Journal of Experimental Psychology: Applied

Testing Effects for Self-Generated Versus Experimenter-Provided Questions

Sarah J. Myers, Hannah Hausman, and Matthew G. Rhodes

Online First Publication, August 17, 2023. https://dx.doi.org/10.1037/xap0000487

Journal of Experimental Psychology: Applied

© 2023 American Psychological Association
ISSN: 1076-898X

https://doi.org/10.1037/xap0000487

Testing Effects for Self-Generated Versus Experimenter-Provided Questions

Sarah J. Myers[1], Hannah Hausman[2], and Matthew G. Rhodes[1]
[1] Department of Psychology, Colorado State University
[2] Department of Psychology, University of California, Santa Cruz

Given the finding that retrieval practice improves memory, it is frequently suggested that students test themselves while studying. This study examined whether participants benefit from testing if they create and use their own test questions. In Experiment 1, participants read passages, generated questions about the passages, and then either answered their questions as they created them (the procedure used in previous studies) or after a delay. In Experiments 2 and 3, participants either generated questions and answered them after a delay (i.e., self-testing), answered experimenter-provided questions, or reread the passages before taking a final test administered shortly after learning or following a 2-day delay. The experiments found no benefits of answering one's own questions after a delay. In fact, those who self-tested tended to have worse performance on a final assessment of learning than the other learning conditions. Exploratory analyses suggested that participants' questions often did not target material that was on the later criterion test, which may explain why self-testing was not beneficial. The present study suggests that testing may not benefit learning if students create their own test questions.

Public Significance Statement

Although practice testing is an effective learning strategy, we found that testing by creating and answering one's own review questions (i.e., self-testing) was not beneficial to learning. In the study, self-testing did not lead to better performance on a final test over studied passages compared to answering provided review questions or rereading the passages.

Keywords: self-testing, retrieval practice, generation

The testing effect (i.e., the finding that retrieving information from memory improves memory for that information) has been well-established by previous research with a variety of materials (see H. L. Roediger & Karpicke, 2006a; Rowland, 2014, for reviews). Accordingly, many researchers suggest that students test themselves while studying (e.g., Adesope et al., 2017; Rhodes et al., 2020; Roediger, Putnam, et al., 2011; Yang et al., 2021). In most past studies on the testing effect, participants are given prompts or practice test questions. Within genuine educational settings, however, students are not always provided with practice

Sarah J. Myers https://orcid.org/0000-0003-0939-4571

This work was supported by the National Science Foundation Graduate Research Fellowships under Grant 006784-00002 to Sarah J. Myers. The authors report no conflicts of interest.

Testing one's memory via practice questions or quizzing is a powerful strategy to boost learning and memory of course materials. Therefore, it is often recommended to students that they test themselves while studying. In educational settings, students may write their own practice questions if none are provided. However, few scientific studies have examined whether practice testing is beneficial if students write and answer their own practice questions. We explored this question in the present study. In three experiments, college students read short passages covering scientific topics. They then completed one of four study strategies: (a) writing their own questions and the answer to those questions at the same time, (b) writing their own questions and then answering those questions later from memory, (c) answering provided questions, or (d) rereading the passages. Answering one's own questions from memory was less beneficial to later memory of the passages than answering provided questions or rereading. This suggests that testing may not be beneficial in every situation, and instructors should supplement calls for students to test themselves with materials for students to use for self-testing.

Neither the study design/hypotheses nor the analytic plan were preregistered. Project materials, data, and supplementary analyses are available on the Open Science Framework at https://osf.io/4chs3/?view_only=ed6ee5678fac437b827a6eecd8a09280 (Myers et al., 2023).

Sarah J. Myers played a lead role in conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, validation, visualization, writing–original draft, and writing–review and editing. Hannah Hausman played a supporting role in conceptualization, formal analysis, funding acquisition, methodology, writing–original draft, and writing–review and editing. Matthew G. Rhodes played a lead role in supervision and a supporting role in conceptualization, formal analysis, funding acquisition, methodology, resources, and writing–original draft.

Correspondence concerning this article should be addressed to Sarah J. Myers, Department of Psychology, Colorado State University, Behavioral Sciences Building, 410 West Pitkin Street, Fort Collins, CO 80523, United States. Email: Sarah.Jean.Myers@colostate.edu

test materials (e.g., review questions, quizzes). Under these circumstances, students might craft their own test questions for retrieval practice (i.e., self-testing), but it is unclear whether testing benefits would still be observed. Past research on self-testing (e.g., Bae et al., 2019; Denner & Rickards, 1987; Owens, 1976; Weinstein et al., 2010) has found mixed results and, importantly, has not incorporated opportunities for engaging fully in retrieval practice to answer those questions. Thus, this study examined whether students benefit from self-testing when self-testing involves both generation and retrieval practice.

Effective Learning Strategies Within Self-Testing

Self-testing allows for the combination of several different learning strategies that have been shown to be effective (see Dunlosky et al., 2013; Rhodes et al., 2020, for a review of effective learning strategies). For example, creating questions incorporates elaboration (e.g., Dornisch et al., 2011), summarizing (e.g., Friend, 2000), and an opportunity for generation. The *generation effect* is the finding that retention is superior when participants produce material themselves compared to merely viewing the material (see Mulligan & Lozito, 2004, for a review). For example, words are more likely to be remembered on a subsequent test if they were initially generated from a fragment (e.g., "f_ ie_d") than if they were initially studied (e.g., "friend"; Slamecka & Graf, 1978; Watkins & Sechler, 1988).

Generation effects have been explored with educationally realistic materials as well. For example, some researchers have found that participants who generated headers for a passage learned the passage better than participants who were provided with headers (Brooks et al., 1983; Jonassen et al., 1986). Other research suggests that generating test questions, specifically, can enhance learning (see Song, 2016). For example, Kelley et al. (2019) required psychology students to generate multiple-choice questions for textbook chapters (although they did not answer these questions). Students performed better on a later exam if they wrote practice questions that targeted the same material, compared to exam questions about material not addressed in their review questions (see also Foos et al., 1994; Shakurnia et al., 2018).

In addition to generation, self-testing may also provide an opportunity for students to engage in *retrieval practice* by attempting to answer their generated questions from memory, another effective learning strategy (H. L. Roediger & Karpicke, 2006a). Many studies have demonstrated that retrieving information from memory improves later memory for that material compared to restudying. In a meta-analysis of these studies, Rowland (2014) estimated an overall benefit of testing over restudying of $g = 0.50$ (see also Adesope et al., 2017). The benefits of testing have also been documented with educationally relevant materials, including foreign language vocabulary (Carrier & Pashler, 1992), online videos (Butler & Roediger, 2007), and classroom quizzes (e.g., Carpenter et al., 2009; H. L. Roediger, Agarwal, et al., 2011; see Sotola & Crede, 2021; Trumbo et al., 2021; Yang et al., 2021, for reviews).

Self-testing may afford participants an opportunity for both generation activities (i.e., writing their own practice questions) and retrieval practice (i.e., answering those questions from memory). By allowing for both activities, learning benefits from self-testing could extend beyond those of either generation or retrieval practice alone. Roelle et al. (2022) noted how the cognitive processes involved in generation and retrieval should complement one another. Specifically, generation aids learners in developing an overarching mental representation and integrating new information into their prior knowledge (Fiorella & Mayer, 2016), while retrieval consolidates learned information and makes retrieved memories more durable (Karpicke, 2017). Because generation and retrieval practice increase learning through different cognitive processes, combining the two strategies via self-testing should allow learners to gain additive benefits from both. However, past studies on self-testing may have limited the potential benefits of using both generation and retrieval practice during self-testing.

Past Findings on Self-Testing

Most previous research on self-testing has focused on teaching students how to generate their own questions (see Rosenshine et al., 1996, for a review). However, although advised to self-test (e.g., Adesope et al., 2017), students are typically not given training on how to construct test questions. Only a handful of studies have directly examined the benefits of self-testing without extensive training, and these studies have yielded conflicting results. Some research has found no difference between self-testing and rereading (Bae et al., 2019; Owens, 1976). Other studies find that generating questions is better than rereading (Denner & Rickards, 1987; Weinstein et al., 2010) and incurs similar benefits as answering provided questions (Dornisch et al., 2011; Weinstein et al., 2010). Still other research has reported that generating one's own questions was more beneficial than being given questions (Davey & McBride, 1986; Foos et al., 1994).

Notably, for studies finding that self-testing was more beneficial than answering provided questions, additional benefits of self-testing were limited to questions that could be answered within one sentence of the studied passage (i.e., factual questions; Davey & McBride, 1986; Denner & Rickards, 1987) or test questions that targeted the same material as the self-generated questions (Foos et al., 1994). It seems that only with training on how to write questions do the benefits of self-testing generalize to conceptual final test (FT) questions (Bugg & McDaniel, 2012; Ebersbach et al., 2020).

In all, past work generally suggests that self-testing can be at least as beneficial as answering provided practice questions, particularly for factual information. However, an important methodological decision might have reduced the potential of self-testing as a learning strategy in these past studies. In all past self-testing studies, participants wrote questions and their corresponding answers at the same time. This would provide a generation activity but limit opportunities for retrieval practice (recalling information from memory). As one example, Weinstein et al. (2010) had participants read a passage and then complete one of three study strategies: generating questions and answers over the passage (their generate condition), answering provided questions about the passage (answer condition), or rereading the passage. Those in the generate and answer conditions had access to the passage while completing their practice strategy. Weinstein et al. (2010) argued that this procedure mimics an open-book test, which can still be beneficial to learning (Agarwal et al., 2008). Although this applies to those who answered provided questions, this methodology might have restricted potential learning benefits when answering self-generated questions. To elaborate, participants in the generate condition would have most likely reviewed the passage to choose an important topic, written a question, and then immediately answered it. This would have reduced the opportunity to rely on one's memory of the passage to answer questions. Therefore, participants in the answer condition had more of an opportunity to attempt retrieval (or at least searching through the passage) than those in the generate condition.

