The Effects of Retrieval Practice Across Levels of Thinking and Retention Interval on Reading Comprehension

Corrin Alicia Nero[1], Norehan Zulkiply[2*]

[1, 2] Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak
94300 Kota Samarahan, Sarawak, Malaysia
aliciacorrin96@gmail.com
znorehan@unimas.my
[*]Corresponding Author

https://doi.org/10.24191/ajue.v17i4.16222

Received: 30 August 2021
Accepted: 30 September 2021
Date Published Online: 31 October 2021
Published: 31 October 2021

Abstract: The present study examined the effect of different types of retrieval practice on reading comprehension across levels of thinking and retention interval in a classroom setting. One hundred undergraduates divided into two retention interval groups (short- and long-retention interval) were asked to read a passage on a topic in Cognitive Psychology and were then required to engage in a retrieval practice learning strategy using the two types of question format (production test and recognition test) and different levels of thinking (lower-order thinking and higher-order thinking). A three-way mixed ANOVA statistical test was used to analyse the data and found no significant differences in reading comprehension across the different types of retrieval practice, suggesting that the performance when using the recognition test is equivalent to when using the production test. The difference in reading comprehension between the different types of retention interval also was not observed, indicating that students in the short-retention interval group retained just as much information as those in the long-retention interval group. Additionally, the present study observed a significant difference in students' reading comprehension between different levels of thinking, signifying that the students' performance for the lower-level thinking questions was better than that for the higher-level thinking questions. The present finding contributed to the existing body of knowledge in which it suggested that the performance in reading comprehension when using a recognition test, particularly a well-constructed one, with competitive alternatives was equivalent to when using a production test.

Keywords: Levels of thinking, Question format, Reading comprehension, Retention interval, Retrieval practice

1. Introduction

The test in educational settings is conventionally used as a tool in the assessment of learning (Moreira et al., 2019; Brame & Biel, 2015) as well as to provide a summarised picture of an individual's knowledge (Stenlund et al., 2016). However, a growing number of studies have revealed that a test can do more than just measure one's knowledge; that is, it can boost learning as well as promote long-term memory retention. When learners engage in the practice of retrieving previously studied information, it can result in greater long-term memory retention compared to merely rereading or restudying. This phenomenon is known as the testing effect (Carpenter et al., 2006; McDaniel et al, 2007) or sometimes it is referred to as the retrieval practice effect (Karpicke & Roediger, 2007).

The explanation behind the effectiveness of a test is based on the retrieval effort hypothesis that is driven by the desirable difficulties framework which states that a more difficult but successful

retrieval has more beneficial effects on memory compared to easier, successful retrievals (Pyc & Rawson, 2009). A study to test the hypothesis was conducted by Pyc and Rawson (2009) to examine and theoretically discuss the effects of retrieval practice in promoting memory retention by using a cued-recall test of 70 Swahili-English word pairs. The study was conducted by manipulating two variables — interstimulus interval (ISI, defined as the number of items between each following practice trial) and criterion level (the number of correctly retrieved items). Findings from the study revealed that participants' performance in the final test increased when the difficulty of retrieval increased. As for the second manipulated variable (criterion variable), it was revealed that when the criterion level increased, retrieval became less difficult, thus resulting in a decrease in the final test performance. These findings therefore confirmed the assumptions on the retrieval effort hypothesis which states that successful but difficult retrievals are more advantageous in enhancing memory as opposed to successful but less difficult retrievals. Prior to Pyc and Rawson's study (2009), a study by Gardiner et al. (1973) also provided evidence that retrieval difficulty can enhance memory retention. Participants in their study were tested on 50 definition-word pairs and findings showed that the difficult retrieval of words in the initial test was significantly better than in the final recall test, thus indicating that greater retrieval effort yields greater benefits in memory. With regard to the desirable difficulty framework from which the retrieval effort hypothesis is derived, the framework suggests difficult retrievals are more desirable than easy ones (Pyc & Rawson, 2009). Based on the desirable difficulty framework, learning strategies that require more effort and are cognitively demanding during the encoding or retrieval practice phase as compared to activities that require minimal cognitive demand (i.e., restudying or rereading) may hinder short-term recall but can assist in constructing the necessary networks in memory to improve long-term recall (Bjork, 1994).

