Journal of Applied Research in Memory and Cognition 7 (2018) 116–122

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

The Effect of Question Placement on Learning from Textbook Chapters

Oyku Uner*, Henry L. Roediger III

Washington University in St. Louis, United States

Retrieval practice enhances learning of short passages, but its effectiveness for authentic educational materials such as textbook chapters is not well established. In the current experiment, students studied a 40-page textbook chapter on biology. Retrieval practice with correct-answer feedback was manipulated within subjects: some questions appeared only after a chapter section, others only after the whole chapter, and yet others at both times. Two groups served as controls: the reread group read the feedback presented in the retrieval practice condition, and the other group simply read the chapter once. Students took a final test two days later. Practicing retrieval resulted in greater recall relative to the two control groups. On the final test, the two single testing conditions produced comparable benefits, but testing twice produced the greatest benefit. Retrieval practice is effective in learning from authentic text material and placement of the initial test does not matter.

General Audience Summary

In educational settings, testing is typically used to assess knowledge of students; however, research has shown that testing can be a powerful tool to enhance learning. This outcome is referred to as the retrieval practice effect, or the testing effect. Most laboratory studies examining this effect use simple materials, but it is not clear whether testing can be an effective study strategy when students read entire textbook chapters, which is the task faced by many students in introductory courses. Because a textbook chapter is lengthy and complex, a critical issue is where to place practice tests: after each section, after the whole chapter, or both? In the current study, we asked students to study a biology textbook chapter and we tested them two days later with short-answer questions from the chapter. One group of students read the chapter once, another group read the chapter and then reread critical information from the chapter, and a final group read the chapter and answered practice questions on it. The questions could occur after each section, after the entire chapter, or both. We found that answering questions once while reading the chapter increased recall two days later relative to the two control groups. Where the questions were placed did not matter on the final test; however, answering questions twice increased recall more than answering questions once. When studying from textbook chapters, students can use self-testing to improve their grades. Whether they test themselves during reading of the chapter or after reading the chapter does not matter, so long as feedback is provided. To receive the greatest benefit, students should test themselves more than once.

Keywords: Retrieval practice, Testing effect, Learning from text, Question placement

Author Note

Henry L. Roediger III, Department of Psychological & Brain Sciences, Washington University in St. Louis, United States.

We thank the James S. McDonnell Foundation for their financial support in a Collaborative Activity Grant. We also thank Maggie Clapp for her assistance with data collection and data scoring.

* Correspondence concerning this article should be addressed to Oyku Uner, Department of Psychological & Brain Sciences, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63130-4899, United States. Contact: uner@wustl.edu

In educational settings, testing is typically used to assess knowledge but can also be a powerful tool to enhance learning (Roediger & Karpicke, 2006b). Roediger and Karpicke (2006a) have shown in the laboratory that practicing retrieval enhances retention on tests delayed a few days or a week relative to restudying the material. McDaniel, Roediger, and McDermott (2007) further showed that testing can boost performance in academic settings. In fact, a large body of research has shown that testing is an effective technique in both the laboratory and the classroom (see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013, for a review). Many educators are now familiar with and implement retrieval practice in their classes (Wooldridge, Bugg, McDaniel, & Liu, 2014).

The impetus for the present research is to examine how best to use retrieval practice in learning of authentic educational materials. Many of the laboratory studies have used relatively impoverished materials such as paired associates (e.g., Carrier & Pashler, 1992; Kuo & Hirshman, 1996). The usual technique is to provide a test for some subjects or some materials after study and then to measure how the initial testing affects performance on a later test. The oft-obtained finding is that initial testing, especially with feedback, improves performance on the final test relative to either a restudy control condition or a control with only one study phase (see Karpicke, 2017, for a review).

Research with authentic educational materials is not entirely lacking. For example, some researchers have employed lectures and accompanying slides (Butler & Roediger, 2007; Szpunar, Khan, & Schacter, 2013; Weinstein, Nunes, & Karpicke, 2016). Others have used longer text passages to better simulate the tasks students face when studying (Wissman & Rawson, 2015; Wooldridge et al., 2014). Because most authentic educational materials are lengthy and complex, researchers have also explored the use of different testing schedules, interpolating questions during study, rather than testing at the end. Our research focuses on whether interspersed testing during reading of a chapter or testing concentrated after reading the text is more beneficial for subsequent test performance.