Interestingly, in Weinstein et al.'s (2010) study, those in the answer and generate condition performed similarly on a final test, and both outperformed the rereading condition. However, using the reasoning explained previously, those who answered provided questions most likely benefited from retrieval practice to a greater extent than those who generated questions. In contrast, those in the generation group most likely benefited more from the generation activity than those given questions. If those who generated questions had a delay before answering their questions (without the passage), this would presumably allow self-testing to involve not only generation but full engagement in retrieval practice. Indeed, several studies have shown that delaying an initial test (i.e., increasing the time between initial exposure to material and a practice test) increases the benefit of testing compared to taking a test soon after studying (Jacoby, 1978; Rawson et al., 2015; Whitten & Bjork, 1977; see also Kornell et al., 2011). In the present study, one goal was to determine whether adding a delay between having students write their own questions and answering those questions (and requiring them to answer their questions from memory) would allow self-testing to become more beneficial than answering provided questions.

Many prior self-testing studies also incorporated a practice phase whereby participants read a practice passage and answered experimenter-provided questions (e.g., Bugg & McDaniel, 2012; Weinstein et al., 2010). Participants were then told that they should generate questions like those they answered in the practice phase. This practice and instructions might have influenced the types of questions that participants generated. Within classroom contexts, students are unlikely to receive training in generating test questions. Therefore, to mimic students' real-world experiences as much as possible, we did not provide instructions on writing questions or sample questions for participants to view. Given participants' freedom in creating questions, we explored whether certain qualities of their generated questions were associated with their performance on the final test.

The Present Study

The purpose of the present study was to demonstrate how self-testing could be more beneficial if students answered their generated questions after a delay (thus encouraging retrieval practice) rather than generating questions and answers at the same time (the procedure used in, e.g., Denner & Rickards, 1987; Weinstein et al., 2010). Experiment 1 sought to directly compare these two activities to one another (delayed answering vs. simultaneous answering), while Experiments 2 and 3 compared self-testing with delayed answering to other study strategies—answering experimenter-provided questions and rereading. Across the three experiments, we anticipated that full self-testing via generating questions and then answering them later from memory, would be superior to all comparison strategies.

Experiment 1

In Experiment 1, we investigated whether participants benefited more from self-testing if they answered their questions after a delay compared to generating questions and answers simultaneously. Participants initially read two passages, generated questions over each passage, answered their questions (immediately or after a delay), and then completed a final test. Based on findings that testing leads to larger benefits with a longer delay between study and initial test (Jacoby, 1978; Pyc & Rawson, 2009; Rawson et al., 2015; Whitten & Bjork, 1977), participants who answered their questions after a delay should perform better on the final test than participants who generated and answered questions simultaneously.

Method

Transparency and Openness

Materials. Because the passages and tests used in the present experiments were developed by another research group (Thiede et al., 2011), we have not made the materials publicly available.

Data and Analytic Methods. Data are available at https://osf.io/4chs3/?view_only=76aca8996e9c47f09e2da40e7347657a (Myers et al., 2023). Data were analyzed using JASP Version 0.11.1 (JASP Team, 2019) and *R* Version 1.1.414 (R Core Team, 2022). *R* code to reproduce the split-violin plots is available at https://osf.io/4chs3/?view_only=76aca8996e9c47f09e2da40e7347657a. However, analysis code was not available for this version of JASP.

Participants

Participants were 145 undergraduate students from Colorado State University (CSU) who participated in exchange for course credit. Twenty-three participants were removed from analyses due to not completing the experiment ($n = 4$), already seeing the passages ($n = 7$), or not following instructions ($n = 12$). Therefore, data from 122 participants (65 in delayed answering, 57 in simultaneous answering) were used in analyses. A sensitivity analysis using G*Power (Faul et al., 2007) indicated that this sample size was sufficient to detect an effect size of $d = .51$ in a two-tailed, between-groups test. Participants (40 men, 81 women, 1 nonbinary) were between 17 and 27 ($M = 18.92$, $SD = 1.38$) years old. Two participants did not provide their age.

Materials

We used two short passages about monetary policy (549 words) and ice ages (1,052 words), as well as the corresponding final tests, created by Thiede et al. (2011). The final tests comprised 10 multiple-choice questions (five factual and five conceptual) for each passage. Factual questions (e.g., How much of the earth is covered by glaciers during an ice age?) were defined as questions that could be answered using one sentence of the passage, whereas conceptual questions (e.g., What might the Fed do if it wants to affect the economy in a way that is similar to that of lowering income taxes?) required participants to integrate information across two or more sentences and/or make inferences beyond material covered in the passage. Questions were originally categorized as factual and conceptual by Thiede et al. (2011). However, based on the definitions used in the present study, one factual question on the final test was rescored as conceptual for each passage, and one conceptual question for each passage was changed to factual. The experiment took place in Qualtrics and all experiments were approved by the institutional review board before data collection began.

Procedure

After providing consent and answering a demographic survey, participants were randomly assigned to either generate questions and write the answers simultaneously while having access to the passage (simultaneous answering condition) or to generate questions and answer them after a delay without the passage (delayed answering condition). See Figure 1 for a diagram of the procedure.

Participants alternated between Passages 1 and 2 for the different stages of the experiment. First, participants were given 5 min to read one of the passages (order of passages was counterbalanced). For the next 7 min, those in the delayed answering condition generated questions over the first passage while those in the simultaneous answering condition reread the passage. The participants then completed these same activities for Passage 2. After that, participants returned to Passage 1 for additional activities. Specifically, those in the delayed answering condition were shown their generated questions for 7 min and attempted to answer them from memory, without access to the passage. During these 7 min, the simultaneous answering condition saw Passage 1 again and was asked to type in questions and answers over the passage. Then, both groups were provided the passages, their questions, and answers over Passage 1 and were given 5 min to self-score their questions using the passage. This served as an opportunity for feedback (Agarwal et al., 2008). They were asked to type “C” for each question they believe they answered correctly and “I” for each question they believe they answered incorrectly. The participants then completed these activities (generating/answering questions and self-scoring) for Passage 2.

When generating questions over the passages, all participants were given the following instructions:

> You will create your own questions about the passage you just read. Think of this task as if you are creating a quiz to help you prepare for an upcoming exam over the passage.

Participants were only given a blank textbox on the screen to type in their questions. They received no instructions regarding how many or the types of questions (multiple choice or short answer, factual or conceptual, etc.) to write. Participants were also encouraged to guess if they did not know the answer to a question.

After completing these learning activities over the two passages, participants were told that they would take a multiple-choice test on each passage. To assess students’ own interpretations of their learning, they were asked to predict how many of the 10 final test questions they would answer correctly for each passage (providing a global judgment of learning [JOL]).[1] After completing a 3-min distraction phase of solving math problems, participants received a final test over the first passage they read. All participants then completed a final test over the second passage. Questions were presented in a unique random order for each participant and the final tests were self-paced. Because the passages have been used in other experiments at the university, participants were also asked if they had seen these passages before. Participants were then debriefed and released.

Results

We employed both frequentist and Bayesian methods to analyze the data. Analyses include the corresponding p value, a standardized effect size measure (Cohen’s d or η_p^2), and the Bayes factor (BF_{10}). Bayes factors are a ratio of the likelihood of the provided data given the alternative hypothesis (i.e., a difference between conditions) to the likelihood of the data given the null hypothesis (i.e., no difference between conditions). A Bayes factor of 1 means that the data are equally likely under the alternative and null hypotheses. Unlike null hypothesis significance testing, Bayes factors can also indicate that the null hypothesis is more probable than the alternative hypothesis (i.e., when $BF_{10} < 1$) and is reported as the reciprocal ratio, denoted as BF_{01}. Following suggestions from Rouder et al. (2009), we used the Jeffreys-Zellner-Siow prior to calculate Bayes factors because it requires the fewest assumptions about the range of the true effect size.

In addition to reporting multiple analysis metrics, we also formed our data analysis plan based on the questions we were interested and interpret results more qualitatively instead of setting only a p-value criterion. Because one of our research questions involved how self-testing impacted factual and conceptual questions differently (as prior studies had found differences based on type of question; Bugg & McDaniel, 2012; Denner & Rickards, 1987), we added type of final test question as a second variable in addition to study strategy. In addition, we planned to run follow-up t tests even if the Strategy × Type of Question interaction was not statistically significant ($p < .05$).

Judgments of Learning

Analyses of participants’ JOLs and calibration are provided in the supplementary analyses at https://osf.io/4chs3/?view_only=76aca8996e9c47f09e2da40e7347657a. Participants tended to be underconfident in their performance, and predictions did not differ between the two study strategies.

Passage and Order Effects

See https://osf.io/4chs3/?view_only=76aca8996e9c47f09e2da40e7347657a for full analyses. In all experiments, factual monetary policy questions were easier than factual ice age questions, whereas conceptual ice age questions were easier than conceptual monetary policy questions. Minimal differences were found based on the order that passages were read.

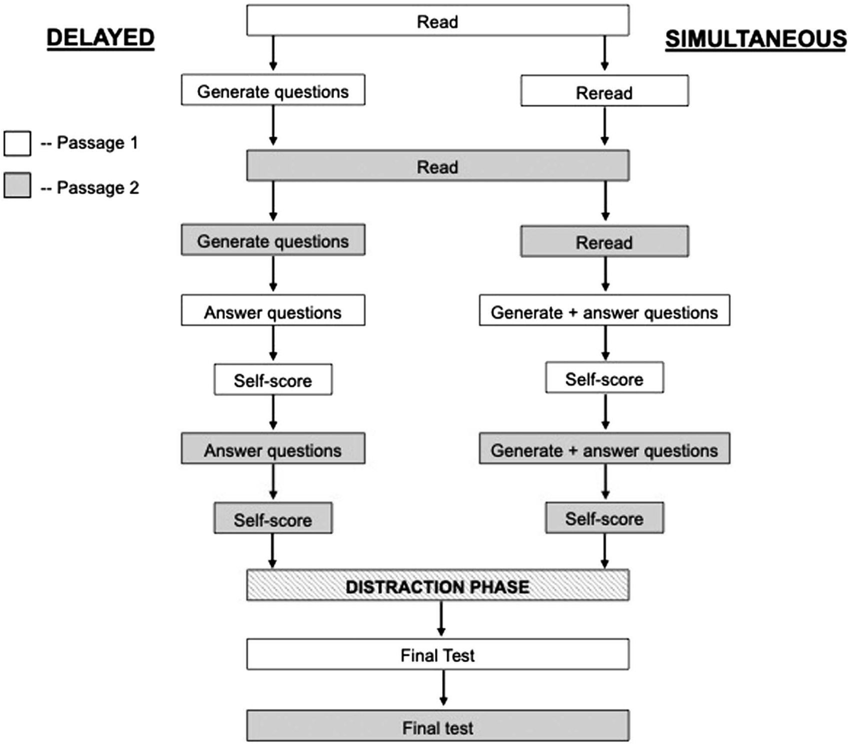
Final Test Performance

Figure 2 presents the average final test performance for the simultaneous and delayed answering conditions. Because some previous self-testing studies have shown differences between conceptual and factual questions (e.g., Bugg & McDaniel, 2012; Davey & McBride, 1986; Denner & Rickards, 1987; see also

[1] Participants provided JOLs after completing the study activities for both passages. Thus, more time had elapsed between their study activities and JOL for the first passage than the second. This was corrected in Experiment 2.