There are two commonly used tests or retrieval practice formats in examining testing effect — the recognition test and production test (Larsen & Butler, 2013; Butler & Roediger, 2007; Greving & Richter, 2018). The production test is usually in the form of free recall, short- answer essay and fill-in-the-blank, whereas the recognition test is in the form of multiple-choice questions (MCQ) and "true or false" (Larsen & Butler, 2013). Moreira et al. (2019) noted that the differences with regard to the effectiveness of the different types of test in drawing out the testing effect remains a question for both laboratory and classroom contexts. In short, some studies reported that tests which involve a production task would provide greater benefit than a recognition task (Butler & Roediger, 2007; Greving & Richter, 2018) whereas some studies reported otherwise (Little et al., 2012; Smith & Karpicke, 2014). In a recent study by Greving and Richter (2018) which examined the testing effect in a university lecture, the participants were asked to either answer short-answer questions, MCQs, or read summarising statements about core lecture content (restudy). The retention of the learning content was measured at different times (1 week, 12 weeks and 23 weeks after the last lecture) and the results revealed that short-answer testing benefitted learning in higher education contexts more than multiple-choice testing, independent of the time of test (Greving & Richter, 2018). Previous findings which found that the short-answer test is more advantageous than MCQs are supported by the retrieval effort hypothesis which suggests that a more difficult and effortful retrieval process results in greater benefits (Gardiner et al., 1973; Pyc & Rawson, 2009). In contrast, Smith and Karpicke (2014) found there was little or no advantage for answering short-answer questions over MCQs. This finding emerged due to unsuccessful retrievals when using the short-answer test, whereby participants seemed to perform better in MCQ than in short-answer testing for most of the time. They explained that although a more difficult test is said to be more beneficial, success of retrieval is equally as important for later performance (Smith & Karpicke, 2014). Additionally, Little et al. (2012) discovered that properly constructed MCQs, that is, including competitive incorrect options in the choices of answers can result in productive retrieval processes. However, this study was done in a short-term basis and one of the questions raised was whether the benefits of multiple-choice testing can be retained for a longer period of time.

In terms of retrieval practice and levels of thinking, there have been an increasing number of studies which have examined the effects of retrieval practice on different levels of thinking (i.e., lower-order thinking and higher-order thinking). Particularly, these studies examined whether transfer of knowledge from one level to another level (e.g., lower-level thinking to higher-level thinking or vice versa) would occur if students were instructed to engage in a retrieval practice learning strategy (McDaniel et al., 2013; Dobson et al., 2018; Agarwal, 2019). Findings from these studies showed that although the levels of thinking are usually hierarchical, superior performance on lower-level thinking

tasks does not necessary enhance performance in higher-level thinking tasks (McDaniel et al., 2013; Dobson et al., 2018; Agarwal, 2019). In these studies, levels of thinking are categorised according to *The Taxonomy of Educational Objectives*, a framework published by Bloom, Engelhart, Furst, Hill and Krathwohl in 1956 (Agarwal, 2019), otherwise known as the Bloom's Taxonomy. A revision on the Bloom's taxonomy was made in 2001 and based on the revised taxonomy, higher-order thinking falls under the *apply*, *analyse*, *evaluate*, and *create* categories, whereas lower-order thinking falls under the *remember* and *understand* categories (Anderson et al., 2001; Agarwal, 2019). Looking at most previous studies on retrieval practice and levels of thinking however, there has yet to be a study that particularly investigates the effect of different types of question format used in retrieval practice (production test versus recognition test) on different levels of thinking. For instance, most studies that observed the effect of retrieval practice on levels of thinking usually examined the effect of the transfer of learning and most of the time they used only one question format, which was either a production test or a recognition test.

With regard to the presence of a retrieval practice effect on retention interval, numerous studies have shown that in comparison to rereading or restudying, retrieval practice either in the form of a short-answer test, a multiple-choice test or a free recall has greater benefits for long-term memory retention. (Carpenter et al., 2006; Karpicke & Roediger, 2007; McDaniel et al., 2007). As for the effects of type of test on retention interval, most studies however, did not observe the testing benefits across different retention intervals, that is, they either observed over a short-retention interval or a long-retention interval. One study that observed the testing effect using different types of tests on different retention intervals was conducted by Stenlund et al. (2016). It was revealed that retrieval practice using short-answer items was more beneficial for long-term retention. On the other hand, a study by Little et al. (2012) found that a recognition test is also capable of eliciting productive retrieval processes. However, it observed the effect on a short-retention interval and therefore, it remains unclear whether the benefits of multiple-choice testing can be retained for longer period of time.

