited. To address this issue, text summarization techniques can be applied to automate the assessment process.

The adoption of marking or underlining is a metacognitive skill that enables learners to identify the essential concepts and focus on them during the review process (Van Horne et al., 2016). Similarly, self-regulated

learning (SRL) is a learning process that includes multiple metacognitive skills and positively affects learning performance (Michalsky & Schechter, 2013; Siadaty et al., 2012). We speculated that learners with SRL skills are more likely to adopt and optimize the effect of the text marking strategy. For instance, students who are good at task strategies may mark the critical concepts to enhance their memory retention. Therefore, this study also evaluated the correlation between text marking and SRL.

2. Literature review

2.1. Precision education

Precision education involves four phases: diagnosis, prediction, treatment, and prevention. The process of identifying students' learning patterns and behavior, predicting learning outcomes based on the collected data, providing timely intervention, and preventing them from failing in the course is similar to the procedure when a doctor gives a patient a treatment based on their symptoms, hence the name precision education. Contrary to the traditional one-size-fits-all approach to teaching, precision education aims to provide individual students with personalized feedback and treatment according to their learning profile. Empirical studies have shown the effectiveness of precision education. For example, Lu et al. (2018) applied learning analytics early on in a blended calculus course to predict students' academic performance and identified seven critical factors affecting their performance. Hu et al. (2014) developed early-warning systems to predict at-risk students during a course in progress. They found that time-dependent variables are essential to predict student online learning performance. In this study, we explored whether the content marked by students in an e-book can predict their learning performance.

2.2. Text summarization techniques for key concept extraction

Automated text summarization has been studied since the late 1950s (Luhn, 1958). Many studies have focused on extractive summarization using statistical methods (Das & Martins, 2007). TextRank is an extractive and unsupervised text summarization technique introduced by Rada Mihalcea and Paul Tarau (Mihalcea & Tarau, 2004). It has been used to summarize meeting transcripts (Garg et al., 2009) and assess web content credibility (Balcerzak et al., 2014). Rose et al. (2010) introduced rapid automatic keyword extraction (RAKE), an unsupervised, domain-independent, and language-independent method for text summarization. They applied RAKE to a corpus of news articles to extract keywords that are essential to documents. While these studies achieved information retrieval by using traditional machine learning algorithms, researchers from Google built an unsupervised learning architecture called Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2018). It is a deep learning model developed on top of the Transformer architecture (Vaswani et al., 2017), which beats nearly all existing models in the natural language processing (NLP) field for various tasks (Devlin et al., 2018).

2.3. SRL in learning analytics

SRL is a learning process involving cognitive and metacognitive strategies that enhance students' motivation to learn and reflect on their learning process, thereby contributing to their comprehension of studied materials (Michalsky & Schechter, 2013; Siadaty et al., 2012). Through SRL, students can come to deeply understand complex topics in the learning process (Jacobson & Archodidou, 2012; Järvelä et al., 2015; Labuhn et al., 2008). However, their behaviors and attitudes reflect SRL, which also contributes to their self-confidence (Artino & Jones, 2012; Stefanou et al., 2014).

After realizing the importance of SRL in online learning, studies have investigated whether the use of SRL strategies influences students' learning and metacognitive skills. For example, Littlejohn et al. (2016) discovered that learners' motivations and goals influence how they conceptualize the purpose of a course, which in turn influences their perception of the learning process. Similarly, Kizilcec et al. (2017) explored learners' SRL skills based on their overall course achievement, engagement in course content, and survey responses across six massive open online courses. They compared various SRL strategies and learners' actual behaviors and found that goal setting and strategy planning were associated with the attainment of personal course goals. Moreover, learners with high self-reported SRL were more likely to review the learned content.

Learning analytics have potential advantages in the examination of student SRL in online learning environments (Järvelä et al., 2016). The use of markers on e-book readers is considered a cognitive and metacognitive strategy (Van Horne et al., 2016). High-skill readers who have good metacognitive knowledge can identify and isolate essential concepts, whereas low-skill readers have difficulties in marking the most relevant information, leading to the overuse or misuse of the marking strategy. In this study, we collected data on students' self-reported SRL by asking them to complete an SRL questionnaire before the experiment. Their marker records were logged in the BookRoll database, and the learning analytics approach was applied to investigate their marking patterns and frequencies and SRL skills. We hypothesized that learners who are high-skill readers and frequent marker users have higher SRL skills. This approach may provide a new method to analyze the relationship between different groups of students stratified by reading and SRL skills.

