THESIS

TESTING EFFECTS FOR SELF-GENERATED VERSUS EXPERIMENTER-GENERATED QUESTIONS

Submitted by

Sarah J. Myers

Department of Psychology

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Spring 2020

Master's Committee:

Advisor: Matthew Rhodes

Anne Cleary
James Folkestad

ABSTRACT

TESTING EFFECTS FOR SELF-GENERATED VERSUS EXPERIMENTER-GENERATED QUESTIONS

Those familiar with the testing effect (i.e., the finding that practicing retrieval improves memory) frequently suggest that students test themselves while studying for their classes. However, it is unclear whether students benefit from testing if they are not provided with testing materials. Few studies have examined whether generating one's own test questions improves performance, and none of these studies have given participants a full retrieval opportunity. The proposed experiments bridged this gap between testing effect and question generation research by allowing participants to generate questions and attempt to answer those questions after a delay. In Experiment 1, participants generated test questions over passages and either answered their questions as they created them or after a delay. In Experiment 2, participants either generated questions and answered them after a delay (i.e., self-testing), answered experimenter-generated questions, or restudied the material. Both experiments found no benefits of self-testing compared to the other conditions. In fact, those who self-tested tended to have worse final test performance than the other conditions. Analyses of the questions that participants created suggest that students may benefit more from self-testing when they generate more questions and those questions target material that is on the final test. Although further research is needed to confirm these conclusions (e.g., longer delays between study activities and final test), the current study suggests that testing may not always benefit learning if students must create their own questions.

ACKNOWLEDGEMENT

I would first like to thank my advisor, Dr. Matthew Rhodes, for his assistance with planning, conducting, and analyzing these experiments. Dr. Rhodes was always available for questions and provided encouraging and well-considered feedback throughout the entire research project. I would also like to thank my committee members, Dr. James Folkestad and Dr. Anne Cleary for their intriguing questions, feedback, and general interest in the research project.

I would like to thank my labmate, Hannah Hausman, for always being another resource and helping hand when I needed it, as well as my countless research assistants. Without them, this project would not have ever been finished. Lastly, I would like to thank my friends and family for their unfailing support and encouragement to complete and write this project.

TABLE OF CONTENTS

INTRODUCTION

The testing effect (i.e., the finding that taking a test improves one's memory for material compared to restudying) has been well-established in previous research (for reviews, see Roediger & Karpicke, 2006a; Roediger, Putnam, & Smith, 2011; Rowland, 2014). It has also been established in a variety of settings and for different types of materials, such as foreign language vocabulary and prose passages. Due to the amount of research supporting testing as one of the most effective learning strategies, many researchers suggest that students test themselves while studying (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). However, if students are not provided with testing material (e.g., quiz questions), it is unclear whether they would still benefit from testing. Little research has directly examined whether students show improved performance when they self-test by creating and answering their own test questions compared to other study strategies. Further, the few studies that have investigated self-testing have not given participants a full retrieval opportunity. The experiments reported bridged the gap between research on retrieval practice and self-generating questions by determining the overall effectiveness of incorporating both of these two similar, but distinct, strategies. Thus, the experiments I report examined whether students benefit from generating their own questions to use for retrieval practice (i.e., testing their memory) when testing material is not available to them.

The Testing Effect

Research on the testing effect has extensively documented the finding that tests serve as more than an evaluation tool to measure one's memory or abilities; taking a test over material improves memory for that material compared with simply restudying it (Roediger & Karpicke,

2006a). Direct benefits of testing to memory have been well-established since the early twentieth century. For example, Spitzer (1939) visited elementary schools across Iowa, gave students short passages to study, and then tested the students on different schedules. For example, some students received tests immediately after reading the passages, seven days later, and 63 days later. Other students completed tests seven and 21 days after initial exposure to the passages. Students who took a test (without feedback) within a week of studying the passages performed better on a final test given 63 days after the original study phase than students who were not tested prior to the final test, providing evidence that taking a test improves retention. However, an alternative explanation for Spitzer's (1939) findings is that students who took tests before the final test performed better because they were exposed to the study material more than students who did not take any practice tests. In modern studies, this explanation has been accounted for by controlling for the amount of exposure time. For example, Roediger and Karpicke (2006b) controlled for exposure by having participants either recall as much as they could from a studied passage (i.e., a free recall test) or reread the passage. After 2 days or 1 week, participants who took a practice test performed better than those who reread on a final test over the passage (although rereading led to superior performance after a 5-minute delay), thus suggesting that testing enhances memory.