Although questions are generally placed at the end of study, some prior research has also shown benefits of interspersed testing. For instance, educational research on adjunct questions (i.e., questions provided within a text passage) has indicated that answering questions shortly after reading a relevant portion of a passage enhances retention of tested information, relative to answering questions before, to restudying the passage, or to not answering any questions at all (e.g., Bruning, 1968; Hamaker, 1986; Rothkopf, 1970). More recently, Szpunar et al. (2013) showed that providing interpolated tests increased note-taking and reduced mind-wandering during learning and enhanced performance on a final test. Thus, adjunct or interpolated questions may not only aid retention of the tested material but may also keep the focus on learning.

Few studies have compared the benefits of answering adjunct or interpolated questions to answering questions at the end of study. Duchastel and Nungester (1984) compared these conditions relative to a restudy control. They found that the tested groups performed better than the restudy group on a test two weeks later, but that placement of the test did not matter on the initial and the delayed tests. The authors argued that the length and complexity of a passage may affect this comparison, as they used a 1700-word passage that was not college-level. They suggested that the potential benefits of inserting questions within a passage over placing them at the end might be observed when the passage is lengthier and more complex.

More recently, Wissman and Rawson (2015) addressed the issue of question placement using relatively brief text passages. Subjects studied passages ranging from 779 to 1333 words, and were asked to recall the passage after retention intervals ranging from 15 min to two days. Critically, some subjects were quizzed after each section of the passage and some were quizzed at the end of the passage. On the initial test, subjects performed better at interpolated recall than recall at the end of the passage, but the benefits of both placements were equivalent on recall on a delayed test.

In the present experiment, we extended prior research regarding placement of retrieval practice by employing lengthier and more complex materials much like a student's typical reading assignment, a textbook chapter. Dividing the study material into smaller chunks might be beneficial in keeping students engaged with the material throughout study (Szpunar et al., 2013). In addition, waiting until the end of reading a chapter might decrease initial retrieval success, potentially reducing long-term benefits of testing unless feedback is given (Pyc & Rawson, 2009; Rawson & Dunlosky, 2011). Although prior research contrasted interpolated questions with questions provided after study, in the natural setting of reading a textbook chapter, such an either/or arrangement need not be used. Students may be best served answering questions both during learning and after learning, because repeated retrieval boosts the testing effect (Karpicke & Roediger, 2008; Pyc & Rawson, 2009). This issue represented another focus of the current research.

In the current experiment, we asked how three schedules of retrieval practice would affect learning of a biology textbook chapter on evolution. We compared interpolated testing, end-of-study testing, or testing on both occasions. We used a read-only condition and a read and restudy condition as controls, neither of which included any tests until the final test. Briefly, subjects studied a 40-page chapter with 18 sections in one of three conditions. One group simply read the chapter (read-only condition). A second group (the tested condition) read the chapter while answering questions under three schedules manipulated within-subjects: questions were given either after each section of the chapter, after the entire chapter, or on both occasions. Correct-answer feedback was given after subjects provided each response. A third group read the chapter and studied the correct answer statements that served as feedback to the questions for the tested group (a reread condition that focused on the critical material that would be tested rather than the whole chapter). All subjects returned two days later to take a final test.

We predicted that the tested group would perform better on the final test compared to the two control groups due to benefits of retrieval practice on long-term retention, the standard testing effect. The primary interest was in the three schedules of testing. We predicted that repeated questions would result in better

performance compared to the single testing conditions on the final test, due to increased benefits of repeated retrieval practice compared to one practice episode (Karpicke & Roediger, 2008; Pyc & Rawson, 2009). We also predicted that the end-of-section questions during learning would result in better initial recall relative to end-of-chapter questions, replicating prior work. Because feedback was given, however, we predicted that the two question placement conditions would show no differential effect on the final test. Weinstein et al. (2016) and Wissman and Rawson (2015) reported similar results while we were conducting our experiment; we discuss these studies after presenting our results.

Method

The experiment used three groups of subjects. The condition of most interest was the testing condition in which the schedule of questions (within chapter, after chapter, both) was manipulated within subjects, with correct-answer feedback on all questions. One control group read the chapter once and the other control group read the chapter and in addition read the feedback statements given to the tested group on the same schedule.