Figure 1

Procedure for the Delayed and Simultaneous Answering Conditions



Note. White boxes represent tasks for Passage 1. Gray boxes represent tasks for Passage 2.

Hausman & Rhodes, 2018), we separated analyses based on these two types of final test questions in the present study.

A 2 (strategy: simultaneous, delayed) × 2 (type of final test question: conceptual, factual) mixed-design analysis of variance (ANOVA) was conducted on final test performance, with strategy manipulated between participants and type of question manipulated within participants. Overall, participants answered more factual questions correctly (M = 79.28, SE = 1.48) than conceptual questions (M = 58.92, SE = 1.41), $F(1, 120) = 149.10$, $p < .001$, $\eta_p^2 = .55$, $BF_{10} = 5.79 \times 10^{20}$. On average, final test performance for participants in the delayed answering (M = 67.77, SE = 1.61) and simultaneous answering (M = 70.44, SE = 1.72) conditions did not differ, with the Bayes factor indicating the null hypothesis was almost four times more likely than the alternative, $F(1, 120) = 1.28$, $p = .26$, $\eta_p^2 = .01$, $BF_{01} = 3.95$. The interaction did not reach conventional significance, and the Bayes factor indicated the null and alternative hypotheses were equally likely, $F(1, 120) = 3.18$, $p = .07$, $\eta_p^2 = .03$, $BF_{01} = 1.20$. Planned comparisons were still conducted to compare performance between the two study strategies for factual and conceptual questions separately. Following recommendations from Wickens and Keppel (2004), we did not adjust the α level from 0.05 because these were planned tests.

Although participants in the simultaneous answering condition slightly outperformed those in the delayed answering condition for factual questions, the difference was not significant, and the Bayes factor indicated the null and alternative hypotheses were equally likely, $t(120) = 1.90$, $p = .06$, $d = 0.35$, $BF_{01} = 1.01$. For conceptual questions, those in the simultaneous and delayed answering conditions did not differ, $t(120) = 0.11$, $p = .91$, $d = 0.02$, $BF_{01} = 5.15$.

Analysis of Generated Questions

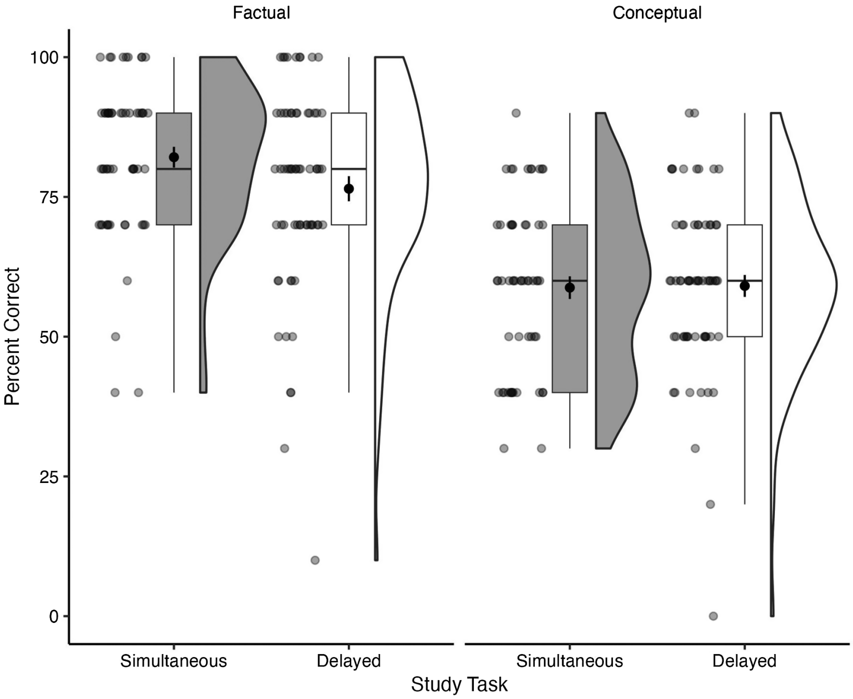
With minimal instructions, participants were given the freedom to create questions in whatever manner they wished.[2] To further explore which factors might make self-testing more beneficial, we categorized several dimensions of generated questions. Most students wrote short-answer questions with only a few multiple-choice and true/false questions generated, so we did not consider the question format. However, there was considerable variation in whether questions were conceptual (C) or factual (F) and for whether the question targeted the same information as a final test question (on final test) or not (not on final test; see Table 1 for example), so we used these as metrics of variability. Participants' accuracy (i.e., whether they answered each of their generated question correctly) was also assessed. A group of scorers rated each generated question on these parameters. See Table 2 for a summary of these measures.

Scorers were trained via descriptive criteria for each score, several practice scoring activities with explanation feedback, and a 10-question test to verify that they scored at least 85% of the sample questions correctly. To assess interrater reliabilities (IRRs) between the two initial scorers, we used Krippendorff's α (Krippendorff,

[2] A sample of students' generated questions can be viewed at https://osf.io/4chs3/?view_only=76aca8996e9c47f09e2da40e7347657a.

Figure 2

Percent of Factual and Conceptual Questions Answered Correctly on the Final Test in Experiment 1 for Those Who Answered Their Questions Simultaneously or After a Delay



1970). Two scorers scored each question, but it was not the same two scorers for every question. Accordingly, we did not use a fully crossed design (Hallgren, 2012), making Krippendorff's α a more appropriate metric (Hayes & Krippendorff, 2007; Krippendorff, 1970). IRRs were lower than desired and outside Krippendorff's (2009) acceptable range of 0.67 (C/F: α = .43, on final test/not on final test: α = .66, accuracy: α = .59). Because reliabilities were lower than desired, all questions were scored by two scorers, and controversies were settled by a third scorer. Krippendorff also acknowledged that these were conservative cutoffs and recognized that acceptable IRR estimates will vary depending on the research question (Krippendorff, 2018; see Hallgren, 2012). Because our analyses were exploratory and were used to understand why self-testing might not have been beneficial (a finding we did not anticipate), we still report findings even though IRRs were low. However, analyses based on the quality of generated questions should be considered exploratory, and conclusions should be taken as tentative. Further, given that we did not manipulate the number of questions or the content of questions, the analyses that follow are correlational and do not identify a causal relationship.

Differences Between Delayed and Simultaneous Answering Conditions. Participants in the delayed answering condition, on average, created more questions than participants in the simultaneous answering condition, $t(120) = 5.33$, $p < .001$, $d = 0.97$, $BF_{10} = 2.88 \times 10^4$. This was expected since those in the delayed answering condition had 7 min to create questions and 7 min to answer those questions, whereas those in the simultaneous answering condition only had 7 min to both create and answer questions. The two conditions did not differ in the percentage of factual questions, $t(120) = 1.57$, $p = .12$, $d = 0.30$, $BF_{01} = 1.70$, or questions that targeted final test material, $t(120) = 1.39$, $p = .16$, $d = 0.25$, $BF_{01} = 2.16$. Although both conditions answered a majority of their questions correctly, those in the simultaneous answering condition were more accurate than those in the delayed answering condition, $t(120) = 3.39$, $p < .001$, $d = 0.62$, $BF_{10} = 30.33$. This was also expected because only participants in the simultaneous answering condition had access to the passage while answering their questions.

Exploratory Regressions. Linear regressions were also conducted to determine whether different characteristics of the generated questions predicted final test performance. Because the two study strategy conditions used different procedures, analyses were run separately for each condition. Final test performance (collapsed across factual and conceptual final test questions) was regressed on the number of questions created, percentage of created questions that were factual (percent factual) and that targeted final test material (percent on final test), and accuracy of answers to created questions. Table 3 presents the unstandardized (b) and standardized (β) regression coefficients, standard error (SE), p value (p), and Bayes factor (BF_{10}) for each model. In the delayed answering condition, the percent of created questions that overlapped with final test material significantly predicted final test performance while controlling for the other factors, such that a 1 percentage-point increase in questions that targeted final test material was associated with a 0.28 percentage-point increase in final test performance. The number of questions created also predicted final test performance, such that creating one additional

Table 1

Sample of Participants' Generated Questions That Were Considered "Overlapping" and the Corresponding Final Test Question

Generated question	Generated answer	Final test question	Final test answer
(C) What are three variables that determine how much the sun's rays will affect the earth's temperature?	Distance from sun to earth, angle of sun's rays hitting earth, and earth's axis tilt	(C) What can cause less solar radiation to reach earth?	The earth's tilt
(C) What is an Ice Age and when was the last known Ice Age?	An ice age is when about a third of the earth is covered with ice and snow; the last known one was over 10,000 years ago.	(F) How much of the earth is covered by glaciers during an ice age?	About a third
(C) What is the purpose of a monetary policy?	A monetary policy overlooks the amount of money in the banks reserve and try to keep it fitting for the current state of the economy	(F) Which of the following does monetary policy affect?	The amount of money circulating in the economy
(C) How does inflation happen?	When people are buying more than there is product	(C) Which of the following is a cause of inflation?	Production cannot keep up with consumer demand

Note. (C) = the question is conceptual; (F) = the question is factual.

question was associated with a 1.5 percentage-point increase in final test performance. Percentage of factual questions and accuracy did not significantly predict final test performance. For those in the simultaneous answering condition, no individual factors significantly predicted final test performance.