The testing effect has proven robust in a number of studies. In a meta-analysis of 61 independent studies on the testing effect, Rowland (2014) estimated an overall benefit of testing over restudying of $g = 0.50$. Testing improves performance over restudying even when participants are not given corrective feedback (Roediger & Karpicke, 2006b; Rowland, 2014; but see Kang, McDermott, & Roediger, 2007). The benefit of testing without feedback is even more impressive when one considers total exposure to the material: a group who receives a restudy

opportunity is re-exposed to all the to-be-learned material whereas a group who receives a test without feedback is only re-exposed to the material they can accurately retrieve. If exposure alone accounted for the benefits of testing, one would predict that restudying would be more beneficial than testing for learning. Nevertheless, testing without feedback is still more effective than restudying (g = 0.39; Rowland, 2014).

The benefits of testing extend beyond the lab to educationally-relevant materials including foreign language vocabulary (Carrier & Pashler, 1992), text passages (Butler, 2010; Kang et al., 2007), and online lectures (Butler & Roediger, 2007). Testing has also improved overall performance in middle school, high school, and college classrooms, as well as medical residencies (e.g., Larsen, Butler, & Roediger, 2009; McDaniel, Anderson, Derbish, & Morrisette, 2007; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014). Testing not only promotes learning for a variety of material and contexts, but it can also benefit learning of related material beyond what is originally tested (Chan, McDermott, & Roediger, 2006; Rohrer, Taylor, & Sholar, 2010). Additionally, testing enhances participants' ability to apply learned information to new situations and contexts (i.e., transfer of learning; Butler, 2010; for a review, see Carpenter, 2012; but see Pan & Rickard, 2018, for a different conclusion) and may even mitigate text anxiety (Agarwal, D'Antonio, Roediger, McDermott, & McDaniel, 2014).

Testing with Self-Generated Questions

Based on the vast amount of research supporting the benefit of testing, researchers and educators suggest that students test themselves while studying for their classes (e.g., Dunlosky et al., 2013; Roediger et al., 2011; Rhodes, Cleary, & DeLosh, 2020). However, much of the research on testing does not necessarily support this suggestion. In almost all of these studies, participants in the testing conditions are given questions or cues provided by the experimenter.

Experimenters are usually aware of what questions participants will be given on a final test and thus may be more aware than participants of which material in a passage is considered important. Indeed, in some studies, participants in testing conditions receive the same list of questions for their practice test and the final test (e.g., Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). Therefore, it is unclear whether testing would still benefit learning if students were not provided with targeted questions. To benefit from retrieval practice if not given test questions, participants must be able to identify important information in the study material and write adequate, sufficiently challenging questions that they can then use to engage in retrieval practice.

Although research directly examining whether students can benefit from self-testing is sparse, related research on the *generation effect* suggests that it would be possible for students to benefit from creating their own test questions. Specifically, the generation effect refers to the finding that participants remember material better if they produce that material themselves compared to viewing the material (for reviews, see Bertsch, Pesta, Wiscott, & McDaniel, 2007; Mulligan & Lozito, 2004). By extension, students who create their own test questions might benefit from generating material over being given questions.

Most research on the generation effect has been conducted using word lists as the stimuli to be remembered. In these studies, words are more likely to be remembered if they were generated from a word fragment (e.g., "f_ ie_d") than if they were simply viewed (e.g., "friend"; Watkins & Sechler, 1988; see also Slamecka & Graf, 1978). Based on these findings, students who generate their own test questions may actually benefit more than students who are given test questions. Nonetheless, research examining generation effects for educational materials has been mixed. For example, Kelley, Chapman-Orr, Calkins, and Lemke (2019) found that requiring students in a cognitive psychology class to create multiple choice questions for each textbook

chapter improved their test scores. Another study reported that participants who generated their own headers for a passage outperformed participants who followed headers that were provided by the experimenter (Brooks, Dansereau, Holley, & Spurlin, 1983). However, Jonassen, Hartley, and Trueman (1986) reported that participants only benefitted from generating their own section headers if the headers they created were judged as "good." Therefore, students may not reap generation benefits when they create their own test questions if they cannot identify important material or write sufficient questions. Further, although research on both the testing and generation effects suggests that self-testing would improve memory, studies do not always support this conclusion. Most previous research on self-testing has focused on teaching students how to generate their own questions (for a review, see Rosenshine, Meister, & Chapman, 1996). However, the typical student does not have training on how to construct test questions, so these studies do not provide evidence for whether students can benefit from generating test questions (without instruction) while studying.