Subjects

Eighty-one Washington University undergraduates from the Psychology subject pool participated in the study. The number of subjects was determined based on the number of subjects used in recent studies in our lab on the testing effect. Subjects received either 0.5 course credits or $5 per 30 min of participation. Five subjects who completed the first session of the experiment did not return for the second session or started the second session late. We included data of 81 subjects (28 in the tested condition) in our analyses of recall in the first session. We had 76 subjects' data for the final test. The study was approved by Washington University's Institutional Review Board.

Materials

Subjects studied a 40-page chapter titled *Evolution and Natural Selection* from a biology textbook (Phelan, 2009, chap. 8). Eighteen out of the original 21 sections were used for experimental purposes; sections 3 and 5 were removed because they were not as biologically oriented as other sections and sections 13 and 14 were combined to make sections of similar length. There were six short-answer questions corresponding to each section, resulting in a total of 108 questions for the chapter. Sample questions are: "Evolution can result from many processes. Name two of them" (answer: mutation, genetic drift, migration and natural selection) and "Penguins and dolphins have flippers but do not have a common ancestor. Their flippers are evidence of what?" (answer: convergent evolution). These questions were taken from the textbook's accompanying database *Prep-U* and had been used in a study by Wooldridge et al. (2014) that used the same chapter. Three questions were created by the first author to provide equivalent numbers of questions per section.

Design

Subjects attended two sessions that were two days apart, and they were randomly assigned to one of three between-subjects conditions during the first session: read-only control ($n = 26$), tested ($n = 28$) or reread control ($n = 27$). The placement of questions was manipulated within-subjects for the tested group, where a third of the questions were placed after sections (section questions), another third occurred at the end of the chapter (chapter questions), and the final third were included both after sections and at the end of the chapter (repeated questions). The dependent measures were percentage of correct answers during the first and the second sessions for the tested group (examining the three testing schedules) and the percentage of correct answers for all groups on the final criterial test that occurred 48 h after the first session.

Procedure

Session 1. Subjects were tested on computers in the laboratory in groups of up to six. They were told that they would be reading a textbook chapter, that they might be asked short-answer questions at various times during their study, and that correct-answer feedback on the questions would be provided. The chapter was presented on the screen one page at a time with 166–718 words on a page (less if it contained a picture, more if there was no picture), in 12-point Bembo font. Each page remained on the screen for at least 20 s, but subjects moved on to the next one whenever they were finished reading the current page on the screen. In the read-only control condition, subjects read the whole chapter with 2-min breaks between sections that were included to equate time with the other two conditions. In these unfilled breaks, subjects sat in front of the screen without engaging in any particular task and the next section appeared on the screen once two minutes were over. This condition serves as a baseline against which to measure rereading and testing.

Subjects in the tested condition read the chapter under the same conditions, but they answered questions after sections and after the chapter instead of taking 2-min breaks. Subjects answered four short-answer questions after each section, two of which were also repeated at the end of the chapter. After reading the chapter, subjects answered 72 questions, half of which were questions they had answered after sections and half of which were new. This arrangement resulted in three different test placement conditions with 36 questions per condition: after sections (section questions), after the chapter (chapter questions), and repeated questions (occurring both after sections and after the chapter). Test placement was counterbalanced such that each question served in all three tested conditions an equal number of times.

The questions were presented on the screen one at a time with a response box underneath for subjects to type in an answer. Each question remained on the screen for at least 10 s, but subjects moved on to the next screen when they finished responding. Subjects could move on if they did not type in an answer. Correct-answer feedback was presented on the screen after each question, and subjects could study the feedback as long as they wanted.

In the reread control condition, instead of answering questions, subjects were presented the statements that served as feedback for the tested group after sections and after the chapter. The design for the reread condition instantiated that for the tested condition: Four statements were reread after each section, two of which were repeated after study of the chapter, and other statements were presented only at the end of the chapter. The presentation of these statements was counterbalanced such that each statement was presented only after sections, after the chapter, or at both times an equal number of times. Thus, the reread control group's exposure to the material was equated to that of the tested group. Subjects studied the statements at their own pace for at least 5 s. The first session took 94 min on average for all subjects to complete.