Final Test Performance on Overlapping Questions. Writing practice questions that targeted the same material as a final test question was a significant predictor for participants in the delayed answering condition. To further explore this finding, we used a conditional analysis to examine how participants performed on final test questions that were targeted in their practice questions (overlap) versus performance on final test questions that were not addressed in practice questions (no overlap). A 2 (strategy: delayed, simultaneous) × 2 (final test question: overlap, no overlap) mixed-design ANOVA indicated that participants answered more overlapping questions correctly on the final test (M = 87.31, SE = 1.50) than nonoverlapping questions (M = 64.47, SE = 1.50), $F(1, 117) = 138.49$, $p < .001$, $\eta_p^2 = .33$, $BF_{10} = 5.79 \times 10^{20}$. A lack of an interaction, $F(1, 117) = 0.40$, $p = .53$, $\eta_p^2 < .001$, $BF_{01} = 1.20$, suggests that this was true for both the delayed and simultaneous answering conditions, although the Bayes factor indicated the null and alternative hypotheses were equally likely. Thus, performance was better on final test questions that were also addressed in practice questions, suggesting that targeting final test materials during practice makes self-testing more beneficial.

Most other experiments on retrieval practice examine the benefits of testing when initial and final test questions are identical or target the same material, what we refer to as overlapping questions (e.g., Hinze & Wiley, 2011; McDaniel et al., 2013; see Pan & Rickard, 2018). Therefore, we also isolated accuracy on only final test questions that overlapped with participants' created questions to determine whether traditional testing effects (i.e., where practice and final test questions overlap) differed between our conditions. Those in the delayed answering (M = 84.84, SD = 19.76) and simultaneous answering conditions (M = 89.95, SD = 16.34) did not significantly differ in performance on overlapping final test questions, $t(117) = 1.53$, $p = .13$, $d = 0.28$, $BF_{01} = 1.79$.

Discussion

Contrary to our hypotheses, final test performance was largely equivalent between those who answered their generated questions

Table 2

Summary of Measures Regarding Participants' Generated Questions

Strategy	No. of questions/ passage	Percent factual	Percent conceptual	Percent on FT	Percent not on FT	Percent correct	Percent incorrect
Experiment 1							
Delayed	6.95 (2.48)	57.6% (23.6%)	42.4% (23.6%)	35.2% (15.4%)	64.8% (15.4%)	84.2% (16.8%)	15.8% (16.8%)
Simultaneous	4.88 (1.67)	64.5% (25.0%)	35.5% (25.0%)	38.8% (12.1%)	61.2% (12.1%)	93.0% (10.1%)	7.0% (10.1%)
Experiment 2							
Self-test	7.05 (2.59)	55.9% (23.2%)	44.1% (23.2%)	26.3% (13.8%)	73.7% (13.8%)	82.6% (15.9%)	17.4% (15.9%)
Answer	8.00 (0.00)	50.0% (0.0%)	50.0% (0.0%)	50% (0.0%)	50% (0.0%)	67.3% (15.9%)	32.7% (15.9%)
Experiment 3							
Self-test	7.01 (2.45)	65.1% (21.7%)	34.9% (21.7%)	36.0% (15.4%)	64.0% (15.4%)	78.6% (17.9%)	21.4% (17.9%)
Answer	8.00 (0.00)	50.0% (0.0%)	50.0% (0.0%)	50% (0.0%)	50% (0.0%)	73.2% (20.2%)	26.8% (20.2%)

Note. No. of questions/passage = average number of questions per passage participants generated; Percent factual = percentage of factual questions generated; Percent conceptual = percent of conceptual questions generated; calculated as 100 − percent factual. Percent on FT = percentage of questions that targeted material on the final test; Percent not on FT = percentage of questions that did not target material on the final test; calculated as 100 − percent on FT. Percent correct/Percent incorrect = percentage of questions answered correctly and incorrectly (100 − percent correct). Please also note that the questions provided by the experimenter in the answer condition were created so that 50% were factual and 50% targeted material on the final test. Standard deviations are in parentheses. FT = final test.

Table 3

Linear Regressions of Final Test Performance Regressed on Number of Questions Generated, Percentage of Questions That Were Factual and That Targeted Materials on the Final Test, and Accuracy in Answering Questions

Variable	*B*	β	*SE*	*p*	BF_{10}
Exp. 1—delayed (R^2 = 0.21, *p* = .006)					
Number of questions	**1.47**	**0.27**	**0.65**	**.03**	**3.13**
Percent factual	−0.04	−0.06	0.07	.60	0.42
Percent on FT	**0.28**	**0.32**	**0.11**	**.01**	**5.29**
Accuracy	0.14	0.17	0.10	.17	0.86
Exp 1—simultaneous (R^2 = 0.39, *p* = .08)					
Number of questions	0.98	0.13	1.01	.34	0.63
Percent factual	−0.13	−0.26	0.07	.07	1.71
Percent on FT	0.22	0.22	0.14	.12	1.21
Accuracy	−0.25	−0.22	0.15	.11	1.23
Exp 2—self-test (R^2 = 0.35, *p* < .001)					
Number of questions	**1.42**	**0.25**	**0.58**	**.02**	**3.80**
Percent factual	−0.09	−0.14	0.07	.17	0.70
Percent on FT	**0.43**	**0.41**	**0.12**	**<.001**	**46.12**
Accuracy	0.18	0.19	0.11	.12	0.92
Exp 3—self-test (R^2 = 0.58, *p* < .001)					
Number of questions	**2.05**	**0.30**	**0.58**	**<.001**	**43.70**
Percent factual	−0.09	−0.12	0.07	.17	0.61
Percent on FT	−0.02	−0.02	0.12	.85	0.25
Accuracy	**0.52**	**0.56**	**0.11**	**<.001**	**7.11×10^4**

Note. Bold = significant (*p* < .05). Note that BF_{10}'s < 1 provide more support for the null hypothesis. Exp 1 = Experiment 1; Exp 2 = Experiment 2; Exp 3 = Experiment 3; *SE* = standard error; BF = Bayes factor; FT = final test.

after a delay and those who simultaneously generated and answered questions. For those in the delayed answering condition, regression analyses revealed that the number of questions and percentage of questions that targeted final test material were positively associated with final test scores. We will return to these factors in the General Discussion section. No components of the generated questions significantly predicted final test scores for those in the simultaneous answering condition.

We did not anticipate final test performance would be equal between the simultaneous and delayed answering conditions. However, other differences were present between conditions in addition to whether questions were answered immediately or after a delay. First, for participants who wrote and answered their questions at the same time (simultaneous answering condition), more time passed between initial exposure to the passage and writing questions (as well as exposure to the second passage). This might have impacted students' question generation. Those in the delayed answering conditions also had 14 min to generate and answer questions (7 min for generation, 7 for answering), whereas those in the simultaneous answering condition were only given 7 min total to generate and answer their questions. Importantly, those in the simultaneous answering condition also had an extra exposure to the passages compared to those in the delayed answering condition. Those in the delayed answering condition saw the passage for the initial reading, during question generation, and when self-scoring. Those in the simultaneous answering condition, however, saw the passage for the initial reading, a rereading opportunity, generating and answering their questions, and self-scoring. This additional exposure may have boosted learning for those in the simultaneous answering condition. Thus, although future work may seek to further isolate these factors, in Experiments 2 and 3, we shifted our focus to compare self-testing with a delay to other study strategies.

Experiment 2

Experiment 2 sought to determine how generating questions and answering them after a delay (i.e., self-testing; identical to the delayed answering condition from Experiment 1) compared to two other study strategies—answering provided questions and rereading. Although a delay between generating and answering questions did not lead to added benefits of self-testing in Experiment 1, including the delayed self-testing condition in Experiment 2 allowed us to isolate the effects of self-testing by comparing retrieval practice with provided versus self-generated questions.

Experiment 2 therefore compared the three study strategies that were examined in Weinstein et al. (2010), but with an added delay between generating and answering questions for the self-testing condition. Unlike Weinstein et al. (2010), we expected self-testing to enhance learning more than answering experimenter-provided questions because the delay between creating and answering questions should require participants to rely on retrieval.

Method

Participants

Participants were 228 undergraduate students from CSU who completed the experiment in exchange for course credit. Twenty-six participants were removed from analyses because of not completing the experiment (*n* = 2) or having already seen the passages (*n* = 24). Data from 202 participants (68 in self-test, 66 in answer, 68 in reread) were used in analyses. A sensitivity analysis indicated that this sample size was sufficient to detect an effect size of *f* = .22 in a one-way ANOVA, assuming an α of .05, power of .80, and two-tailed test. Participants (79 men, 118 women, two nonbinary, one preferred not to say) were between 17 and 32 (*M* = 19.29, *SD* = 1.97) years old. Two participants' demographic information was not recorded.