Only a handful of studies have directly examined the benefits of self-testing without training and the results of these studies are inconclusive. For example, Denner and Rickards (1987) found that participants who generated and answered test questions and those who answered experimenter-generated questions had higher performance than a restudy group on final test questions for which answers could be found within one sentence of the studied passage (i.e., factual questions). However, only those who answered experimenter-generated questions had enhanced performance on questions requiring integration of information across multiple sentences (i.e., conceptual questions). Similarly, Davey and McBride (1986) reported better performance on a final test for participants who generated and answered their own questions compared to those who answered experimenter-generated questions, but the increased

performance was restricted to factual but not conceptual questions. Owens (1976) found no difference in performance between those who self-tested and those who restudied material.

Although the previously mentioned studies have, at best, only reported benefits of self-testing under specific circumstances, a few studies have found support for self-testing. Weinstein, McDermott, and Roediger (2010) observed that participants who generated and answered their own questions or who answered experimenter-provided questions after reading a passage outperformed participants who restudied the passage on a final test. However, there was no difference in final test performance between answering experimenter-generated questions and self-testing. Similarly, Bugg and McDaniel (2012) compared memory for passages between those who restudied and those who generated either factual or conceptual questions. Results indicated that, for conceptual final test questions, those who created and answered conceptual questions on the initial test outperformed those who created factual questions or restudied. For factual final test questions, there was no difference in performance between the two self-testing conditions and the restudy condition. Foos, Mora, and Tcakz (1994) reported that participants who either self-generated questions with or without answers not only outperformed those who restudied but also outperformed participants who answered experimenter-generated questions. Foos et al. (1994) explain these findings via the generation effect: participants who generated their own questions received an added benefit from generation over participants who answered experimenter-generated questions.

Although promising, the previous literature on self-generating questions still leaves questions to be answered. In particular, there are only a handful of studies (e.g., Davey & McBride, 1986; Denner & Rickards, 1987; Weinstein et al., 2010) that have considered self-testing without training, which all produced different results. Furthermore, these previous studies

still do not fully answer the question of whether students benefit from *testing* themselves on their own questions. To elaborate, no prior study has given participants a true retrieval attempt when answering their generated questions. In the previously described studies, participants created their own questions and answered those questions simultaneously while having access to the passage. In order to have a true retrieval attempt, participants should create their own questions while having access to the passage and then answer those questions after a delay without the passage. Under these circumstances, participants would have to engage in episodic retrieval, potentially allowing them to receive the full benefits of retrieval practice (Rawson, Vaughn, & Carpenter, 2015; Whitten & Bjork, 1997).

One prior study (i.e., Weinstein et al., 2010) attempted to account for this discrepancy, but the authors' explanation does not take into consideration an important difference between their conditions. Specifically, Weinstein et al. (2010) compared final test performance of three different conditions. One condition (generate) read passages and then were told to write test questions over the passages and answers to those questions simultaneously, while having access to each passage. Another condition (answer) read passages and then received experimenter-generated questions to answer while also having access to each passage. The final condition (restudy) read passages and then were given a second opportunity to reread the passages. Thus, all study tasks were completed with access to the passages, with the authors suggesting that this situation mimicked an open-book test. Although some research has shown that open- and closed-book tests benefit performance similarly (Agarwal et al., 2008; Agarwal & Roediger, 2011), only one of the two testing conditions compared in Weinstein et al.'s (2010) study (i.e., answer vs. generate) was truly analogous to an open-book test. Participants who answered experimenter questions (i.e., the answer condition) would have completed an open-book test because they

received the questions and then could review the passage to find the answers. While doing this, some participants most likely would have attempted to retrieve the answer to each question before looking for the answer in the passage. In contrast, participants who generated and answered their own questions (i.e., the generate condition) did not necessarily complete a process analogous to an open-book test. When instructed to generate test questions and answers, these participants likely would have reviewed the passage to find important questions and would have written the answers immediately after writing the question. Thus, they would be viewing specific material within the passage on which the answer was based while writing the answers to their questions. Therefore, the generate condition in Weinstein et al. (2010) affords little possibility for episodic retrieval. Due to this difference in Weinstein et al.'s (2010) procedure, participants in the answer condition had more of an opportunity to attempt retrieval than those in the generate condition. This would allow those who answered experimenter questions to take more advantage of retrieval practice than those who generated and answered their own questions.