2.4. The association between learning analytics, learning behaviors, and marking strategy

Using markers is considered a metacognitive learning behavior that positively influences reading comprehension (Van Horne et al., 2016). According to Pearson et al. (2016), the prime tasks of students when reading a textbook are to (a) focus attention and (b) engage in encoding activities in a manner that will increase the probability of understanding and retrieving the high pay-off ideas and relationships. Thus, they should identify the most relevant information in the text and test themselves to ensure that the material is understood and remembered. Nist and Hogrebe (1987) argued that students cannot remember everything they read. Thus, the use of text marking can help them identify and isolate key concepts. Specifically, college students are often required to learn and remember a large amount of information for assessments over a relatively long period. The strategy of marking or underlining the most relevant information can help them reduce the amount of information they need to review. Nist and Hogrebe (1987) further argued that instead of merely engaging in passive reading, text marking allows students to actively engage with the learning materials, which is believed to improve their memory retention. Chickering and Gamson (1987) also suggested the importance of active learning—learning that engages the student and makes them active participants in the learning process.

Studies have demonstrated the importance of marking skill in reading and learning performance. Bell and Limber (2009) noted differences in reading skills based on how much students mark, with low-skill readers marking more than high-skill readers. Nian et al. (2019) explored correlations between student reading engagement and learning outcomes and applied machine learning to predict learning outcomes. Yin et al. (2019) analyzed reading pattern behaviors, such as deleting markings or bookmarks after adding them. These findings suggest that instructors must train students on how to effectively mark a text. Many researchers have simply measured the frequency of text marking in digital textbooks without considering the content being marked, thus neglecting overmarking or mismarking by students, which can confound the results.

To overcome this limitation and investigate whether marked content has a strong correlation with learning performance, we used a method to automatically summarize the text from learning materials to calculate the marking score, which was referred to as marking quality in this study. We further measured students' reading skills using their marking score and marking frequency and evaluated whether it affects learning performance and SRL. This study addressed the following research questions:

- Can machines extract concepts that are approximate to the key concepts extracted by humans for marker grading?
- What is the relationship between marking quality, marking frequency, and learning outcomes?
- Is there a difference among students with varying levels of reading skills in learning performance and SRL?

3. Methods

3.1. Research context

A 12-week accounting course offered to undergraduate students at a university in Taiwan was the research context. The course was mandatory for students majoring in accounting but open to other majors as an elective course. The course used an e-book system, BookRoll, in which students read the e-books uploaded by the instructor. Thirteen slides were uploaded to BookRoll. BookRoll is an e-book reading system (Flanagan & Ogata, 2017) developed by Kyoto University. Students' e-book reading actions in BookRoll have been introduced in detail by Ogata et al. (2015) and Flanagan and Ogata (2018). One hundred thirty-two undergraduate students in the department of accounting took the course. Students completed a questionnaire

containing SRL questions at the beginning of the course. Students took a midterm exam in the middle of the semester and a final exam at the end of the semester. All students completed the course, but two failed to complete the SRL questionnaire, and consequently, we used the data from 130 students for analysis.

3.2. Procedure

In the first week, the students completed the prequestionnaire, which comprised items related to self-regulation. The instructor then introduced the syllabus of the course and demonstrated how to use BookRoll functions, such as opening slides, using markers, and adding memos. Before each class, students were required to preview the learning materials uploaded by the instructor on BookRoll. They were also encouraged to use the various functionalities in BookRoll, such as using markers or adding memos, as their use of those interactive tools counted toward their learning activity score. Instructors can highlight the sentences they want student to pay more attention or post notes to provide furhter explanation. At weeks 8 and 12, the students took a midterm and final exam, respectively. All activities by students on the e-books were logged in the BookRoll database. We applied BERT to automatically extract the concepts from learning materials and used them as reference answers to calculate the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002) of student markings as their marking score. The students received a BLEU score ranging from 0 to 1 each week, and the sum of scores over 12 weeks constituted their final marking score. We also determined the frequency of using markers by collecting the total number of markings made and preserved by students at the end of the experiment. We calculated the Spearman correlation coefficient between the final marking score, marking frequency, and students' learning performance, which was measured using their midterm (30%) scores, final exam(40%) scores, and performance in learning activities during the class (30%), to investigate whether marking quality and marking frequency are correlated with learning performance.

In addition, we applied k-means clustering (Lloyd, 1982) to categorize students into four groups: high-skill readers who prefer marking, high-skill readers who do not like marking, low-skill readers who prefer marking, and low-skill readers who do not like marking. We also compared differences in students' awareness of SRL between the groups. The two features used in k-means were students' reading skills and activeness in using markers. Reading skill was measured using students' final marking score, whereas activeness in using markers was assessed using marking frequency.

3.3. Preprocessing

Automatic extraction of key concepts by using text summarization models requires text preprocessing. We applied Python's pdfminer package to convert the PDF learning materials to plain text files and extract the sentences. Both instructor and students' markers were collected using BookRoll. We removed special characters throughout and converted the text to lowercase. In addition, we ignored the deleted markings and used only the marking that were preserved by the students. Furthermore, we added two special tokens, [CLS] and [SEP], before and after the input for the BERT model, respectively. The [CLS] token in the output of BERT stores the embeddings of the text that represent its syntax and meaning. The [SEP] token serves as the separator between sentences. To simplify the preprocessing steps for BERT, the open-source transformers package developed by the Huggingface team was used.