With all the varying findings reported in past studies, further exploration is needed to investigate the effects of retrieval practice across levels of thinking and retention interval. Across a plethora of studies on retrieval practice effect over the years, it is a widely established fact that retrieval practice enhances retention compared to rereading or restudying (Carpenter et al., 2006; McDaniel et al., 2007; Rowland, 2014; Karpicke, 2017). Although there is convincing evidence of the effects of retrieval practice relative to restudying or rereading, the generalisability to instructional settings and the circumstances in which the effects arise is still up for discussion (Greving & Richter, 2018). The few studies that have looked into the different types of test (production test vs. recognition test) have produced mixed findings, where some studies found that tests which involved a production task would provide a greater benefit than those that used a recognition task (Butler & Roediger, 2007; Greving & Richter, 2018) whereas others did not (Little et al., 2012; Smith & Karpicke, 2014). Additionally, there is a need to further explore the issue concerning retrieval practice and levels of thinking because there has yet to be a study that particularly investigates the effect of the different types of question format used during retrieval practice on levels of thinking. Considering the student's development in higher-order thinking is an important element of education, educators and scientists recently have been actively developing strategies and instructional approaches that can increase higher-order thinking (Agarwal, 2019). By further examining the effects of different question formats on levels of thinking, it can shed light on how educators can design learning materials that will improve the students' learning experience either for lower-order thinking questions or higher-order thinking questions. With respect to the effects using different types of test format across different retention intervals, most studies did not observe the testing benefits across different retention intervals; that is, they either observed on a short-retention interval or a long-retention interval. A study by Stenlund et al. (2016) found that retrieval practice using production task has more beneficial effects for long-term retention. Conversely, a study by Little et al. (2012) revealed that the multiple-choice test (recognition task) resulted in productive retrieval processes when compared to the cued-recall test (production task), of which is contrary to the popular belief that the production task brings greater benefit than the recognition task. However, the study observed the effect over a short-retention interval; therefore, it remains uncertain whether the benefits of multiple-choice testing can be retained for a longer period of time.

2. Method

2.1 Participants

A total of 201 participants among the first-year students enrolled in the Cognitive Psychology course subject was involved in the study. To ensure that the experiment was equitable, pre-screening of the participants was done based on their Malaysian University English Test (MUET) result. Students with MUET Band 3 and above were selected to participate in the study. MUET Band 3 was set to be the minimum requirement for participation to ensure that participants had at least a moderate level of English proficiency so as to reduce the chances of language proficiency affecting their performance. Out of 201 students who were initially recruited, 100 students were identified to meet the MUET requirement, therefore participated in the actual experimental manipulation of this study.

2.2 Design

The experiment used a 2 (Retrieval practice: production test, recognition test) x 2 (Levels of thinking: higher-order thinking, lower-order thinking) x 2 (Retention Interval: short-retention interval, long-retention interval) mixed-subjects design. Retention interval was varied between subjects whereas retrieval practice phase and levels of thinking were varied within subjects. The experiment included the following phases: a) Study phase: participants were asked to read a passage on Problem Solving, b) First distractor task: participants were asked to do a word search puzzle, c) Retrieval practice phase: participants were asked to answer 10 recognition test questions (MCQ) and 10 production test questions (short-answer) on the reading passage. Each of the ten questions consisted of five lower-order thinking questions and five higher-order thinking questions. For this phase, participants were asked to answer a total of 20 questions, d) Second distractor task: participants were asked to do a word search puzzle, and e) Final test: participants were required to answer a total of 20 multiple-choice questions (MCQ), comprising 10 lower-order thinking questions and 10 higher-order thinking questions, similar to the questions in the retrieval practice phase.

2.3 Materials

The materials used in the present study are discussed according to the phases of the study.

2.3.1 Study phase

The material for the study phase consisted of a reading passage on one of the chapters in the Cognitive Psychology course content — Problem Solving. The length of the reading passage was 980 words which was appropriate for university students (e.g., Agarwal, 2019; Little & Bjork, 2012).

2.3.2 Retrieval practice phase

In this phase, participants were asked to restudy the passage using retrieval practice that comprised both production and recognition tests. Then, participants were asked to answer a total of 20 questions. The production test used 10 short-answer questions, whereas the recognition test used 10 multiple-choice questions (MCQs). In both the production and recognition tests, the questions asked tested the participants' comprehension on the reading passage given earlier. Specifically, each of the 10 questions in each test (production and recognition) consisted of five lower-order thinking questions and five higher-order thinking questions. Based on the findings by Little et al. (2012) which revealed that well-constructed MCQs can elicit a productive retrieval process, the present study therefore decided to follow the same method, that is by including competitive alternatives in the MCQs to ensure that the MCQs cannot be attributed to being relatively easier than short-answer questions. All the questions used in the present study were validated by the course instructor (an expert in Cognitive Psychology) to

ensure that each of them matched the level of thinking (LOT vs. HOT) required. As for the scoring, MCQs were scored with one point for each correct response and zero point for any incorrect response. For the short- answer questions, a grading rubric was created by consulting the course instructor. A fully correct response that contained keywords or phrases based on the rubric was graded with one point whereas half a point was given for a partially correct response. Zero points were given for any incorrect response or no response.

2.3.3 Final test phase

A total of 20 MCQs were asked. They comprised 10 lower-order thinking questions and 10 higher-order thinking questions, similar to those asked during the retrieval practice phase but with slight changes. The structure of the questions in this phase was slightly modified but the context of the questions remained similar to the ones in the retrieval practice phase (see sample of questions used in Appendix A). MCQ was chosen for the final test format because there have been past studies that observed a testing effect when MCQ was used as the final test (Dobson & Linderholm, 2015; Carpenter et al., 2016). Furthermore, the MCQ type was chosen in the final test because it merely acts to evaluate participants' performance based on the different types of retrieval practice administered to them. Therefore, the types of format used for the final test does not matter (Karpicke, 2017). Moreover, there were previous studies on testing effect that used different formats for both the retrieval phase and final test phase (see Dobson & Linderholm, 2015; Carpenter et al., 2016). Additionally, several literatures have also noted that the format of the initial test did not have to match the format of final test for testing benefits to occur (McDermott et al., 2014; Blunt & Karpicke, 2014; Karpicke & Blunt, 2011; Karpicke, 2017). The scoring for the MCQs in the final test followed the ones used in the retrieval practice phase — each question was scored with one point for a correct response and zero points for any incorrect response.