Given this difference in the opportunity for retrieval practice between groups, it is surprising that participants who generated questions and those who answered experimenter questions performed similarly on a final test (Weinstein et al., 2010). For reasons described previously, performance for participants who answered experimenter-generated questions most likely was driven to a greater extent by the benefits of retrieval practice than for participants in the generate condition. Accordingly, some other factor may have been driving performance for the self-testing group, leading to similar performance between the two groups. One possibility is that the generate group had an added benefit due to the generation effect, which offset the extra benefit of retrieval practice for the answer group. By extension, if participants in the generate and answer conditions had equal opportunities to engage in retrieval practice, the generate group

should have had better performance on a final test than the answer group. That is, adding a delay between generating questions and attempting to answer those questions should be a more effective study strategy than creating and answering questions simultaneously.

Indeed, several studies show that delaying an initial test (i.e., increasing the time between initial exposure to material and the first test) increases the benefit of testing compared to taking a test soon after studying (Jacoby, 1978; Rawson, Vaughn, & Carpenter, 2015; Whitten & Bjork, 1977; see also Kornell, Bjork, & Garcia, 2011). Because of this, students might benefit more from self-testing if they had a break between generating test questions and attempting to answer those questions (i.e., an initial test). In previous studies (e.g., Weinstein et al., 2010), participants generated and answered questions simultaneously, leaving them unable to take advantage of a delayed initial practice test. I addressed these concerns in the experiments reported by combining typical testing effect research procedures with research on self-generating questions. These data indicate whether participants can truly benefit from testing themselves if they use their own questions compared to prior studies.

The Current Study

The current study consisted of two experiments. In Experiment 1, I determined whether participants benefitted more from generating questions and answering them after a delay compared to generating and answering questions at the same time with access to the passage (the procedure used in e.g., Denner & Rickards, 1987; Weinstein et al., 2010). I hypothesized that participants who answered questions after a delay would perform better than those who generated and answered questions simultaneously. That is, participants who have a delay should attempt to retrieve answers from memory, whereas those who generated and answered questions simultaneously would be less likely to engage in retrieval practice.

In Experiment 2, I used a similar design as Weinstein et al. (2010) to compare the effectiveness of three study strategies: generating and answering questions, answering experimenter-generated questions, or rereading material. In contrast to Weinstein et al. (2010), Experiment 2 added a delay between generating and answering questions. This allowed participants who generated and answered their own test questions (self-test condition) and those who answered experimenter-generated questions (answer condition) to have similar opportunities for retrieval practice. Additionally, both of these groups did not have access to the passage when they attempted to answer questions (although they had an opportunity for feedback), thus encouraging both groups to attempt to retrieve answers from memory rather than searching for the answers in the passage. I hypothesized that participants in the self-test and answer conditions would perform better than those who only reread the passage (reread condition) since both the self-test and answer conditions would benefit from retrieval practice. Moreover, those in the self-test condition were expected to outperform those in the answer condition because the self-test condition had the combined benefits of both generation and retrieval practice.

Another difference between the proposed study and previous self-testing research is that the proposed study did not provide participants with any examples or instructions on what types of questions they should generate. Although some prior studies were not investigating the effects of training on self-generating questions, several of these studies included a practice phase where participants attempted to answer experimenter-generated questions and then completed a final test over a practice passage (e.g., Bugg & McDaniel, 2012; Weinstein et al., 2010). Participants were then told they should generate questions similar to the ones they answered in the practice phase. This does not give an accurate depiction of how students would generate test questions in

their own studying as students are unlikely to be given any training in how to create test questions. Thus, the proposed study did not include a practice phase so that participants were not influenced in how they created their own test questions, thus increasing ecological validity.

In addition, I analyzed whether the benefits of self-generating or answering experimenter-generated questions differed depending on the type of question: conceptual or factual. Because some previous studies have shown differences in the effects of self-testing between conceptual and factual questions (e.g., Bugg & McDaniel, 2012; Davey & McBride, 1986; Denner & Rickards, 1987), it is important to consider these two types of questions separately in the proposed study. Therefore, experimenter-generated and final test questions included both factual and conceptual questions so that analyses could be performed for both types of questions. Furthermore, participants' generated questions were scored as either conceptual or factual. Participants may also benefit from self-testing differently depending on whether they target information that is included on the final test in their generated questions. Therefore, I also scored participants' generated questions for whether or not they targeted material on the final test.