3.4. Text summarization

We compared three text summarization techniques for automatically extracting key concepts from learning materials. For the traditional algorithms—TextRank and RAKE—the preprocessed text was passed as the input to the models and the keywords in the learning materials were extracted. TextRank is an unsupervised machine learning algorithm used to extract keywords from a text. It applies the idea of the PageRank algorithm (Page et al., 1999) developed by Google for webpage search to calculate the weights of each keyword. RAKE is an unsupervised, language-independent machine learning algorithm that extracts both keywords and key phrases. It splits the text into sentences using special characters and breaks downs each sentence at stop words. The idea behind this is that keywords usually contain multiple words and are surrounded by stop words. After extracting the candidate phrases, each candidate phrase is weighted by the co-occurrence of the words it contains. We selected the top 15 keywords based on the weights and the sentences, which include those words related to the key concepts in the materials. In this study, we adopted open-source TextRank4ZH and Rake_For_Chinese to implement these two algorithms.

BERT is a state-of-the-art technique that adopts the popular two-step transfer learning (Torrey & Shavlik, 2010) used in the NLP field. First, a general model that can understand the basic syntax of the natural language is generated during the pretraining phase, which can then be used for feature extraction or fine-tuned to perform downstream tasks. The self-attention mechanism allows BERT to understand the contextual meaning and learn the syntax structure in the text. For BERT, a pretrained model developed by the transformers team was adopted. We first passed the text of materials to the BERT encoder and extracted the last hidden state of the [CLS] token to acquire a 768-dimension embedding representing the key concepts of materials. To overcome the limit of the number of words that can be sent to the model at one time, we divided the text by pages, retrieved the embeddings of each page, and summed them to acquire the embeddings of the complete text. Next, we applied the same procedure to obtain the embeddings of each sentence in the text. Finally, we calculated the cosine distance between the embeddings of materials and the embeddings of each sentence in the vector space to measure whether the sentence is similar to the key concepts in the materials; we selected the sentences closest to the materials as key concepts.

To compare the performance of the three models, we used the markers provided by the instructor as the reference answer to evaluate the quality of summaries by the machine using BLEU 1, BLEU 2, BLEU 3, BLEU 4 (Papineni et al., 2002), and METEOR (Denkowski & Lavie, 2014) scores. BLEU is a precision-based measure for evaluating a hypothesis translation of a text to reference translations. BLEU-n is a variant of the BLEU score that applies up to a specified n-gram for counting co-occurrences. METEOR is a recall-oriented metric that evaluates the generated sentences using stemming and synonymy matching along with standard exact word matching. Both BLEU and METEOR scores range from 0 to 1. The higher the score is, the better the machine-generated sentences are. In this study, we calculated four scores for the summaries generated by different models every week and summed the scores as the performance of each model.

4. Results

4.1. Analysis of machine grading and human grading

Table 1 presents the scores of each model in different metrics. BERT performed the best in all metrics except for BLEU-4. Therefore, we adopted BERT to extract key concepts from the learning materials, with these extractions serving as the reference answer for marker grading in the following experiment. Considering that most keywords in the key concepts and student markings contained only one or two words, BLEU-1 was chosen as the metric when grading markers.

Table 1. Evaluation of TextRank, RAKE, and BERT

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
TextRank	7.18	5.39	4.21	**3.16**	.25
RAKE	7.4	5.48	4.13	3.12	.26
BERT	**7.98**	**5.97**	**4.31**	3.03	**.43**

Table 2. Cohen's kappa coefficients of different grading results

Human	Machine		
	< 0.2	0.2–0.8	> 0.8
< 0.2	17	0	0
0.2–0.8	5	43	0
➢ 0.8	0	16	7

Table 3. Classification reports of the BERT model

Marking score	Precision	Recall	*F1*
< 0.2	0.77	1.0	0.87
0.2–0.8	0.73	0.9	0.80
> 0.8	1.00	0.3	0.47

Student markings were graded, and the sum of marking scores was used as their final marking score. Before analyzing the correlation between marking score and learning performance, the machine-grading results were compared with human-graded marking to ensure its reliability by calculating Cohen's kappa coefficient, which is often used to measure interrater agreement. Table 2 presents the grading results for machines and humans. The Cohen's kappa coefficient was 0.57, indicating a moderate agreement between the two grading results. The Spearman correlation between the two grading results ($rs = 0.84^{***}$, $p < .001$) indicated that two grading

approaches were highly correlated. Table 3 presents the classification reports of the BERT model. The model achieved the best *F1* score of 0.87 when the marking score was < 0.2 and the worst *F1* score of 0.47 when the marking score was > 0.8, meaning that our model is good at identifying bad markers but may sometimes generate low scores for students who actually mark well. To make sure the marking score used in this study would reflect students' actual performance, the machine grading results were cross-examined by the instructor. If the instructor did not agree with the scores, the scores were adjusted. On the basis of our results, we conclude that machine grading can be used to automatically grade student markings but still requires assistance from humans for a specific group of students.