2.3.4 Distractor task phase

The word search puzzle was used as the distractor task. The level of difficulty of the puzzle was moderate and suitable for university students. The function of this task was to clear participants' working memory and to prevent them from mentally rehearsing the studied material.

2.4 Procedure

The experiment followed a mixed-subjects design in which the participants were randomly assigned into two retention interval groups (i.e., 50 participants in the short-retention interval group and 50 participants in the long-retention interval group). In each group, participants were exposed to all the conditions (retrieval practice phase and levels of thinking). All participants were given a consent form for them to sign as an indication that they agreed to participate in this study. Participants were informed that they could opt out before, during, and after the experiment for any reason. Given that the experiment was conducted using online tools, all participants were given an integrity declaration form for them to sign as an indication that they would not refer to sources from the Internet or use other assistance during the course of the experiment. Once they had given their consent, the participants were briefed on what they were required to do. After the briefing, the procedures for the experiment took place. These involved five phases: study phase, first distractor task, retrieval practice phase, second distractor task and final test phase. For the study phase, participants were asked to read one reading passage on Problem Solving (a chapter in the Cognitive Psychology course content) which was presented on screen using Google Doc. They were informed that they would be tested on the information. However, they were not given the details of the test. The participants were given 15 minutes to read the passage. Once the 15 minutes time limit was over, participants were asked to solve a word search puzzle task in the first distractor task phase for 1 minute. The purpose of this phase was to clear out the participant's working memory. After the 1-minute distractor task finished, the next phase, the retrieval practice phase, involving the manipulation of the retrieval practice phase (production test and recognition test) took place. Participants were asked to answer 10 recognition test questions (MCQ) and 10 production test questions (short-answer questions) on the reading passage. The questions were presented using Google

Form Quiz. Each of the 10 questions consisted of five lower-order thinking questions and five higher-order thinking questions. For this phase, participants were required to answer a total of 20 questions within 15 minutes. Next, the participants were asked to undertake the second distractor task in which they had to solve another word search puzzle for 1 minute. Once the 1-minute time limit for the task was up, the next phase was the final test phase. Similar to the retrieval practice phase, the questions in the final test phase were also presented using Google Form Quiz. Participants in the short-retention group were tested immediately after the second distractor task whereas those in the long-retention interval group were tested after a delay of three weeks. All participants had to answer a total of 20 MCQs for the final test phase. After completing the final test phase, the participants were debriefed and allowed to dismiss. The whole experiment took approximately 60 minutes.

3. Results

A three-way mixed ANOVA statistical test was used to analyse the collected data, mainly focusing on the final test performance to measure the participants' performance based on the different types of retrieval practice phases administered to them. The results showed that there was no significant difference in students' reading comprehension across the different types of retrieval practice (production test vs. recognition test), $F(1, 98) = 1.31$, $p = .255$. Although the difference in reading comprehension across different types of retrieval practice was not significant, the mean scores of students using the production test were slightly higher ($M = 2.83$) than those for the recognition test ($M = 2.72$), indicating that students must have benefitted a little more from the production test compared to the recognition test.

There was also no significant difference in students' reading comprehension observed between the different types of retention interval (short-retention interval vs. long-retention interval), $F(1, 98) = 1.30$, $p = .257$, suggesting that students' performances in reading comprehension between the two groups were no different from each other. Based on the mean scores however, the students' reading comprehension performance in the short-retention interval were slightly higher ($M = 2.88$) compared to the performance in the long-retention interval ($M = 2.67$).

In terms of difference in students' reading comprehension between lower-order thinking and higher-order thinking, results showed that there was a significant difference, $F(1, 98) = 9.88$, $p = .002$, indicating that by using retrieval practice during the retrieval practice phase regardless of the formats used, students' performance in reading comprehension was significantly better for lower-order thinking questions ($M = 2.93$) compared to higher-order thinking questions ($M = 2.62$).

As for the interaction levels, the results showed there was no significant interaction between types of retrieval practice and levels of thinking, $F(1, 98) = 0.02$, $p = .877$, signifying that the effect of different types of retrieval practice on reading comprehension was almost similar for both lower- and higher-order thinking questions. The mean scores however showed that test performance for lower-order thinking was slightly better when using the production test during the retrieval practice phase ($M = 2.98$) compared to using the recognition test ($M = 2.88$). Likewise, for the higher-order thinking, the mean score for the production test was slightly higher ($M = 2.68$) than that of the recognition test ($M = 2.55$).