Session 2. Subjects returned two days later and were tested on the same computer. The second session was the same for all subjects. They answered all 108 short-answer questions that the tested group had answered during the first session. Questions were presented in a random order and no feedback was provided during this session. After answering all questions, subjects were thanked for their participation and were given a debriefing form. The second session lasted 43 min on average.

Scoring

Answers were scored such that they were given either 0 points (incorrect answer), 1 point (partially correct answer), or 2 points (correct answer). All scores were converted to percentages for further analyses. Due to an error in the program, answers for one question were not collected. Therefore, all analyses are based on 107 questions instead of 108. Two raters, who were blind to conditions, scored one-fourth of the answers from the second session (i.e., 107 questions for 19 subjects) and the inter-rater reliability was calculated by correlating scores of each rater. Pearson's r showed good agreement between the two raters ($r = .85$, $p < .001$), and therefore the remaining answers from the second session and all the answers from the first session were scored by one rater, the first author.

Results

We first report initial test performance of the tested group regarding the three question placements (section questions, chapter questions, repeated questions). Then we report data from the delayed test, first comparing retention of the three groups (read-only, tested, reread) and next comparing performance of the tested group on the three types of questions. We last report analyses of time on task.

Comparison of Testing Schedules on the Initial Test

We examined the tested group's initial performance on the three types of retrieval practice questions: questions after each section, after the whole chapter, and at both times. We expected initial test performance to replicate prior findings in showing better recall on end-of-section questions relative to end-of-chapter questions provided for the first time. We thought that questions repeated at the end of the chapter would show a benefit relative to the end-of-chapter questions tested for the first time, due to the earlier test with feedback. We report data from 28 subjects in the tested group who completed the first session. Figure 1 shows that recall in the learning phase was best on the repeated questions ($M = 58.1$, $SD = 12.23$), which was only slightly better than recall on the initial section questions ($M = 56.3$, $SD = 12.54$). Items in both these conditions were recalled better than those occurring for the first time after the chapter ($M = 50.9$, $SD = 11.62$). A one-way repeated measures ANOVA revealed a significant main effect of question placement on recall during the learning phase, $F(2, 54) = 8.34$, $\eta_p^2 = .24$, $p = .001$. Post hoc comparisons using a Bonferroni correction showed significant differences between repeated questions and end-of-chapter questions, $d = .60$, $p = .003$, and between section questions and first time end-of-chapter questions, $d = .45$, $p = .042$. These latter results replicate prior findings where interpolated recall is initially greater than recall at the end of study.

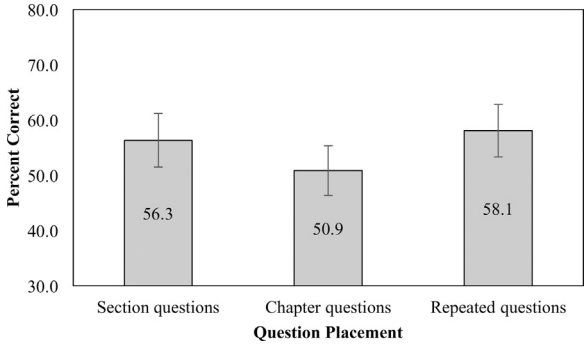

Figure 1. Mean percentage of correct answers of the tested group on the initial test regarding the three question placements. Error bars represent 95% confidence intervals.

Comparison of Groups on the Delayed Test

Next, we examined performance of the three groups (read-only, tested, reread) on the delayed test. We report data from 76 subjects who completed both sessions of the experiment. Figure 2 shows that the tested group performed best ($M = 61.5$, $SD = 13.71$), followed by the reread control group ($M = 44.8$, $SD = 13.10$), which in turn outperformed the read-only control group ($M = 34.2$, $SD = 11.52$). A one-way between-subjects ANOVA revealed a significant main effect of group on delayed