4.2. Analysis of marking frequency and quality

To compare differences between high and low achievement groups in the score and number of markings, the Kruskal–Wallis H-test was performed. The Shapiro–Wilk test of marking score and marking frequency indicated a nonnormal distribution of these variables. We divided the students into high (n = 44) and low achievement (n = 44) groups by taking the upper and lower thirds of the learning performance from the 130 participants. The high achievers did not perform differently in marking score (H = 0.07, p > .05) and marking frequency (H = 0.93, p > .05) than low achievers (Table 4). Because no clear linear relationship was observed between marking score, marking frequency, and learning performance, the Spearman correlation coefficient was calculated to measure the relationship among them (Table 5). No correlation was noted between marking score and learning performance (rs = −0.01, p >.05). A weak but nonsignificant correlation was noted between marking frequency and learning performance (rs = 0.12, p >.05). Finally, the marking score was highly correlated with marking frequency (rs = 0.68***, p < .001). However, this result does not indicate that students can get a high marking score by simply marking more. In Figure 1, when the number of markings is below a threshold of 200, the marking score improves rapidly as the students make more markings. However, once the number exceeds 200, the improvement decreases drastically. We further averaged the marking scores of all students through the semester to investigate whether their ability to identify the key concepts from the learning materials varied in different periods (Figure 2). From Weeks 1 to 6, the marking score was relatively moderate and stable, meaning that students may not have found the motivation to use the marking strategy or were still unfamiliar with the system. The average of marking scores reached the peaks in Weeks 7 and 11, indicating that the students were more familiar with the system and were marking more frequently before the midterm and final exams in Weeks 8 and 12. In Week 9, no new slides were uploaded because the instructor was reviewing the midterm exam with students. To answer the research question, we conclude that, consistent with previous studies (Fowler & Barker, 1974; Hoon, 1974; Idstein & Jenkins, 1972), neither marking frequency nor marking score has a significant correlation with learning performance. However, a strong relationship exists between marking score and marking frequency, which leads into the discussion of our third research question.

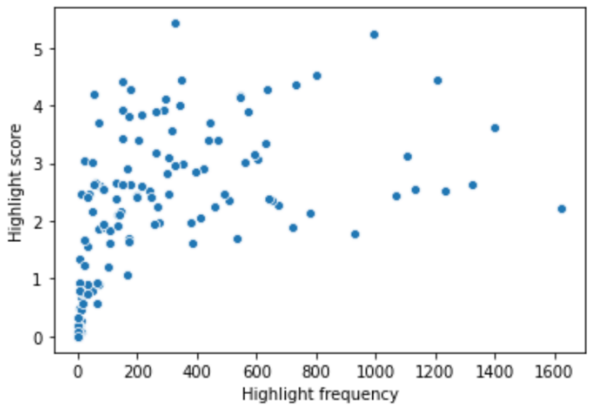

Figure 1. Marker score and marking frequency

Table 4. Results of marking score and marking frequency

	High achievement			Low achievement			
	N	*M*	*SD*	*N*	*M*	*SD*	*H*
Marking score	44	2.40	1.43	44	2.31	1.31	0.07
Marking frequency	44	271.0	439.76	44	152.5	301.25	0.93

Table 5. Correlations between marking score, marking frequency, and learning performance

	1.Marking score	2.Marking frequency	3.Learning performance
1.Marking score	-	0.68***	−0.01
2.Marking frequency		-	0.12
3.Learning performance			-

Note. ***$p < .001$.

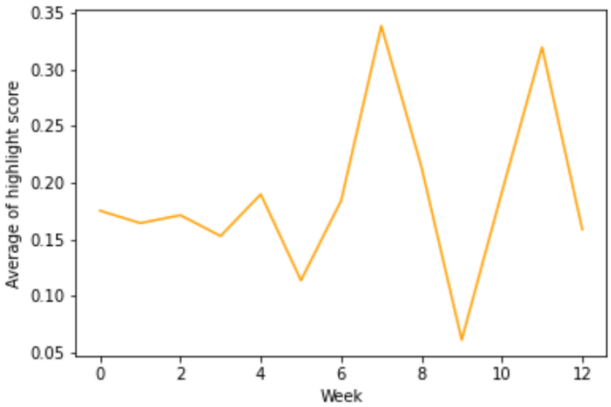

Figure 2. Average marking score throughout the semester

4.3. Analysis of learning performance

To determine whether different reading skills influence students' learning performance and SRL, we used k-means clustering to classify students into four groups: inactive marker users who are high- (Group A, $N = 60$) or low-skill readers (Group B, $N = 41$) and active marker users who are high- (Group C, $N = 27$) or low-skill readers (Group D, $N = 2$). Because Group D only contained two samples, it was not considered in the following test. The Shapiro–Wilk test results revealed a value of 0.97 ($p < .05$), indicating that this sample did not show a normal distribution.