The results also revealed that there was no significant interaction between types of retrieval practice and retention interval, $F(1, 98) = 0.06$, $p = .804$, indicating that the effect of different types of retrieval practice on reading comprehension was almost similar for both short- and long-retention intervals. While there was no significant interaction effect observed between types of retrieval practice and retention interval, for those in the short-retention interval group, the mean score for the production test was a little higher ($M = 2.92$) than that of the recognition test ($M = 2.83$). Similarly, for those in the long-retention interval group, the mean score for production test was slightly higher ($M = 2.74$) compared to the recognition test ($M = 2.60$).

Furthermore, there was also no interaction effect observed between retention interval and levels of thinking, $F(1, 98) = 0.90$, $p = .345$, suggesting that the effect of different retention intervals on reading comprehension was equivalent for both lower- and higher-order thinking. Although no significant interaction effect was observed for the lower-order thinking, those in the short-retention interval group scored slightly higher ($M = 3.08$) than those who were in the long-retention group ($M = 2.78$). The same

applies for the higher-order thinking in which those in the short-retention interval group fared better (*M* = 2.67) than those in the long-retention interval group (*M* = 2.56).

In terms of the interaction effect between the types of retrieval practice, levels of thinking and retention interval, the results showed that there was no significant interaction between these three variables, $F(1, 98) = 0.003$, $p = .959$, signifying that the effects of the types of retrieval practice and levels of thinking on student's reading comprehension was similar in both short- and long-retention intervals. Although no significant interaction was observed, Table 1 shows that regardless of the retrieval practice formats used, students in both the short- and long-retention interval groups seemed to perform slightly better in lower-order thinking questions (*M* = 3.04, *M* = 3.12, *M* = 2.72, *M* = 2.84) than in higher-order thinking questions (*M* = 2.62, *M* = 2.72, *M* = 2.48, *M* = 2.64). Also, by comparing the scores between different types of retrieval practice groups, the students seemed to collectively perform a bit better when using the production test compared to using the recognition test. Additionally, by comparing the scores from the two groups (short-retention interval and long-retention interval), it can be seen that those in the short-retention interval group scored slightly higher than those in the long-retention interval group.

Table 1. Means and standard deviation of interaction between types of retrieval practice, levels of thinking and retention interval

Retention interval	Types of retrieval practice	Levels of thinking	*M*	*SD*
Short-retention	Recognition test (MCQ)	Lower-order thinking	3.04	1.19
		Higher-order thinking	2.62	1.21
	Production test (SA)	Lower-order thinking	3.12	1.21
		Higher-order thinking	2.72	1.26
Long-retention	Recognition test (MCQ)	Lower-order thinking	2.72	1.03
		Higher-order thinking	2.48	1.36
	Production test (SA)	Lower-order thinking	2.84	1.33
		Higher-order thinking	2.64	1.32

4. Discussion

With regard to the effect of different types of retrieval practice on reading comprehension, the statistical analysis revealed that there was no significant difference in students' reading comprehension across the different types of retrieval practice (production test vs. recognition test), indicating that the students' performance in reading comprehension using the recognition test (MCQ) was almost similar to the performance when using the production test (short-answer questions) in the retrieval practice phase. This finding is in line with a few studies which found that short-answer and MCQs are equally effective in enhancing the retention of studied materials (McDermott et al., 2014; Smith & Karpicke, 2014). Particularly, Smith and Karpicke's (2014) study which investigated the effects of different question formats on learning (short-answer, MCQ or hybrid questions) found that there was little to no difference in students' performance across the three different formats. They reasoned that although a more difficult test is said to be more beneficial, success of retrieval is equally as important for later performance. Following their explanation, the findings from the present study thus indicated that both the production and the recognition test resulted in similar retrieval success in students' reading comprehension performance. Another plausible reason as to why no significant difference was observed in students' reading comprehension across different types of retrieval practice is because the recognition test questions in the present study included competitive alternatives. Following Little et al.'s (2012) study which revealed that properly constructed MCQs (i.e., questions which include competitive incorrect choices of answers) are also capable of eliciting productive retrieval processes; therefore, it was possible that the reading comprehension performance when using the recognition test in the present study was on par with using the production test because the recognition test questions were almost as difficult as those used in the production test. In terms of interaction level, the present study found no significant interaction between types of retrieval practice and levels of thinking, indicating that the

effect of different types of retrieval practice on reading comprehension was almost similar for both lower- and higher-order thinking questions. Additionally, there was also no significant interaction observed between types of retrieval practice and retention interval, signifying that the effect of different types of retrieval practice on reading comprehension was almost similar for both short- and long-retention intervals. Looking at the mean scores on the other hand, results showed that the mean score for the production test was slightly higher than that of the recognition test, suggesting that the production test may have provided an added advantage than the recognition test. The mean scores also showed that for both lower- and higher-order thinking questions, students scored slightly higher when using the production test during the retrieval practice phase compared to when students used the recognition test. Also, for both short- and long-retention interval groups, the students' seemed to perform slightly better when they used the production test instead of using the recognition test in the retrieval practice phase. In the present study, the production test involved the production of answers which requires more effort than the recognition test. Therefore, following the desirable difficulty framework, when learners engage in a more difficult but successful retrieval process, it leads to more beneficial effects on memory compared to easier, successful retrievals (Pyc & Rawson, 2009).