Lastly, I also asked participants after the study phase how they thought they would perform on the final test over each passage. This metacognitive question provided information regarding students' beliefs about the effectiveness of the study strategies: generating and answering their own questions, answering experimenter-generated questions, and rereading (Dunlosky & Metcalfe, 2008; Karpicke & Roediger, 2008; Nelson, 1996; Rhodes, 2019). Based on previous findings (Weinstein et al., 2010), I hypothesized that those in the self-test conditions would have higher predictions than those in the answer and reread conditions. In all, because the proposed study combined self-generating test questions and fully engaging in retrieval practice, I

was able to examine whether students could benefit from testing themselves if they created their own test questions.

EXPERIMENT 1

Experiment 1 determined whether participants gain added benefits from self-testing if they answer their questions after a delay compared to generating questions and answers to those questions at the same time. Participants initially read two passages, generated questions over the passages, answered their generated questions (either immediately or after a delay), and then completed a final test over each passage. Based on findings that testing leads to larger benefits with a longer delay between study and initial test (Jacoby, 1978; Whitten & Bjork, 1977), participants who answered their questions after a delay were expected to perform better on the final test than participants who generated and answered questions simultaneously. Furthermore, participants who answered questions after a delay had more opportunity to retrieve answers from memory than those who generated questions and answers simultaneously while having access to the passage.

Methods

Participants

Participants were 145 undergraduate students (76 in the self-test condition, 69 in the simultaneous condition) from Colorado State University who participated in exchange for course credit. Twenty-three participants were removed from analyses due to technical difficulties or not completing the experiment (n = 4), participants already seeing the passages (n = 7), or participants not following the experimental instructions (n = 12). Therefore, data from 122 participants (65 in the self-test condition, 57 in the simultaneous condition) were used in analyses. A sensitivity analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that this sample size was sufficient to detect an effect size of d = .51, assuming an

alpha of .05, power of .80, and a two-tailed test. Participants (40 men, 81 women, 1 non-binary) were between 16 and 27 (M = 18.85, SD = 1.33) years old. Two participants did not provide their age.

Materials

Two short passages on the topics of monetary policy (549 words) and ice ages (1052 words) were used for study materials (see Appendix A; Thiede, Wiley, & Griffin, 2011). Ten multiple-choice questions for each passage developed by Thiede et al. (2011) were used for the final test (see Appendix B). Five of these questions had answers that were explicitly stated within one sentence of the text (i.e., factual questions). The other five questions required participants to integrate information across at least two sentences (i.e., conceptual questions). Final test questions were originally classified as factual and conceptual by Thiede et al. (2011). However, based on the definitions used in the current study[1], one factual question on the final test was re-scored as conceptual for each passage, and one conceptual question for each passage was changed to factual. The entire experiment was conducted in Qualtrics and both experiments were approved by the Colorado State institutional review board before data collection began.

Procedure

After providing consent, participants were randomly assigned to either generate questions and write the answers immediately while having access to the passage (simultaneous condition) or to generate questions and answer them after a delay without the passage (self-test condition). Appendix C depicts the procedure used in Experiment 1. First, participants were given 5 minutes

[1]Factual questions were defined as questions that could be answered using one sentence of the passage, whereas conceptual questions were defined as questions that required participants to integrate information across two or more sentences.

to read one of the two passages. Then, participants in the self-test condition 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. Please do NOT type the answers to the questions you write."

After reading these instructions, participants in the self-test condition had 7 minutes to type in their questions and could scroll to see the passage underneath the response box. To equate exposure time, participants in the simultaneous condition reread the passage during this time. All participants then had 5 minutes to read the second passage and 7 minutes to generate their own questions (self-test condition) or reread the passage (simultaneous condition). The order of passages was counterbalanced across participants so that half of the participants in each condition read the ice age passage first and half read the monetary policy passage first.

Next, participants in the self-test condition were shown their generated questions over the first passage on the screen (although they did not have access to the first passage). They were instructed to answer their questions from memory and were encouraged to guess if they could not remember an answer. Participants in the simultaneous condition saw the first passage again and were instructed to generate questions and answers over the first passage. Instructions were the same as the self-test condition, except that participants in the simultaneous condition were told to type their questions with the answers. All participants had 6 minutes for these activities. After this, all participants saw their questions, answers, and the first passage for 6 minutes. During this time, participants were instructed to self-score their questions using the passage as a means of 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. After scoring their questions for the first passage, participants in the self-test condition had 6

minutes to answer the questions they generated over the second passage and participants in the simultaneous condition had 6 minutes to generate questions and answers for the second passage. After that, both groups had 6 minutes to score their answers.