Materials

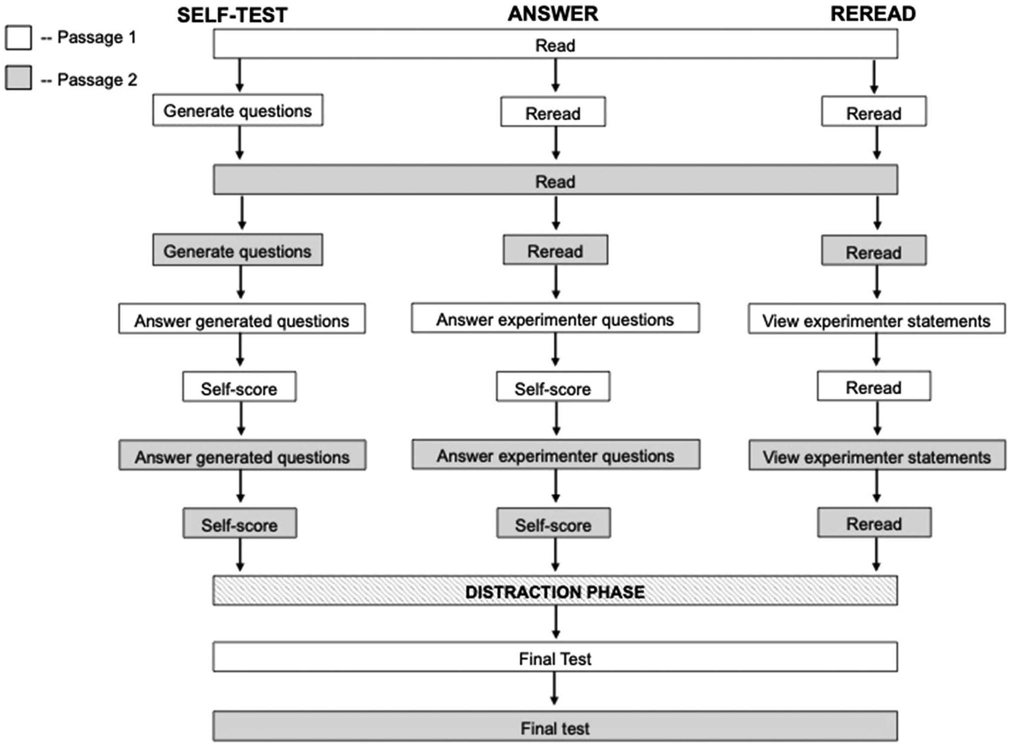
The passages and final tests were the same as in Experiment 1, again administered in Qualtrics. For the answer condition, we developed eight new short-answer questions to use as the experimenter-provided questions (e.g., How much of the earth's land surface can be covered by glaciers during ice ages?). Four of these questions were conceptual, and four were factual. Two of each type of question targeted information that was on the final test and two targeted material that did not appear on the final test, although no initial and final questions were identical. Based on this design, 40% (four of the 10) of final test questions covered materials that were targeted in the provided questions.

Procedure

Figure 3 depicts the procedure. After providing consent and answering demographic questions, participants were randomly

Figure 3

Procedure for the Self-Test, Answer, and Reread Conditions



Note. White boxes represent tasks for Passage 1; gray boxes represent tasks for Passage 2. In Experiment 2, the delay between the distraction phase and final tests was about 15 min; in Experiment 3, the delay was 2 days.

assigned to either the self-test, answer, or reread condition. First, all participants were given 5 min to read one of the two passages. Then, participants in the self-test condition were given 7 min to generate questions about the passage using the same instructions from Experiment 1. To equate exposure time, participants in the answer and reread conditions reread the passage for 7 min. All participants then followed this same procedure for the second passage. Order of passages was counterbalanced across participants.

Next, both the self-test and answer conditions answered questions over the first passage for 7 min, which served as a practice test. The self-test condition answered their own questions, whereas the answer condition answered provided questions. During this phase, participants in the reread condition saw the experimenter-generated questions rewritten as statements (e.g., One third of the earth's land surface can be covered by glaciers during ice ages.). After this stage, participants in the self-test and answer condition had 5 min to score their answers using the procedure described in Experiment 1. Participants in the reread condition reviewed the passage again for 5 min. Next, participants provided a global JOL, predicting how many of the ten final test questions over Passage 1 they would answer correctly. Participants then completed these procedures again for the second passage. Participants followed the same procedure as Experiment 1 for the distractor phase and final tests. Last, participants were asked if they had previously seen either of the study passages and then debriefed.

Results

Judgments of Learning

Participants tended to be underconfident, and metacognitive accuracy was similar between the three study conditions (see https://osf.io/4chs3/?view_only=76aca8996e9c47f09e2da40e7347657a).

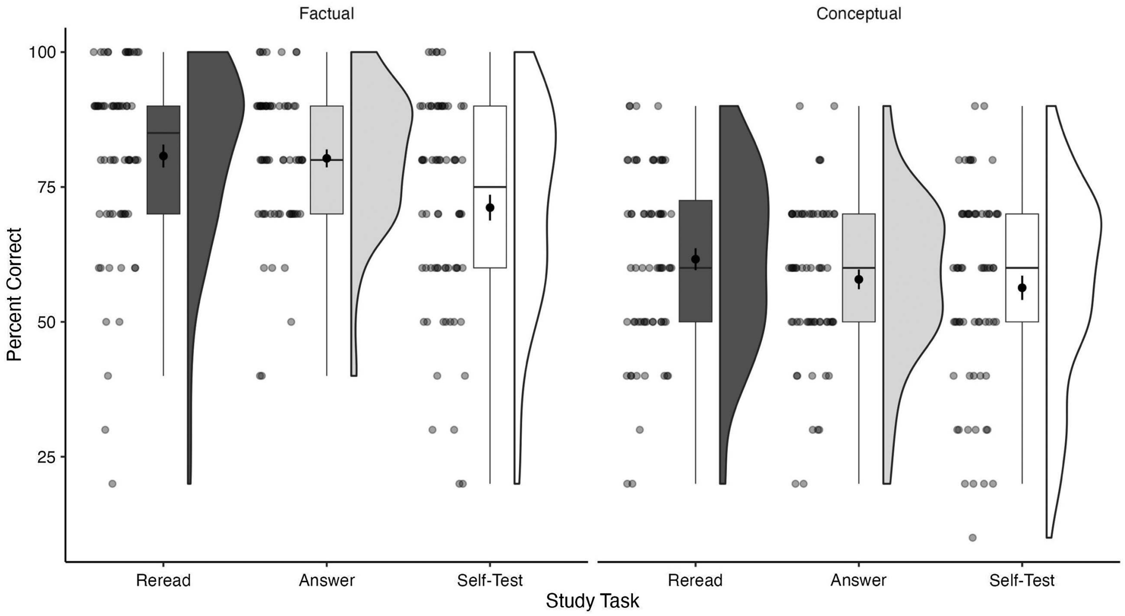
Final Test Performance

A 3 (strategy: self-test, answer, reread) × 2 (type of final test question: conceptual, factual) mixed-design ANOVA was conducted on participants' final test performance (see Figure 4). Participants answered more factual questions correctly ($M = 77.41$, $SE = 1.20$) than conceptual questions ($M = 58.61$, $SE = 1.18$), $F(1, 199) = 194.93$, $p < .001$, $\eta_p^2 = .50$, $BF_{10} = 3.23 \times 10^{28}$. The main effect of study strategy was also significant, and the Bayes factor indicated the alternative hypothesis was twice as likely as the null, $F(2, 199) = 5.10$, $p = .01$, $\eta_p^2 = .05$, $BF_{10} = 2.22$.

The Type of Question × Strategy interaction did not reach conventional significance, and the Bayes factor indicated the null hypothesis was more probable, $F(2, 199) = 2.64$, $p = .07$, $\eta_p^2 = .03$, $BF_{01} = 1.96$. However, planned comparisons were conducted to compare performance between the study strategies for factual and conceptual questions separately. For factual questions, participants in the answer and reread conditions outperformed those in the self-test condition, $t(132) = 3.13$, $p = .002$, $d = 0.54$,

Figure 4

Percent of Factual and Conceptual Questions Answered Correctly on the Final Test in Experiment 2 for Those Who Reread the Passages, Answered Provided Questions (Answer), or Answered Their Own Questions (Self-Test)



$BF_{10} = 14.37$ and $t(134) = 3.00$, $p = .003$, $d = 0.51$, $BF_{10} = 10.17$, respectively. The answer and reread conditions did not differ in factual performance, $t(132) = 0.16$, $p = .87$, $d = 0.03$, $BF_{01} = 5.35$. For conceptual questions, those in the answer condition did not differ from either the self-test, $t(132) = 0.54$, $p = .59$, $d = 0.09$, $BF_{01} = 4.74$, or reread conditions, $t(132) = 1.37$, $p = .17$, $d = 0.24$, $BF_{01} = 2.31$. Although the reread condition numerically outperformed the self-test condition on conceptual final test questions, this difference was not statistically significant, and the Bayes factor indicated both hypotheses were equally likely, $t(134) = 1.75$, $p = .08$, $d = 0.30$, $BF_{01} = 1.36$.

Analysis of Generated Questions

Questions generated by participants in the self-test condition were scored in the same manner as in Experiment 1. In addition, responses participants in the answer condition gave to experimenter-provided questions were scored as correct or incorrect. IRRs were low to moderate (C/F: $\alpha = .39$, on final test/not on final test: $\alpha = .69$, accuracy: $\alpha = .56$ for self-test condition, $\alpha = .76$ for answer condition). All questions were scored by two scorers, and controversies were settled by a third scorer.

Table 2 presents details about the types of questions generated and the accuracy of answering those questions. Similar to Experiment 1, a linear regression indicated that the percent of questions that targeted final test material predicted final test performance such that a 1 percentage-point increase in questions that targeted final test material was associated with a 0.43 percentage-point increase in final test performance. The number of questions generated also significantly predicted final test performance. Percent of factual questions and accuracy in answering self-generated questions did not significantly predict final test performance.

For those in the answer condition, answering more experimenter-generated questions correctly was associated with better performance on the final test ($r = .63$, $p < .001$). An independent-samples t test indicated that those in the self-test condition were significantly more accurate in answering their initial questions than those in the answer condition, $t(132) = 5.07$, $p < .001$, $d = 0.88$, $BF_{10} = 1.09 \times 10^4$.

Final Test Performance on Overlapping Questions

We again examined participants' accuracy on final test questions that were targeted in a practice question/statement compared to final test questions that were not reviewed during practice using a 3 (strategy: self-test, answer, reread) × 2 (final test question: overlap, no overlap) mixed-design ANOVA. Participants answered more overlapping questions correctly ($M = 85.68$, $SE = 1.21$) than nonoverlapping questions ($M = 60.25$, $SE = 1.21$), $F(1, 194) = 314.28$, $p < .001$, $\eta_p^2 = .62$, $BF_{10} = 1.28 \times 10^{43}$. There was no interaction, $F(2, 194) = 1.26$, $p = .29$, $\eta_p^2 = .01$, $BF_{01} = 6.11$, suggesting practice questions or statements that overlapped with final test questions were more beneficial regardless of study condition. Performance did not differ between conditions when final test performance was restricted to only overlapping questions, $F(2, 194) = 0.26$, $p = .77$, $\eta_p^2 = 0.003$, $BF_{01} = 15.27$.