The Kruskal–Wallis H-test was employed to evaluate students' learning achievement in the different groups. Table 6 shows the results of the learning achievement according to final grades. The medians and standard deviations were 81.0 and 7.79, respectively, for Group A; 83.0 and 7.63, respectively, for Group B; and 85.0 and 10.14, respectively, for Group C. The final grades of the three groups were not significantly different ($H = 4.13$, $p > .05$). However, the post hoc Conover test indicated that the learning performance of Group C was significantly higher than that of Group A. This implies that high-skill readers who prefer marking are more likely to also achieve better final grades than other high-skill readers who do not prefer marking.

Table 6. The Kruskal test result of learning performance of three groups

Group	*N*	*M*	*SD*	*H*	Post hoc test (Conover)
A	60	81.0	7.79	4.13	
B	41	83.0	7.63		
C	27	85.0	10.14		C > A*

Note. *$p < .05$.

4.4. Analysis of self-regulation

The SRL questionnaire comprised items regarding six aspects: goal setting, environment structure, task strategies, time management, help-seeking, and self-evaluation. We summed the scores of the six aspects to represent the total SRL score. Table 7 shows the results of the students' self-regulation along with the scores of the six aspects of SRL across the three groups. The Kruskal–Wallis H-test indicated a nonsignificant effect on students' self-regulation in all groups ($H = 4.87$, $p > .05$). The medians and standard deviations of Groups A, B, and C were 122.0 and 20.11, 118.0 and 17.42, and 131.0 and 18.95, respectively. However, the post hoc test indicated a significant difference between Groups B and C in self-regulation, implying that high-skill readers who prefer marking exhibit a better awareness of self-regulation than low-skill readers who do not prefer marking.

To further investigate the students' awareness of their self-regulation in each aspect, the Kruskal test was employed again. As presented in Table 7, no significant difference was observed in goal setting ($H = 1.78$, $p > .05$), help-seeking ($H = 1.54$, $p > .05$), or self-evaluation ($H = 2.06$, $p > .05$). However, a significant difference was noted between the three groups in environment structure ($F = 6.59$, $p < .05$). The Conover post hoc test indicated that Group C obtained a significantly higher score in the awareness of environment structure ($N = 27$, $M = 24.0$, $SD = 2.81$) than Group B ($N = 41$, $M = 22.0$, $SD = 3.30$). A significant difference in task strategies ($H = 7.32$, $p < .05$) was detected between the three groups. Group C performed significantly better than Groups A and B. Also, the time management scores between the three groups were significantly different ($F = 6.13$, $p < .05$). Group C obtained a significantly higher score than Group A.

Table 7. Kruskal–Wallis test results of SRL in the three groups

Dimension	Group	*N*	*M*	*SD*	*H*	Post hoc test (Conover)
SRL	A	60	122.0	20.11	4.87	
	B	41	118.0	17.42		
	C	27	131.0	18.95		C > B*
Goal setting	A	60	26.5	4.67	1.78	
	B	41	27.0	3.96		
	C	27	28.0	3.68		
Environment structure	A	60	24.9	3.38	6.59*	
	B	41	22.0	3.30		
	C	27	24.0	2.81		C > B*
Task strategies	A	60	19.0	3.68	7.32*	
	B	41	18.0	3.68		
	C	27	21.0	3.61		C > A*; C >B**
Time management	A	60	13.0	3.14	6.13*	C > A*
	B	41	14.0	3.20		
	C	27	15.0	3.11		
Help-seeking	A	60	20.0	4.17	1.54	
	B	41	20.0	3.75		
	C	27	21.0	4.12		
Self-evaluation	A	60	20.0	3.66	2.06	
	B	41	20.0	3.33		
	C	27	21.0	3.87		

Note. $^{**}p < .01$, $^{*}p < .05$.

5. Discussion

5.1. Automatic concept extraction and marker analysis of learning performance

Research question 1: Can machines extract concepts that are approximate to the key concepts extracted by humans for marker grading?

Consistent with the previous study (Devlin et al., 2018), BERT outperformed other text summarization models in the natural language understanding task in the present study. The text input to BERT included slides made by the instructor. Most of the learning materials consisted of incomplete sentences or phrases in bullet points rather than a passage or paragraphs that would include a contextual meaning. Bidirectional encoding of each sentence allowed BERT to output the embeddings containing information on syntax and semantic meanings and choose the most relevant sentences in the vector space.

Using BERT-generated summarization to evaluate whether students mark key concepts consistent with the instructor's answers, we found our model generated decent results: most students who are poor or moderate at marking can be graded correctly, but students who are good at marking may sometimes receive only moderate scores (0.2–0.8). Overall, the machine can help identify students who are unable to mark the key concepts, and then, the instructor can focus on helping them better identify the relevant concepts. This finding facilitates execution of the diagnosis phase in precision education.