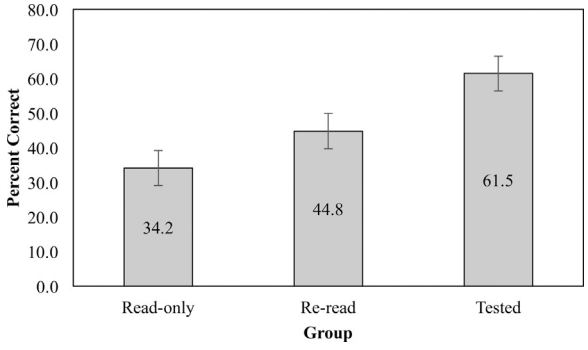

Figure 2. Mean percentage of correct answers on the final test for the three groups. Error bars represent 95% confidence intervals.

test performance, $F(2, 73) = 29.50$, $\eta_p^2 = .45$, $p < .001$. Post hoc comparisons using a Bonferroni correction showed that the tested group performed significantly better than the reread ($d = 1.24$) and read-only ($d = 2.16$) control groups ($p < .001$ for both comparisons) and that the reread control group performed significantly better than the read-only control group ($d = .86$, $p = .013$). The results are standard in the retrieval practice literature, but confirm that retrieval practice enhances long-term recall of a complete textbook chapter.

Comparison of Testing Schedules on the Delayed Test

Of greater interest was the tested group's performance on the delayed test regarding the three types of retrieval practice questions (section questions, end-of-chapter questions, repeated questions). We had predicted better recall after repeated questions on the initial tests compared to the single testing conditions, but possibly similar performance on the section and chapter questions due to the provision of feedback. Two of the 28 subjects in the tested group did not return for the second session, therefore we report delayed recall data from 26 subjects.

Figure 3 shows that final recall was best for the repeated questions condition ($M = 65.4$, $SD = 13.84$) relative to the other two conditions, and that the other two conditions did not differ much from one another ($M = 60.2$, $SD = 15.65$ for the after-section question condition and $M = 58.8$, $SD = 14.20$ for the end-of-chapter question condition). A one-way repeated measures ANOVA revealed a significant main effect of question placement on the delayed test, $F(2, 50) = 8.39$, $\eta_p^2 = .25$, $p = .001$. As predicted, post hoc comparisons using a Bonferroni correction showed that recall on the repeated questions was significantly better than recall for both single testing conditions ($d = .35$ and $p = .01$ for the difference between repeated questions and section questions, $d = .47$ and $p = .001$ for the difference between repeated questions and chapter questions). There was no difference in performance on the section and the chapter questions.

Time on Task

The activities of studying the chapter, answering the questions (if presented) and studying the correct-answer statements (if presented) were self-paced and therefore time spent on the learning phase could affect performance on the delayed test. In our analysis, time spent on the learning phase consisted of reading the text for all three groups along with additional activities specific to each group (unfilled breaks for the read-only group, answering questions and studying the correct-answer statements for the tested group, and studying correct-answer statements for the reread group). The tested group spent the most time on the learning phase ($M = 120$ min, $SD = 26$ min), followed by the read-only control group ($M = 84$ min, $SD = 20$ min), which was followed by the reread control group ($M = 78$ min, $SD = 16$ min). A correlation between time spent on the first session and percentage of correct answers on the delayed test was calculated and Pearson's r showed a significant correlation, $r(74) = .56$, $p < .001$. We also calculated these correlations for each of the groups separately. The correlation between time on task and delayed test was significant only for the tested group, $r(24) = .45$, $p = .021$.

In other studies, time spent on a task, per se, has been shown not to be a critical factor in subsequent memory performance (e.g., Callender & McDaniel, 2009; Rawson & Kintsch, 2005). It seems unlikely that if we had required subjects to spend yet more time reading (but not being tested on) the chapter, they would have learned more. Taking tests and receiving feedback almost certainly were the active ingredients in the additional learning (see too Roediger & Karpicke, 2006b). We acknowledge that the differences among the groups regarding the time spent during the first session may be a limitation, but we believe that the self-paced nature of our experiment better simulates how students study. We are confident that testing, not time on task, is the critical factor producing our results.