With regard to the effect of retention interval on reading comprehension, the results showed that the difference in students' reading comprehension between the different types of retention interval (short-retention interval vs. long-retention interval) was not statistically significant, suggesting that students' performances in reading comprehension between the two groups were equivalent to each other regardless of the types of retrieval practice format and levels of thinking. Based on the mean scores on the other hand, it can be seen that students' reading comprehension scores in the short-retention interval group were slightly higher than those in the long-retention interval group, indicating that the students in the short-retention interval group (immediate testing) were able to recall slightly more information compared to those in the long-retention interval group (delayed testing). Such a finding was likely due to the fact that some students in the long-retention interval group may have forgotten what they had learnt prior to the final test which took place after a delay of three weeks. The present study also observed no significant interaction effect between retention interval and levels of thinking, suggesting that the effect of different retention interval on reading comprehension was equivalent for both lower- and higher-order thinking. Despite no significant interaction effect between the two variables, the mean scores revealed that for both lower- and higher-order thinking questions, the students scored slightly higher in their reading comprehension when the effect was observed under the short-retention interval compared to the long-retention interval, implying that the testing effect was slightly more evident when the students were tested immediately after the retrieval practice phase than when they were tested after a delay. These findings which revealed that those in the short-retention interval group scored slightly better than those in the long-retention interval group were of no surprise as the students in the long-retention interval group may have forgotten some of the information learnt before they took the final test. The reason an individual forgets is because their newly developed memories have yet to have the opportunity to strengthen; therefore, these memories are susceptible to the intrusion of mental activity and mental information (Baddeley & Hitch, 1993; Ferreira et al., 2019). Furthermore, the slight difference between the mean scores of those in the short- and long-retention interval groups can be supported by the recency effect whereby when one takes a test and applies the information learnt immediately, there is a high probability of remembering that information than in the situation where the information is presented the day before (Ferreira et al., 2019). Additionally, while it is true that retrieval practice can improve long term memory retention relative to restudying or rereading (Carpenter et al., 2006; Karpicke & Roediger, 2007; McDaniel et al., 2007), when the retrieval practice effect is concerned, it is evident that the effect is greatest when it is observed in the short-term period (Roediger & Karpicke, 2006; Karpicke, 2017; Ferreira et al., 2019).

As for the effect of levels of thinking on reading comprehension, the present study found a significant difference in students' reading comprehension between lower-order thinking and higher-order thinking, wherein the students' performance in reading comprehension was significantly better for lower-order thinking questions as compared to higher-order thinking questions. From the mean scores, it was evident that the students' performed better in lower-order thinking questions compared to higher-order thinking questions. One plausible explanation for the mean scores for lower-order thinking questions appearing to be slightly higher than the scores for higher-order thinking questions could be due to the fact that the students have not yet mastered the elements of higher-order levels in

the particular topic, given that they were unfamiliar with the topic and had yet to study the topic thoroughly in class. Therefore, some information in the reading passage may have seemed novel to them. Based on Bloom's taxonomy, in order to reach the higher category, an individual has to acquire the cognitive processes of the lower category (Agarwal, 2019). In short, the levels of thinking according to Bloom's taxonomy are said to be hierarchical (Bloom et al., 1956; Dobson et al., 2018). Thus, it is possible that the students were only able to master the lower-level elements during the retrieval practice phase, causing a marginally better performance for lower-order thinking questions than that of higher-order thinking questions.

With respect to the interaction between the types of retrieval practice, levels of thinking and retention interval, the present study found no significant interaction between these three variables, indicating that the effects of the types of retrieval practice and levels of thinking on student's reading comprehension was no different in both short- and long-retention intervals. Although the present study did not observe any significant interaction between types of retrieval practice, levels of thinking and retention interval, the mean scores reported in Table 1 showed that regardless of the retrieval practice formats used, students in both the short- and long-retention interval groups seemed to perform slightly better in lower-order thinking questions than higher-order thinking questions. This finding can be attribute to the long-time belief where factual knowledge must come before skills (Bruner, 1977). Considering that the students have not yet learned the topic explicitly in class and some information in the reading passage may have seemed new to them, they only managed to acquire the cognitive processes in the lower category during the experiment, resulting in slightly better performance for lower-order thinking questions. Furthermore, by comparing the scores between different types of retrieval practice groups for both levels of thinking and retention interval groups, students fared slightly better when using the production test in the retrieval practice phase compared to using the recognition test, signifying that using short-answer tests leads to slightly more successful retrievals than when using multiple-choice tests. Numerous studies in the past have found that the production test, particularly short-answer questions is more beneficial than MCQs because it requires a more concerted retrieval than the latter (Greving & Richter, 2018; Stenlund et al., 2016). Furthermore, by comparing the scores from the two groups (short-retention interval and long-retention interval), it can also be seen that those in the short-retention interval group scored slightly higher relative to those in the long-retention interval group, suggesting that there was a slightly greater immediate testing effect compared to delayed testing effect in both the types of retrieval practice and the levels of thinking. As mentioned earlier, this is probably due to the recency effect where students who take a test and apply the information learnt instantly are prone to recall that particular information better than the information presented the day before (Ferreira et al., 2019).