Research question 2: What is the relationship between marking quality, marking frequency, and learning performance?

The finding that the content being marked and the use of the marker function did not vary greatly between high and low achievers indicates that marking score or frequency alone may not determine success in online learning courses, consistent with previous findings (Fowler & Barker, 1974; Hoon, 1974; Idstein & Jenkins, 1972; Oi et al., 2017). Nevertheless, the text marking strategy may still influence learning success. The instructor needs to guide students in how to use the appropriate text marking strategy in online learning, as many students may not be familiar with the use of many functions in the e-book system. The effectiveness of marking or underlining can be optimized when students are trained on how to use them (Nist & Simpson, 1988; Yue et al., 2015). Therefore, we recommend that instructors introduce the marker function and the text marking strategy at the beginning of the course or encourage students to use this strategy and provide students with the concepts marked by them for comparison. By achieving this, text marking skills can still become a predictor of learning performance insofar as it reflects students' metacognitive skills, which correlate with learning performance (Van Horne et al., 2016).

5.2. Analysis of reading skill, activeness of marking, learning performance, and SRL

Research question 3: Is there a difference among groups with varying levels of reading skills in learning performance and SRL?

We classified students into four groups by using k-means clustering and found that high-skill readers who prefer marking exhibited significantly higher learning performance than high-skill readers who do not prefer marking. Marking is most effective when the reader has maximum faith that the marker can discriminate between essential material and trivia (Fowler & Barker, 1974). Active marker users recognize that marking can help them identify and isolate the most relevant information for later review, thereby enhancing their understanding and long-term memory (Annis & Davis, 1978; Chickering & Gamson, 1987; Nist & Hogrebe, 1987; Rickards & August, 1975).

Furthermore, high-skill readers who prefer marking exhibited a significantly higher self-reported SRL than did low-skill readers who did not prefer marking. This suggests that students with better SRL, metacognitive strategies, and motivation believe that marking is beneficial during their learning process, which leads to better use of this strategy. We further explored students' awareness of SRL in each aspect and found that high-skill readers who are active in marking performed significantly better in environment structure than did low-skill readers who do not prefer marking. Students who can find the environment for them to focus on studying or who are good at using different learning tools during their learning process can benefit from the marker function in e-books. We also found that students who are both high-skill readers and active marker users exhibited significantly better task strategies than other groups. Students who exhibit highly developed task strategies use the marker function to identify the relevant information and ignore the trivia, which enhances their reading efficiency (Nist & Simpson, 1988). Finally, high-skill readers who are active in marking exhibited significantly better time management than did high-skill readers who do not prefer marking. Students who manage their time well prefer to apply the text marking strategy during their learning, as they believe this strategy can improve their reading and learning efficiency (Nist & Simpson, 1988). This finding suggests that the combination of marking quality and marking frequency can be an indicator of students' learning performance and SRL. The instructor can provide timely interventions in terms of training on text marking skills.

6. Conclusion

Our findings provide several contributions to analysis of text marking analysis on an e-book system. First, most of the learning material in this study was PDF slides made by instructors. They extracted the passages or sentences they thought were relevant to the core of the course. These texts were often sentences without contextual meaning or phrases, which is different from a traditional textbook. To our understanding, most research on text summarization has used textbooks, papers, or datasets for competition, whereas our study extracted concepts from slides made by instructors.

Second, this study applied an advanced tool, BERT, to extract concepts from learning materials and use them as reference answers to automatically grade student markings, whereas studies have mainly explored the number, area, or pattern of markings (Yufan et al., 2020). According to Memory (1983) and Meyer et al. (1980), only high-skill readers effectively identify the most relevant materials. High-skill readers pay attention to the essential concepts in the text to a greater degree than low-skill readers do (Lorch & Pugzles-Lorch, 1985). Therefore, we measured not only marking frequency but also students' marking scores to assess their reading skills.

Third, this study revealed the correlation between reading skills, learning performance, and SRL. The combination of marking quality and marking preference has not been used previously. According to our findings, mere frequency of marking did not guarantee high marking scores. Therefore, we considered both indicators when measuring student learning performance and SRL and found that high-skill readers who are active in marking perform significantly better in learning performance, SRL, environment structure, task strategies, and time management.

Finally, the findings of this research provide insights for instructors, students, and researchers in the field of precision education. Instructors can use the proposed system to track students' text marking patterns, predict at-risk students through their marking patterns, and provide timely personalized feedback for individual students. Students can be kept aware of their text marking ability through the system. Weekly monitoring of their marking score and marking frequency enables them to determine whether their learning strategy is effective and whether they should improve their reading skills. Finally, as the current study emphasizes the diagnosis and prediction phase in precision education, future studies can focus on treatment and prevention using text marking patterns—for example, whether students improve their text marking patterns after seeing an e-book in which the key concepts have been marked by the instructor. To summarize, our findings can help instructors and students track text marking skills and demonstrate that the conceptualization of automatically grading markers is effective in the precision education field.