Discussion

Contrary to hypotheses, final test performance was poorest for participants who self-tested relative to those who answered provided questions or reread. Similar to Experiment 1, participants who generated more questions and whose questions overlapped with final test material performed better on the final test.

Experiment 3

Self-testing was less effective than answering provided questions or rereading in Experiment 2. However, typical testing effects (i.e., an advantage for answering provided questions vs. rereading) were not detected in Experiment 2. One potential reason is that the final test was administered only a few minutes after the study activities. Given that testing effects may become more pronounced after longer delays (H. L. Roediger & Karpicke, 2006b; Rowland, 2014), participants in Experiment 3 took the final test 2 days after the study phase. This delay between initial study and the final test also better reflects educational scenarios wherein learning and assessment are usually separated by days or even weeks.

Method

Participants

Participants were 282 undergraduate students from CSU who completed the experiment for course credit. Fourteen participants were removed from analyses because they had already seen the passages (n = 13) or did not follow instructions (n = 1). Data from 268 participants (95 in self-test, 80 in answer, 93 in reread) were used in analyses. A sensitivity analysis indicated that this sample size was sufficient to detect an effect size of f = .19 in a one-way ANOVA. Participants (68 men, 183 women, one nonbinary) were between 17 and 45 (M = 18.96, SD = 2.31) years old. Sixteen participants' demographic information was not recorded.

Procedure

The procedure was the same as Experiment 2 (see Figure 3), except that participants completed the experiment in two parts. In the first portion, participants read the passages and completed the study activities of their corresponding conditions (i.e., generating and/or answering provided questions or reviewing passages). Then, participants were told that they would take a test on the passages in 2 days and were asked, for each passage, how many of the 10 questions they thought they would answer correctly. They then were dismissed from Part 1. Two days later, participants were sent a follow-up email with a Qualtrics link that led to the final test over the two passages (20 questions total).

Results

Judgments of Learning

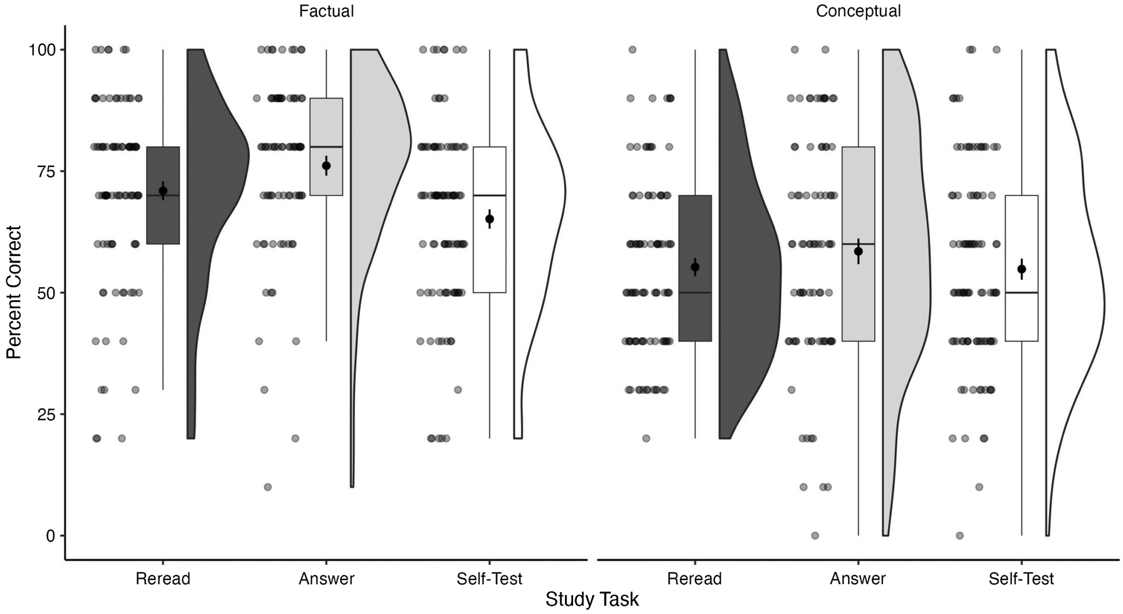
Participants who reread tended to predict they would perform better than they actually did (i.e., were overconfident) and were more confident than those who self-tested and answered provided questions (see https://osf.io/4chs3/?view_only=76aca8996e9c47f09e2da40e7347657a).

Final Test Performance

See Figure 5 for final test performance. A 3 (strategy: self-test, answer, reread) × 2 (type of final test question: conceptual, factual) mixed-design ANOVA indicated that participants answered more factual questions correctly (M = 70.54, SE = 1.22) than conceptual questions (M = 55.99, SE = 1.22), $F(1, 265) = 134.82$, $p < .001$, $\eta_p^2 = .12$, $BF_{10} = 4.06 \times 10^{21}$. The main effect of study strategy, $F(2, 265) = 4.01$, $p = .02$, $\eta_p^2 = .02$, $BF_{10} = 1.35$, and the interaction,

Figure 5

Percentage of Factual and Conceptual Questions Answered Correctly on the Final Test in Experiment 3 for Those Who Reread, Answered Provided Questions, or Self-Tested



$F(2, 265) = 3.08$, $p = .048$, $\eta_p^2 = .01$, $BF_{01} = 1.68$, were also significant.

Although there was a numerical advantage on factual questions for those in the answer condition compared to the reread condition, this difference was not statistically significant, and the Bayes factor was inconclusive, $t(171) = 1.83$, $p = .07$, $d = 0.31$, $BF_{01} = 1.29$. However, aligning with results from Experiment 2, participants in the answer and reread condition outperformed those in the self-test condition on factual questions, $t(173) = 3.85$, $p < .001$, $d = 0.58$, $BF_{10} = 130.21$ and $t(186) = 2.10$, $p = .04$, $d = 0.31$, $BF_{10} = 1.21$, respectively. For conceptual questions, the three conditions did not differ from one another—self-test versus answer: $t(173) = 1.09$, $p = .28$, $d = 0.17$, $BF_{01} = 3.53$, self-test versus reread: $t(186) = 0.15$, $p = .88$, $d = 0.02$, $BF_{01} = 6.25$, and answer versus reread: $t(171) = 1.02$, $p = .31$, $d = 0.16$, $BF_{01} = 3.76$.

Analysis of Generated Questions

Questions generated and answered by participants in the self-test and answer conditions were scored in the same manner. IRRs were low to moderate (C/F: $\alpha = .10$, on final test/not on final test: $\alpha = .56$, accuracy: $\alpha = .67$ for self-test condition, $\alpha = .67$ for answer condition). All questions were scored by two scorers, and controversies were settled by a third scorer.

Table 1 presents details about the types of questions generated, and Table 2 shows linear regressions of which factors predicted final test performance. Number of questions again significantly predicted final test performance. Percent of generated questions answered accurately was also a significant predictor. Contrary to the previous experiments, the percent of questions that targeted final test material did not significantly predict final test performance in Experiment 3. Percent of factual questions was again not a significant predictor.

For those in the answer condition, answering more provided questions correctly was associated with better performance on the final test, $r = .67$, $p < .001$, $BF_{10} = 9.50 \times 10^8$. Although those in the self-test condition numerically answered more of their practice questions correctly than those in the answer condition, this difference was not statistically significant, and the Bayes factor was inconclusive, $t(173) = 1.86$, $p = .07$, $d = 0.28$, $BF_{01} = 1.24$.

Final Test Performance on Overlapping Questions

A 3 (strategy: self-test, answer, reread) × 2 (final test question: overlap, no overlap) mixed-design ANOVA revealed that participants answered more final test questions correctly if it overlapped with practice test questions/statements ($M = 77.89$, $SE = 1.27$) than nonoverlapping questions ($M = 55.40$ $SE = 1.27$), $F(1, 263) = 278.81$, $p < .001$, $\eta_p^2 = .52$, $BF_{10} = 8.36 \times 10^{39}$. A lack of an interaction, $F(2, 263) = 1.29$, $p = .28$, $\eta_p^2 = .01$, $BF_{01} = 7.96$, suggests that this pattern was true for all study conditions.

A one-way ANOVA comparing final test accuracy on overlapping questions did not reach conventional significance, and the Bayes factor somewhat favored the null, $F(2, 263) = 2.75$, $p = .07$, $\eta_p^2 = 0.02$, $BF_{01} = 2.13$. Planned follow-up *t* tests indicated that those in the answer condition ($M = 82.66$, $SD = 20.23$) performed significantly better on overlapping final test questions than those in the reread condition ($M = 75.40$, $SD = 20.81$), $t(171) = 2.32$, $p = .02$, $d = 0.35$, $BF_{10} = 1.94$. Those in the answer condition also performed numerically, but not significantly, better on overlapping questions than those in the self-test condition ($M = 76.05$, $SD = 24.96$), $t(171) = 1.89$, $p = .06$, $d = 0.29$, $BF_{01} = 1.16$. The reread and self-test conditions did not differ, $t(184) = 0.19$, $p = .85$, $d = 0.03$, $BF_{01} = 6.18$.

Discussion

Unlike Experiment 2, Experiment 3 revealed a typical testing effect, but only when the final test performance was isolated to questions that overlapped with the provided practice questions/statements. This comports with past testing effect research in which answering initial questions that are identical or similar to final test questions is more beneficial than rereading (e.g., McDaniel et al., 2013), but practice tests have little benefit for material that is not specifically addressed in the practice test questions (Pan & Rickard, 2018).

The focus of the present study, though, was self-testing. Those who tested themselves with their own questions answered significantly fewer factual questions correctly than those who reread or answered experimenter-provided questions. Thus, self-testing did not lead to testing benefits, even with a longer retention interval (2 days) before the final test. Analyses of the questions generated by participants revealed that number of questions generated and accuracy in answering practice questions were significant predictors of final test performance. Contrary to Experiments 1 and 2, the percent of questions that overlapped final test materials did not significantly predict final test performance.