5. Conclusion

In conclusion, the present study extends past findings on retrieval practice in the classroom setting, particularly examining the effects of different types of retrieval practice across levels of thinking and retention interval on reading comprehension. The present study found no significant differences in reading comprehension across the different types of retrieval practice. This finding contributed to the existing body of knowledge in which it is suggested that the performance when using a recognition test, particularly a well-constructed one, with competitive alternatives is equivalent to when using a production test. A difference in reading comprehension between the different types of retention interval also was not observed, implying that students in the short-retention interval group retained just as much information as those in the long-retention interval group. Additionally, the present study observed a significant difference in students' reading comprehension between different levels of thinking, whereby the students' performance in reading comprehension was significantly better for lower-order thinking questions than higher-order thinking questions. This indicated that even by using retrieval practice, it was of no surprise that when students learned new items, they mastered the lower elements of learning before mastering the higher elements.

Future work can extend the present study by examining the transfer of knowledge between different levels of thinking using different types of retrieval practice format and observing the effect on different retention intervals. The research on transfer of knowledge has been the focus of many recent studies (Dobson et al., 2018; Agarwal, 2019) to examine whether transfer of knowledge from one level

to another level (e.g., lower level- to higher-level thinking or vice versa) would occur if students use retrieval practice as the learning strategy. Thus, it will be noteworthy to know if the transfer of knowledge is dependent on the type of retrieval practice format used and to see whether the effect can be retained for a longer period of time. As well, future work can also extend the present study by including students' reading anxiety scale in the observation. A recent study by Rahmat et al. (2020) which examined the influence of students' fear and perceived difficulties in academic reading found that teaching method was one of the reasons undergraduate students fear reading and become most anxious when they have difficulty in understanding the content of the text they read. Given that the study also suggested to explore the relationship of reading strategies used and reading anxiety, therefore, by employing retrieval practice as a learning strategy in classroom, it would be interesting to know if it contributes to any difference in students' reading anxiety.

6. Acknowledgements

This research was supported by the Special Top Down Grant (F04/SpTDG/1776/2018) from Universiti Malaysia Sarawak.

7. References

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order thinking? *Journal of Educational Psychology*, *111*(2), 189-209. https://doi.org/10.1037/edu0000282

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives* (abridged ed.). New York, NY: Addison Wesley Longman.

Baddeley, A. D. & Hitch, G. (1993). The recency effect: Implicit learning with explicit retrieval? *Memory* & *Cognition*, *21*(2), 146-155. https://doi.org/10.3758/BF03202726

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205). Cambridge, MA: MIT Press.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *The taxonomy of educational objectives: The classification of educational goals* (Handbook 1: Cognitive domain). New York, NY: David McKay Company.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. https://doi.org/10.1037/a0035934

Brame, C. J., & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science course. *CBE—Life Sciences Education*, *14*(2), 1-12. https://doi.org/10.1187/cbe.14-11-0208

Bruner, J. S. (1977). *The process of education*. Cambridge, MA: Harvard University Press.

Butler, A. C., & Roediger, H. L. III (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514-527. https://doi.org/10.1080/09541440701326097

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, *28*(2), 353-375. https://doi.org/10.1007/s10648-015-9311-9

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*(5), 826-830. https://doi.org/10.3758/BF03194004

Dobson, J. L., & Linderholm, T. (2015). Self-testing promotes superior retention of anatomy and physiology information. *Advances in Health Sciences Education, 20*, 149–161. https://doi.org/10.1007/s10459-014-9514-8

Dobson, J., Linderholm, T., & Perez, J. (2018). Retrieval practice enhances the ability to evaluate complex physiology information. *Medical Education*, *52*(5), 513-525. https://doi.org/10.1111/medu.13503

Ferreira, R., Sierra, V. S., & Vega, S. (2019). Immediate testing is more beneficial than delayed testing when learning novel words in a foreign language. *Revista signos: estudios de lingüística*, *52*(100), 290-305. http://dx.doi.org/10.4067/S0718-09342019000200290

Gardiner, F. M., Craik, F. I., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, *1*(3), 213-216. https://doi.org/10.3758/BF03198098

Greving, S., & Richter, T. (2018). Examining the testing effect in university teaching: Retrievability and question format matter. *Frontiers in Psychology*, *9*, 2412. https://doi.org/10.3389/fpsyg.2018.02412

Karpicke, J. D. (2017). Retrieval-based learning: A decade of progress. In J. H. Bryne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., pp 487-509). Oxford, England: Academic Press.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772-775. https://doi.org/10.1126/science.1199327

Karpicke, J. D., & Roediger, H.L., III, (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. https://doi.org/10.1037/0278-7393.33.4.704

Larsen, D. P., & Butler, A. C. (2013). Test-enhanced learning. In Walsh, K. (Ed.), *Oxford textbook of medical education* (pp. 443-452). Oxford: Oxford University Press.