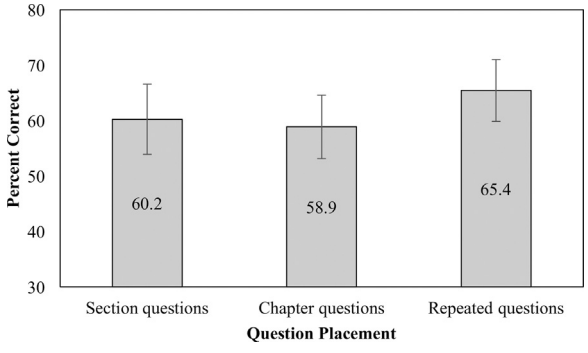

Figure 3. Mean percentage of correct answers of the tested group on the final test regarding the three question placements. Error bars represent 95% confidence intervals.

Discussion

The primary outcome of our experiment is that placement of questions within or after a chapter does not differentially affect retention on a delayed test. We also showed that the questions led to greater learning than control conditions in which subjects read the chapter once or read it once and then reread critical parts that were to be tested later (i.e., the testing effect). Further, we showed that taking two tests—one while reading the chapter and one after reading the chapter—led to better recall than single tests on the final test.

Our findings complement and confirm prior research showing benefits of retrieval practice using authentic educational materials, both with lectures and lecture slides (Butler & Roediger, 2007; Szpunar et al., 2013; Weinstein et al., 2016) and long passages (Wissman & Rawson, 2015; Wooldridge et al., 2014). Thus, testing can be an effective learning strategy for a student's normal means of study: reading textbook chapters.

We compared retrieval practice to both a liberal control condition (reading the chapter only once) and a more conservative control condition (reading the chapter and then rereading facts that were to be tested). Testing on the facts during reading of the chapter, after reading it, or on both occasions revealed benefits compared to both control groups. However, the reread group performed better than the read-only group on the criterial test. Although previous research has shown that rereading

entire chapters does not necessarily enhance learning above and beyond reading once (Callender & McDaniel, 2009), our findings suggest that selectively reading information that would later be tested provides a benefit over simply studying the material once. A key difference in our procedure and that of other researchers is that they had students reread the entire chapter whereas we had students read a set of targeted facts. Thus, no conflict exists between the two sets of findings. Our control condition is not one that could be used by students, because in educational settings teachers are unlikely to tell students exactly what information will be tested on an upcoming exam.

Our primary interest was in the most effective question placement in reading a textbook chapter. We found that answering questions after reading each section and answering questions after reading the entire chapter provided equivalent benefits on the final test. This outcome may seem counterintuitive because of the greater recall on the initial test in the interspersed questions than on the end-of-chapter questions. Typically, higher levels of recall lead to a greater testing effect on delayed tests than do lower levels (see Roediger & Karpicke, 2006b, for a review). However, in our experiment, tests were followed by feedback and this doubtless played a role in equalizing the effects of a single test on the final test. Feedback enables subjects to modify their incorrect answers on a subsequent test (Pashler, Cepeda, Wixted, & Rohrer, 2005) and reinforces their correct answers made with low confidence (Butler, Karpicke, & Roediger, 2008). Previous research has also shown roughly equal benefits from successful retrieval as from failed retrieval with feedback on a later test (Kornell, Klein, & Rawson, 2015). In addition, other researchers have shown test-potentiated learning when feedback is provided after an attempted retrieval (Arnold & McDermott, 2013; Cho, Neely, Crocco, & Vitrano, 2017), relative to conditions like our control in which subjects simply reread the relevant facts.

Our findings regarding question placement replicate those of two recent studies that were published as we conducted our research (Weinstein et al., 2016; Wissman & Rawson, 2015). As discussed earlier, the findings of Wissman and Rawson (2015) suggested that interpolated recall is initially better than recall at the end of study; however, both placements of recall enhanced delayed recall equally well. In three experiments conducted in a laboratory, online, and a classroom, Weinstein et al. (2016) also reported similar findings. In all their experiments, initial tests occurred interspersed throughout learning or after learning, and feedback was provided. Even though performance was initially higher on interspersed questions, the placement of practice tests did not make a difference on the final test that occurred a week (Experiments 1 and 2) or 19 days (Experiment 3) later. The present findings replicate these results using the authentic educational experience of learning from a textbook chapter.

The other major finding from the manipulation of tests is that facts tested both during reading of the chapter and after reading it provided a greater boost to final recall than did a single test. This result adds to prior research showing that repeated testing aids later recall (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2009; Wheeler & Roediger, 1992), although our study was conducted with text materials rather than the simpler materials used in most prior research.