General Discussion

The present study sought to understand whether students benefit from testing if they create their own test questions, particularly if they answer their generated questions after a delay. In the present experiments, those in the self-test condition were expected to perform better than the comparison study conditions because they would reap benefits from both generation (i.e., writing review questions) and retrieval practice (i.e., answering those questions later from memory). However, contrary to predictions, self-testing did not lead to better learning than any of the comparison conditions. In Experiments 2 and 3, self-testing was even less effective for factual final test performance than answering provided questions ($d = -0.54$ to -0.58) and rereading ($d = -0.31$ to -0.51).

Possible Moderators of the Effects of Self-Testing on Learning

The primary finding of the present study was that self-testing was less beneficial for learning facts from texts compared to answering provided questions or restudying. This finding contradicts past studies in which self-testing did not hinder learning or even benefited learning (Bae et al., 2019; Bugg & McDaniel, 2012; Denner & Rickards, 1987; Ebersbach et al., 2020; Foos et al., 1994; Lehman & Lehman, 1984; Weinstein et al., 2010). To our knowledge, no prior study has found that self-testing led to worse performance compared to rereading. However, there are numerous methodological differences among various studies on self-testing, making it difficult to identify the boundary conditions that predict self-testing benefits or decrements. We highlight four moderators of self-testing below that may guide future investigations.

Difficulty of Self-Generated Questions

Prior to conducting the present research, we posited that one possible moderator of self-testing benefits was the degree to which self-testing involved effortful retrieval of information from memory. Prior research on self-testing typically had participants simultaneously generate and answer their own questions (e.g., Weinstein et al., 2010), which affords little opportunity for retrieval practice. Requiring participants to answer their generated questions after a delay should result in more effortful retrieval (e.g., Rawson et al., 2015). This was supported in Experiment 1, given that those in the delayed answering condition answered fewer practice questions correctly than the simultaneous answering condition. Nevertheless, this added difficulty did not result in stronger final test performance later.

One possible explanation is that the delay of approximately 12 min, although more challenging than answering questions immediately, was still insufficient to make retrieval challenging enough. In addition, it is possible that participants generated questions they typically knew the answers to, potentially reducing the effort required to answer those questions. Although answering self-generated questions after a delay was more difficult, performance was still quite high. Across all three experiments, participants in the delayed self-test conditions answered approximately 80% of their generated questions correctly and were more accurate than participants who answered provided questions (who answered around 70% of the questions correctly).

Future research should examine whether increasing the lag between learning, generating, and answering one's questions would increase the benefits of self-testing, in addition to the difficulty of questions students generate for themselves. Regardless of why final test performance was similar between the two conditions in Experiment 1, it remains unclear why the present study revealed a decrement from self-testing (Experiments 2 and 3), whereas previous research found no effect or even a benefit of self-testing compared to rereading (e.g., Denner & Rickards, 1987; Weinstein et al., 2010).

Study Time

One plausible explanation for the discrepancy between the current and past findings reflects the time limits on review activities. In Weinstein et al. (2010; and other past work), participants were allowed to spend as much time as they wanted to complete their respective study activities. Those who self-tested spent at least twice as long on their activity as those who reread or answered provided questions. In the present study, when study time was controlled, self-testing was no longer as beneficial as the other strategies. Bae et al. (2019) similarly found that generating questions was less beneficial than answering provided questions when study time was controlled (although participants did not answer the questions they generated).

Therefore, the benefits of self-testing may only emerge when participants have sufficient time to generate and answer questions that comprehensively review the learning material. Indeed, in the present study, participants in the self-test condition generated significantly fewer questions (Table 1) than the eight provided questions (Exp. 2: $d = 0.37$; Exp. 3: $d = 0.40$), and the number of questions generated significantly predicted final test performance. Nevertheless, given students' limited time to study in real-world educational settings, answering provided questions or using other review activities might be more time-efficient than writing and answering one's own test questions.

Quality of Self-Generated Questions

Self-testing might also have reduced final test performance due to the quality of questions that students wrote. In the present study, students were not given instructions on how many or what type of questions to create, allowing for considerable variability among the practice questions. Compared to the experimenter-provided questions, participants generated fewer questions, which included a lower percentage of conceptual questions and less overlap with material on the final test. Thus, these measures of question quality suggest that students write lower quality review questions for themselves than an expert or instructor could provide.

Research on generating examples corroborates this possibility. Zamary and Rawson (2018) found that asking learners to generate examples was less beneficial than giving participants examples (see also Hamilton, 1989, 1997; Rawson & Dunlosky, 2016). Zamary and Rawson (2018) concluded that generated examples were of lower quality than the examples provided by the experimenters, which might explain why generating examples was not as effective as other learning strategies. Similarly, there was a strong correlation between the quality of generated examples and performance on the final test (Rawson & Dunlosky, 2016; Zamary & Rawson, 2018), consistent with the regression analyses reported for the current experiments, where some qualities of generated questions predicted final test performance.

This might suggest that training students to create higher quality questions would increase the benefits of self-testing, but it also may reflect individual differences such that better or more motivated participants created better quality questions and performed better on the final test. Consistent with this possibility, other work has failed to demonstrate that providing instruction or scaffolding to generate better quality examples leads to larger learning benefits (Rawson & Dunlosky, 2016; Zamary et al., 2016). Thus, research on effective methods to train students to write more effective practice questions is also necessary.

Beyond the general quality of questions, another factor to consider is the alignment of practice and final test questions. In the present study, students often chose to generate short-answer questions, whereas final test questions were multiple choice, introducing a misalignment between question formats. However, past studies have suggested that the format of practice questions (e.g., short answer, essay, multiple choice) does not need to match between an initial and final test for testing effects to be detected (Carpenter & DeLosh, 2006; McDermott et al., 2014; Rowland, 2014; Yang et al., 2021). A more crucial factor may be whether participants generated questions that targeted material that was present on the final test. For instance, participants in the self-test condition were significantly more likely to answer final test questions that overlapped with the questions/statements that they generated in the practice phase compared to those that did not overlap. Participants in self-testing conditions may have reviewed less material that was relevant to the later test because their generated questions often did not target the same information. Thus, self-testing could be an effective strategy, but not if self-testing is on test-irrelevant material (cf. Pan & Rickard, 2018).

Ebersbach et al. (2020) tested this conjecture directly and found a similar benefit of answering provided questions and self-testing when overlap with the final test was controlled. Students in a psychology course watched a lecture with accompanying slides and then reread the slides, answered provided questions, or generated their own questions. Critically, each slide had one bolded term, and participants were instructed to generate and answer one question or were provided one question on these terms. Final test performance was similar between these two conditions and better than in the reread condition. Thus, one way to increase the benefits of self-testing may be to instruct students on the precise information their questions should target. Indeed, students might find it difficult to differentiate between test-relevant and irrelevant materials on their own (Broekkamp et al., 2002; Gernsbacher et al., 1990; Haynes et al., 2015). For example, Broekkamp et al. (2002) found that students and teachers selected considerably different pieces of information from a passage when asked to select what they thought was "most important." Thus, although instructors may aver practices that seem akin to "teaching to the test," an effective self-testing strategy should likely be informed by understanding what concepts should be targeted.

Perhaps a more favorable approach to increase the benefits of self-testing would be to train students on the types of questions to write rather than identifying the exact pieces of information that they should write questions about (Bugg & McDaniel, 2012; Denner & Rickards, 1987; Weinstein et al., 2010). For example, Weinstein and colleagues had participants, see sample questions on a practice passage and then instructed them to write questions that mimicked these questions. The present study did not provide any guidance on the types of questions participants should write, reflecting experiences in the classroom but at the potential cost of poorly constructed questions (Bae et al., 2019). Taken together, the present research and past work on self-testing suggest that training and specific guidelines might be necessary for students to generate high-quality questions about the correct study material, thereby limiting the practical utility of self-testing in authentic educational settings (Song, 2016).

Type of Final Test Question

The present experiments provide further evidence that the type of final test question—factual or conceptual—might be another moderator of the effects of self-testing on learning. Although self-testing impaired factual final test performance, we found no evidence that self-testing significantly affected performance on conceptual questions, positively or negatively. This is consistent with prior research that has shown little to no benefit of retrieval practice (Pan & Rickard, 2018), pretesting (Hausman & Rhodes, 2018), and self-testing (Bugg & McDaniel, 2012; Denner & Rickards, 1987; but see Ebersbach et al., 2020) for test questions that require inferences. Given the practical importance of learning more than facts, future research should identify ways to modify retrieval practice with provided or self-generated questions to enhance conceptual understanding (e.g., Nguyen & McDaniel, 2016).

In sum, the present study suggests that self-testing does not benefit and, under some conditions, may harm learning from texts. Despite incorporating generation and retrieval practice, self-testing may have been less effective than answering provided questions or restudying because participants' generated questions were of a lower quality.

Practical Implications

The present study suggests that students may fail to benefit from testing if they are not given testing materials (and thus must self-test). Furthermore, creating test questions is a time-consuming activity (Weinstein et al., 2010) and may not confer substantial benefits when time on task is controlled for, as seen in the present study (see also Bae et al., 2019). Therefore, instructors and researchers may need to exercise caution when recommending that students test themselves. We suggest that instructors provide students with practice questions. These need not mimic exam questions but should model the type of knowledge, reasoning, and concepts students will need to demonstrate on the exam (e.g., Thomas & McDaniel, 2007). Indeed, such practice would be consistent with students' belief that it is the obligation of instructors, and not students, to provide practice test questions (Krueger et al., 2023). If instructors are not able to provide practice questions, a better and less time-intensive recommendation to students might be to write down everything they remember about a reading, lecture, or topic (i.e., free recall). Free recall has been shown to benefit learning (Rowland, 2014), including improving memory for key ideas in a text relative to both rereading (H. L. Roediger & Karpicke, 2006b) and generating test questions (Bae et al., 2019; but see Yang et al., 2021). Another possible solution is to provide information about what material will be covered on exams, allowing students to generate more questions relevant to exam material (Ebersbach et al., 2020) or at least provide instruction on how to identify important information (e.g., Williams et al., 2016).