Little, J. L., & Bjork, E. L. (2012). The persisting benefits of using multiple-choice tests as learning events. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 683–688). Austin, TX: Cognitive Science Society.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: Fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science*, *23*, 1337–1344. http://dx.doi.org/10.1177/0956797612443370

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14*, 200-206. https://doi.org/10.3758/BF03194052

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*, 360–372. http://dx.doi.org/10.1002/acp.2914

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L. III, & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied, 20*(1), 3–21. https://doi.org/10.1037/xap0000004

Moreira, B. F. T., Pinto, T. S. S., Starling, D. S. V., & Jaeger, A. (2019). Retrieval practice in classroom settings: a review of applied research. *Frontiers in Education, 4(*5), 1-16. https://doi.org/10.3389/feduc.2019.00005

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437-447. https://doi.org/10.1016/j.jml.2009.01.004

Rahmat, N. H., Arepin, M., & Sulaiman, S. (2020). The cycle of academic reading fear among undergraduates. *Asian Journal of University Education*, *16*(3), 265-274. https://doi.org/10.24191/ajue.v16i3.9730

Roediger, H. L. III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rowland, C.A., 2014. The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. Psychol. Bull. 140 (6), 1432–1463. https://doi.org/10.1037/a0037559

Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory*, *22*(7), 784-802. https://doi.org/10.1080/-09658211.2013.831454

Stenlund, T., Sundström, A., & Jonsson, B. (2016). Effects of repeated testing on short-and long-term memory performance across different test formats. *Educational Psychology*, *36*(10), 1710-1727. https://doi.org/10.1080/01443410.2014.953037

Appendix A

Levels of thinking	Questions	
	Retrieval practice phase	Final test phase
Lower order thinking	(Recognition test – MCQ) What is involved when one is experiencing insight during the process of problem solving? A. He applies a methodical process to find the solution to the problem. *B. He suddenly becomes aware of the solution to the problem.* C. He can find more than one solution. D. He compares the current solution to the previous one.	(MCQ) What is involved when one experiences insight during the process of problem solving? *A. The individual suddenly becomes aware of the solution to the problem.* B. The individual can find more than one solution. C. The individual applies a methodical process to find the solution to the problem. D. The individual compares the current solution to the previous one
	(Production test – Short Answer) "Occurs when there is an obstacle between a present state and a goal, and the solution to get around the obstacle is not immediately obvious". The above description refers to …	(MCQ) "The solution to get around the obstacle is not immediately obvious". The above description refers to … A. Representation B. Analogy C. Restructuring *D. Problem*
Higher order thinking	(Recognition test – MCQ) "A car starts from rest and accelerates uniformly over a time of 5.21 seconds for a distance of 110 m. Determine the acceleration of the car." The above mentioned problem demonstrates *A. A well-defined problem* B. Problem solving involving insight C. An ill-defined problem D. Problem solving involving restructuring	(MCQ) "A man's speed with the current (rate of movement in the water) is 15 km/hr and the speed of the current is 2.5 km/hr. The man's speed against the current is?" The above mentioned problem demonstrates A. an ill-defined problem B. problem solving involving insight C. problem solving involving restructuring *D. a well-defined problem*
	(Production test – Short Answer) "Mental illness is expected to be the second biggest health problem affecting Malaysians after heart disease. Based on the latest National Health and Morbidity Survey, every three in 10 adults aged 16 years and above in Malaysia suffer from some form of mental health issues. Like any parts of the globe, mental illness is a rising issue in the country and should be taken seriously." What type of problem is this? Can you elaborate on the reason why you give that answer?	(MCQ) "Mental illness is expected to be the second biggest health problem affecting Malaysians after heart disease. Based on the latest National Health and Morbidity Survey, every three in 10 adults aged 16 years and above in Malaysia suffer from some form of mental health issues. Like any parts of the globe, mental illness is a rising issue in the country and actions should be taken to solve this problem." Which of the following facts can you gather from the above problem description? A. It is an insight problem and the solution to the problem appears all of a sudden.

		B. It is an ill-defined problem and the path to the solution is usually not clear. C. It is a well-defined problem and the path to the solution is usually clear. D. It is a problem that requires restructuring to find a solution.

Note: The correct response is in italics