This study has a few limitations. First, this experiment was conducted in an accounting course that involves many terms and theories that require memory retention. Thus, marking may benefit students when a large amount of information needs to be remembered. Using the marker function allows students to filter out irrelevant information. However, whether marking is as effective when students engage in a course where computing and logic are used more frequently remains unclear. Second, the students were not given any instructions regarding marking strategies in this study during their learning. It has been shown that students well trained in marking performed better in learning performance and learning efficiency than those not trained in marking. Thus, students' marking scores can positively influence learning performance after they receive training; thus, as students improve their ability to identify key concepts, their learning performance can be enhanced.

In summary, our study leveraged the advanced text summarization model to automatically grade student markings and analyzed the effectiveness of text marking strategy on learning performance and SRL. Leutner et al. (2007) stated that it is worth training students in specific metacognitive learning strategies and teaching them to examine and regulate their use of those strategies in learning. Future studies should explore whether teaching students to mark on e-books and SRL strategies affects their learning performance.

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 Special Innovation Program on AI and Big Data 18102059-0.

References

Al-khazraji, A. (2019). Analysis of discourse markers in essays writing in ESL classroom. *International Journal of Instruction, 12*(2), 559-572.

Annis, L., & Davis, J. K. (1978). Study techniques: Comparing their effectiveness. *The American Biology Teacher, 40*(2), 108-110.

Artino Jr, A. R., & Jones II, K. D. (2012). Exploring the complex relations between achievement emotions and self-regulated learning behaviors in online learning. *The Internet and Higher Education, 15*(3), 170-175.

Balcerzak, B., Jaworski, W., & Wierzbicki, A. (2014). Application of textrank algorithm for credibility assessment. In *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)* (Vol. 1. pp. 451–454). doi:10.1109/WI-IAT.2014.70

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

Chickering, A. W., & Gamson, Z. F. (1987). Seven principles for good practice in undergraduate education. *AAHE bulletin, 3*, 7.

Das, D., & Martins, A. (2007). A Survey on automatic text summarization. literature survey for language and statistics. *II Course at CMU*.

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.*

Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation* (pp. 376-380). Baltimore, MD: Association for Computational Linguistics.

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). Christchurch, New Zealand: Asia-Pacific Society for Computers in Education (APSCE).

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

Fowler, R. L., & Barker, A. S. (1974). Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology, 59*(3), 358-364. doi:10.1037/h0036750

Garg, N., Favre, B., Reidhammer, K., & Hakkani-Tür, D. (2009). Clusterrank: A Graph based method for meeting summarization. In *Tenth Annual Conference of the International Speech Communication Association* (pp. 1499-1502). Brighton, United Kingdom: International Speech Communication Association.

Hoon, P. W. (1974). Efficacy of three common study methods. *Psychological Reports, 35*(3), 1057-1058.

Hu, Y. H., Lo, C. L., & Shih, S. P. (2014). Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior, 36*, 469-478.

Idstein, P., & Jenkins, J. R. (1972). Underlining versus repetitive reading. *The Journal of Educational Research, 65*(7), 321-323.

Jacobson, M. J., & Archodidou, A. (2012). The Knowledge mediator framework: Toward the design of hypermedia tools for learning. In *Innovations in Science and Mathematics Education* (pp. 128-172). New York, NY: Routledge.

Järvelä, S., Kirschner, P. A., Panadero, E., Malmberg, J., Phielix, C., Jaspers, J., Koivuniemi, M., & Järvenoja, H. (2015). Enhancing socially shared regulation in collaborative learning groups: designing for CSCL regulation tools. *Educational Technology Research and Development, 63*(1), 125-142.

Järvelä, S., Malmberg, J., & Koivuniemi, M. (2016). Recognizing socially shared regulation by using the temporal sequences of online chat and logs in CSCL. *Learning and Instruction, 42*, 1-11.

Kizilcec, R. F., Pérez-Sanagustín, M., & Maldonado, J. J. (2017). Self-regulated learning strategies predict learner behavior and goal attainment in Massive Open Online Courses. *Computers & Education, 104*, 18-33.

Labuhn, A. S., Boegeholz, S., & Hasselhorn, M. (2008). Fostering learning through stimulation of self-regulation in science lessons. *Zeitschrift fur Padagogische Psychologie, 22*(1), 13-24.

Leutner, D., Leopold, C., & den Elzen-Rump, V. (2007). Self-regulated learning with a text-highlighting strategy. *Zeitschrift für Psychologie/Journal of Psychology, 215*(3), 174-182.

Littlejohn, A., Hood, N., Milligan, C., & Mustain, P. (2016). Learning in MOOCs: Motivations and self-regulated learning in MOOCs. *The Internet and Higher Education, 29*, 40-48.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*(2), 129-137.