If students aim to achieve durable learning when studying a textbook chapter, where they practice retrieval once does not seem to matter, so long as feedback is provided during these tests. In fact, students will retain more if they practice retrieval both during and after learning, relative to testing only once.

Take-Away Message for Students

The practical advice we offer to students is to create questions while they read the text (a helpful procedure in its own right; see Bugg & McDaniel, 2012) and then stop at the end of sections and test themselves on their own questions. Then, when they reach the end of the chapter, they should review the questions they have generated and the ones that may be provided in the textbook. An important piece of advice is that students should provide themselves with feedback by rereading the relevant section of a chapter if they are uncertain as to whether they answered a question correctly. Of course, this advice relies on students having accurate metacognitive awareness and college students are often overly optimistic in their assessments of how complete their answers are (Dunlosky & Rawson, 2012), so it would be even safer to check all answers. Another tactic would be to write answers in a separate file to permit students giving themselves feedback. Although we have emphasized two practice tests as being better than one, students should repeatedly review important material or points at widely spaced intervals to keep knowledge fresh (Karpicke & Bauernschmidt, 2011; Lindsey, Shroyer, Pashler, & Mozer, 2014) and to maintain an accurate knowledge base (Rawson & Dunlosky, 2011). As shown in research with simpler materials than textbook chapters, repeated retrieval is the key to developing long-term knowledge that can be retained well and easily accessed (Karpicke & Roediger, 2007, 2008).

Conflict of Interest Statement

The authors declare no conflict of interest.

Author Contributions

Both authors conceived and designed the studies. Oyku Uner collected, scored and analyzed the data. Both authors interpreted the data and wrote the manuscript.

References

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: Distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 940–945.

Bruning, R. H. (1968). Effects of review and testlike events within the learning of prose materials. *Journal of Educational Psychology*, *59*(1), 16–19.

Bugg, J. M., & McDaniel, M. A. (2012). Selective benefits of question self-generation and answering for remembering expository text. *Journal of Educational Psychology*, *104*(4), 922–931.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918–928.

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

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*(1), 30–41.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642.

Cho, K. W., Neely, J. H., Crocco, S., & Vitrano, D. (2017). Testing enhances both encoding and retrieval for both tested and untested items. *The Quarterly Journal of Experimental Psychology*, *70*(7), 1211–1235.

Duchastel, P. C., & Nungester, R. J. (1984). Adjunct question effects with review. *Contemporary Educational Psychology*, *9*(2), 97–103.

Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, *22*(4), 271–280.

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

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, *56*(2), 212–242.

Karpicke J. D. (2017). Retrieval-based learning: A decade of progress. In J. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: A comprehensive reference* (J. H. Byrne, Series Ed.) (pp. 487-514) (Chapter 2.27).

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1250–1257.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283–294.

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *The American Journal of Psychology*, 451–464.

Lindsey, R. V., Shroyer, J. D., Pashler, H., & Mozer, M. C. (2014). Improving students' long-term knowledge retention through personalized review. *Psychological Science*, *25*(3), 639–647.

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review*, *14*(2), 200–206.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 3–8.

Phelan, J. (2009). *What is life? A guide to biology* (pp. 1–610) New York: W H Freeman.

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.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, *140*(3), 283–302.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend on time of test. *Journal of Educational Psychology*, *97*(1), 70–80.

Roediger, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255.

Roediger, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210.

Rothkopf, E. Z. (1970). The concept of mathemagenic activities. *Review of Educational Research*, *40*(3), 325–336.

Szpunar, K. K., Khan, N. Y., & Schacter, D. L. (2013). Interpolated memory tests reduce mind wandering and improve learning of online lectures. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(16), 6313–6317.

Weinstein, Y., Nunes, L. D., & Karpicke, J. D. (2016). On the placement of practice questions during study. *Journal of Experimental Psychology: Applied*, *22*(1), 72–84.

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*(4), 240–245.

Wissman, K. T., & Rawson, K. A. (2015). Grain size of recall practice for lengthy text material: Fragile and mysterious effects on memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(2), 439–455.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*(3), 214–221.

Received 28 June 2017;
received in revised form 20 August 2017;
accepted 19 September 2017
Available online 17 October 2017