Future self-testing research is also needed in more educationally relevant situations beyond the use of short passages. Students likely have more prior knowledge about and experience with a semester's worth of material, which might allow them to more easily identify important material to target in their questions (although past research suggests students do not select all relevant information from classroom materials; Haynes et al., 2015). Furthermore, motivation to do well on course exams might encourage students to generate better questions and test themselves more diligently (but see Kang & Pashler, 2014). Nevertheless, given our findings that testing with one's own generated questions was not as effective as other strategies, it may be important for instructors to provide practice materials so that students can test themselves effectively.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659–701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*(7), 861–876. https://doi.org/10.1002/acp.1391

Bae, C. L., Therriault, D. J., & Redifer, J. L. (2019). Investigating the testing effect: Retrieval as a characteristic of effective study strategies. *Learning and Instruction*, *60*, 206–214. https://doi.org/10.1016/j.learninstruc.2017.12.008

Broekkamp, H., van Hout-Wolters, B. H., Rijlaarsdam, G., & van den Bergh, H. (2002). Importance in instructional text: Teachers' and students' perceptions of task demands. *Journal of Educational Psychology*, *94*(2), 260–271. https://doi.org/10.1037/0022-0663.94.2.260

Brooks, L. W., Dansereau, D. F., Holley, C. D., & Spurlin, J. E. (1983). Generation of descriptive text headings. *Contemporary Educational Psychology*, *8*(2), 103–108. https://doi.org/10.1016/0361-476X(83)90001-2

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. https://doi.org/10.1037/a0028661

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

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, *34*(2), 268–276. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U.S. history facts. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(6), 760–771. https://doi.org/10.1002/acp.1507

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, *20*(6), 633–642. https://doi.org/10.3758/BF03202713

Davey, B., & McBride, S. (1986). Effects of question-generation training on reading comprehension. *Journal of Educational Psychology*, *78*(4), 256–262. https://doi.org/10.1037/0022-0663.78.4.256

Denner, P. R., & Rickards, J. P. (1987). A developmental comparison of the effects of provided and generated questions on text recall. *Contemporary Educational Psychology*, *12*(2), 135–146. https://doi.org/10.1016/S0361-476X(87)80047-4

Dornisch, M., Sperling, R. A., & Zeruth, J. A. (2011). The effects of levels of elaboration on learners' strategic processing of text. *Instructional Science*, *39*(1), 1–26. https://doi.org/10.1007/s11251-009-9111-z

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. https://doi.org/10.1177/1529100612453266

Ebersbach, M., Feierabend, M., & Nazari, K. B. B. (2020). Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning. *Applied Cognitive Psychology*, *34*(3), 724–736. https://doi.org/10.1002/acp.3639

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Foos, P. W., Mora, J. J., & Tkacz, S. (1994). Student study techniques and the generation effect. *Journal of Educational Psychology*, *86*(4), 567–576. https://doi.org/10.1037/0022-0663.86.4.567

Friend, R. (2000). Teaching summarization as a content area reading strategy. *Journal of Adolescent & Adult Literacy*, *44*(4), 320–329.

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 430–445. https://doi.org/10.1037/0278-7393.16.3.430

Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

Hamilton, R. J. (1989). The effects of learner-generated elaborations on concept learning from prose. *Journal of Experimental Education*, *57*(3), 205–217. https://doi.org/10.1080/00220973.1989.10806506

Hamilton, R. J. (1997). Effects of three types of elaboration on learning concepts from text. *Contemporary Educational Psychology*, *22*(3), 299–318. https://doi.org/10.1006/ceps.1997.0935

Hausman, H., & Rhodes, M. G. (2018). When pretesting fails to enhance learning concepts from reading texts. *Journal of Experimental Psychology: Applied*, *24*(3), 331–346. https://doi.org/10.1037/xap0000160

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Haynes, J. M., McCarley, N. G., & Williams, J. L. (2015). An analysis of notes taken during and after a lecture presentation. *North American Journal of Psychology*, *17*(1), 176–186.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290–304. https://doi.org/10.1080/09658211.2011.560121

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*(6), 639–667. https://doi.org/10.1016/S0022-5371(78)90393-6

JASP Team. (2019). *JASP* (Version 0.11.1) [Computer software].

Jonassen, D., Hartley, J., & Trueman, M. (1986). The effects of learner-generated versus text-provided headings on immediate and delayed recall and comprehension: An exploratory study. *Human Learning*, *5*, 139–150.

Kang, S. H., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition*, *3*(3), 183–188. https://doi.org/10.1016/j.jarmac.2014.05.006

Karpicke, J. D. (2017). *Retrieval-based learning: A decade of progress*. Grantee Submission. https://doi.org/10.1016/B978-0-12-809324-5.21055-9

Kelley, M. R., Chapman-Orr, E. K., Calkins, S., & Lemke, R. J. (2019). Generation and retrieval practice effects in the classroom using PeerWise. *Teaching of Psychology*, *46*(2), 121–126. https://doi.org/10.1177/0098628319834174

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. https://doi.org/10.1016/j.jml.2011.04.002

Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61–70. https://doi.org/10.1177/001316447003000105

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage Publications.

Krippendorff, K. (2009). Testing the reliability of content analysis data. In K. Klaus & M. A. Bock (Eds.), *The Content Analysis Reader* (pp. 350–357). Sage Publications.

Krueger, L., Dyer, J., Schroeder, J., & Carlini, P. (2023). Should students or instructors provide opportunities for testing and why? A mixed methods approach. *College Student Journal*, *56*, 351–357.

Lehman, J. R., & Lehman, K. M. (1984). The relative effects of experimenter and subject generated questions on learning from museum case exhibits. *Journal of Research in Science Teaching*, *21*(9), 931–935. https://doi.org/10.1002/tea.3660210907

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*(3), 360–372. https://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

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. *Psychology of Learning and Motivation*, *45*, 175–214. https://doi.org/10.1016/S0079-7421(03)45005-6

Myers, S. J., Rhodes, M., & Hausman, H. (2023). *Testing effects for self-generated versus experimenter-provided questions*. https://osf.io/4chs3/?view_only=ed6ee5678fac437b827a6eecd8a09280

Nguyen, K., & McDaniel, M. A. (2016). The JOIs of text comprehension: Supplementing retrieval practice to enhance inference performance. *Journal of Experimental Psychology: Applied*, *22*(1), 59–71. https://doi.org/10.1037/xap0000066

Owens, A. M. (1976). *The effects of question generation, question answering, and reading on prose learning* [Unpublished doctoral dissertation]. University of Oregon.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710–756. https://doi.org/10.1037/bul0000151

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

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rawson, K. A., & Dunlosky, J. (2016). How effective is example generation for learning declarative concepts? *Educational Psychology Review*, *28*(3), 649–672. https://doi.org/10.1007/s10648-016-9377-z

Rawson, K. A., Vaughn, K. E., & Carpenter, S. K. (2015). Does the benefit of testing depend on lag, and if so, why? Evaluating the elaborative retrieval hypothesis. *Memory & Cognition*, *43*(4), 619–633. https://doi.org/10.3758/s13421-014-0477-z

Rhodes, M. G., Cleary, A. M., & DeLosh, E. L. (2020). *A guide to effective studying and learning: Practical strategies from the science of learning*. Oxford University Press.

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., III, & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x

Roediger, H. L., III, & Karpicke, J. D. (2006b). 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

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *55*, 1–36. https://doi.org/10.1016/B978-0-12-387691-1.00001-6

Roelle, J., Froese, L., Krebs, R., Obergassel, N., & Waldeyer, J. (2022). Sequence matters! Retrieval practice before generative learning is more effective than the reverse order. *Learning and Instruction*, *80*, Article 101634. https://doi.org/10.1016/j.learninstruc.2022.101634

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, *66*(2), 181–221. https://doi.org/10.3102/00346543066002181

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

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

Shakurnia, A., Aslami, M., & Bijanzadeh, M. (2018). The effect of question generation activity on students' learning and perception. *Journal of Advances in Medical Education & Professionalism*, *6*(2), 70–77.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Song, D. (2016). Student-generated questioning and quality questions: A literature review. *Research Journal of Educational Studies and Review*, *2*(5), 58–70.

Sotola, L. K., & Crede, M. (2021). Regarding class quizzes: A meta-analytic synthesis of studies on the relationship between frequent low-stakes testing and class performance. *Educational Psychology Review*, *33*(2), 407–426. https://doi.org/10.1007/s10648-020-09563-9

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *The British Journal of Educational Psychology*, *81*(2), 264–273. https://doi.org/10.1348/135910710X510494

Thomas, A. K., & McDaniel, M. A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin & Review*, *14*(2), 212–218. https://doi.org/10.3758/BF03194054

Trumbo, M., McDaniel, M. A., Hodge, G. K., Jones, A. P., Matzen, L. E., Kittinger, L. I., Kittinger, R. S., & Clark, V. P. (2021). Is the testing effect ready to be put to work? Evidence from the laboratory to the classroom. *Translational Issues in Psychological Science*, *7*(3), 332–355. https://doi.org/10.1037/tps0000292

Watkins, M. J., & Sechler, E. S. (1988). Generation effect with an incidental memorization procedure. *Journal of Memory and Language*, *27*(5), 537–544. https://doi.org/10.1016/0749-596X(88)90024-1

Weinstein, Y., McDermott, K. B., & Roediger, H. L. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, *16*(3), 308–316. https://doi.org/10.1037/a0020992

Whitten, W. B., II, & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*(4), 465–478. https://doi.org/10.1016/S0022-5371(77)80040-6

Wickens, T. D., & Keppel, G. (2004). *Design and analysis: A researcher's handbook*. Pearson Prentice-Hall.

Williams, J. L., McCarley, N. G., Haynes, J. M., Williams, E. H., Whetzel, T., Reilly, T., Giddens, M., & Bailey, L. (2016). The use of feedback to help college students identify relevant information on powerpoint slides. *North American Journal of Psychology*, *18*(2), 239–256.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin*, *147*(4), 399–435. https://doi.org/10.1037/bul0000309

Zamary, A., & Rawson, K. A. (2018). Which technique is most effective for learning declarative concepts—provided examples, generated examples, or both? *Educational Psychology Review*, *30*(1), 275–301. https://doi.org/10.1007/s10648-016-9396-9

Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction*, *46*, 12–20. https://doi.org/10.1016/j.learninstruc.2016.08.002

Received September 26, 2022
Revision received June 8, 2023
Accepted June 12, 2023 ■