Lorch, R. F., & Lorch, E. P. (1985). Topic structure representation and text recall. *Journal of Educational Psychology, 77*(2), 137.

Lu, O. H. T., Huang, A. Y. Q., Lin, A. J. Q., Ogata, H., & Yang, S. J. H. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educational Technology & Society, 21*(2), 220–232.

Luhn, H. P. (1958). The Automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165. doi:10.1147/rd.22.0159

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.

Meyer, B. J., Brandt, D. M., & Bluth, G. J. (1980). Use of top-level structure in text: Key for reading comprehension of ninth-grade students. *Reading research quarterly, 16*(1), 72-103. doi:10.2307/747349

Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411). Barcelona, Spain: Association for Computational Linguistics.

Michalsky, T., & Schechter, C. (2013). Preservice teachers' capacity to teach self-regulated learning: Integrating learning from problems and learning from successes. *Teaching and Teacher Education, 30*, 60-73.

Nian, M. W., Lee, Y. H., & Wu, J. Y. (2019). Using machine learning to explore the associations among e-reader operations and their predictive validity of learning performance. In *International Conference on Learning Analytics & Knowledge (LAK19)* (pp. 1-6). Tempe, AZ: Association for Computing Machinery.

Nist, S. L., & Hogrebe, M. C. (1987). The Role of underlining and annotating in remembering textual information. *Literacy Research and Instruction, 27*(1), 12-25.

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., Oi, M., Mohri, K., Okubo, F., Shimada, A., Yamada, M., & Hirokawa, S. (2017). Learning analytics for e-book-based educational big data in higher education. In *Smart Sensors at the IoT Frontier* (pp. 327- 350). doi:10.1007/978-3-319-55345-0_13

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). Hangzhou, China: Asia-Pacific Society for Computers in Education.

Oi, M., Okubo, F., Taniguchi, Y., Yamada, M., & Konomi, S. (2017). Effects of prior knowledge of high achievers on use of e-book highlights and annotations. In *Proceedings of the 25th International Conference on Computers in Education, ICCE 2017 - Main Conference Proceedings* (pp. 682-687). Christchurch, New Zealand: Asia-Pacific Society for Computers in Education.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318). Philadelphia, PA: Association for Computational Linguistics.

Pearson, P. D., Kamil, M. L., Mosenthal, P. B., & Barr, R. (Eds.). (2016). *Handbook of reading research*. New York, NY: Routledge.

Rickards, J. P., & August, G. J. (1975). Generative underlining strategies in prose recall. *Journal of Educational Psychology, 67*(6), 860-865. doi:10.1037/0022-0663.67.6.860

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic keyword extraction from individual documents. *Text Mining: Aapplications and Theory, 1,* 1-20.

Siadaty, M., Gasevic, D., Jovanovic, J., Pata, K., Milikic, N., Holocher-Ertl, T., Jeremić, Z., Ali, L., Giljanović, A., & Hatala, M. (2012). Self-regulated workplace learning: A Pedagogical framework and semantic web-based environment. *Educational Technology & Society, 15*(4), 75-88.

Stefanou, C., Lord, S. M., Prince, M. J., & Chen, J. C. (2014). Effect of classroom gender composition on students' development of self-regulated learning competencies. *International Journal of Engineering Education, 30*(2), 333-342.

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (pp. 242-264). Hershey, PA: IGI global. doi:10.4018/978-1-60566-766-9.ch011.

Van Horne, S., Russell, J. E., & Schuh, K. L. (2016). The Adoption of mark-up tools in an interactive e- textbook reader. *Educational Technology Research and Development, 64*(3), 407-433.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*, 5998-6008.

Yang, S. J. H. (2019). Precision education: New challenges for AI in education [conference keynote]. In *Proceedings of the 27th International Conference on Computers in Education (ICCE)* (pp. XXVII-XXVIII). Kenting, Taiwan: Asia-Pacific Society for Computers in Education (APSCE).

Yin, C., Yamada, M., Oi, M., Shimada, A., Okubo, F., Kojima, K., & Ogata, H. (2019). Exploring the relationships between reading behavior patterns and learning outcomes based on log data from e-books: A Human factor approach. *International Journal of Human-Computer Interaction*, *35*(4-5), 313-322.

Yue, C. L., Storm, B. C., Kornell, N., & Bjork, E. L. (2015). Highlighting and its relation to distributed study and students' metacognitive beliefs. *Educational Psychology Review, 27*(1), 69-78.

Yufan, X., Xuewang, G., Li, C., Satomi, H., Yuta, T., Hiroaki, O., & Yamada, M. (2020). Can the area marked in ebook readers specify learning performance. In *Proceedings of the 10th International Conference on Learning Analytics & Knowledge (LAK20)* (pp. 638-648). Frankfurt, Germany: ResearchGate.