Participants were then told they would take a ten-question multiple-choice test over each passage about 5 minutes later. They were asked to predict how many of the ten final test questions they would answer correctly for each passage (providing a global JOL)[2]. After then completing a 3-minute distraction phase of solving math problems, participants received a final test over the first passage they read. Questions were presented in a unique random order for each participant and the final test was self-paced. All participants then completed a self-paced final test over the second passage. Since the ice age and monetary policy passages have been used in other Colorado State University experiments, one final question asked participants if they had seen or read these passages before. Participants were then debriefed and released.

Participant removal

Participants' data were removed if they did not generate any questions, if they provided answers when they were not prompted to, if they responded they had seen the passages before, or if they did not complete the entire experiment.

Scoring and analysis

All participants' generated questions were scored as conceptual or factual (C/F) and for whether the question targeted the same information as a final test question or not (On FT/Not on FT). Furthermore, participants' answers to their own questions were scored by the experimenters as correct or incorrect (accuracy; 0 = incorrect, 1 = correct). Inter-rater reliabilities were lower

[2]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 they read than the second. This was corrected in Experiment 2.

than adequate - C/F: $k = .43$, On FT/Not on FT: $k = .63$, accuracy: $k = .71$. Therefore, all questions were scored by two scorers and controversies were settled by a third scorer.

Data were analyzed using JASP Version 0.11.1 (JASP team, 2019) and SPSS Version 24 (IBM Corp., 2016). I 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_{2p}), and the Bayes factor (*BF*). The Bayes factor (BF_{10}) is 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). Subsequently, 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, Speckman, Sun, Morey, and Iverson (2009), I used the JZS prior to calculate Bayes factors because it requires the fewest prior assumptions about the range of the true effect size. I interpreted Bayes factors using recommendations from Wagenmakers (2007), whereby Bayes factors provide weak ($1 < BF \leq 3$), positive ($3 < BF \leq 20$), strong ($20 < BF \leq 150$), or very strong ($BF > 150$) evidence in favor of one hypothesis over the other.

Results

Order Effects

A 2 (order: ice age first, monetary policy first) x 2 (strategy: self-test, simultaneous) x 2 (type of question: conceptual, factual) mixed-factor analysis of variance (ANOVA) was conducted to analyze the effect of the order in which participants read the two passages. Order and strategy were between-participant variables, whereas type of question was manipulated

within-participants. Only results pertaining to the effect of order are reported in this section. Overall, those who studied the ice age passage first (M = 67.42, SE = 1.63) and those who studied the monetary policy passage first (M = 70.84, SE = 1.68) did not differ in final test performance, $F(1, 118) = 2.15$, $p = .15$, $\eta_{2p} = .02$, $BF_{01} = 2.99$. Order also did not interact significantly with strategy, $F(1, 118) = 2.63$, $p = .11$, $\eta_{2p} = .02$, $BF_{01} = 1.82$, nor type of question, $F(1, 118) = 1.09$, $p = .30$, $\eta_{2p} = .01$, $BF_{01} = 3.10$. Furthermore, the 3-way interaction was not significant, $F(2, 118) = 0.95$, $p = .33$, $\eta_{2p} = .01$, $BF_{01} = 2.42$. Because order did not have a significant effect on final test performance, it was dropped from further analyses.

Performance by passage

A 2 (passage: ice age, monetary policy) x 2 (strategy: self-test, simultaneous) x 2 (type of question: factual, conceptual) mixed-factor ANOVA was also conducted to examine differences in final test performance based on the passages studied. Passage and type of question were manipulated within-participants, while strategy was manipulated between-participants. Only results pertaining to the passage are presented. Overall, participants performed better on questions about the ice age passage (M = 73.30, SE = 1.65) than the monetary policy passage (M = 64.91, SE = 1.27), $F(1, 120) = 22.62$, $p < .001$, $\eta_{2p} = .16$, $BF_{10} = 150.50$. This was qualified by a significant passage x type of question interaction, $F(1, 120) = 140.18$, $p < .001$, $\eta_{2p} = .54$, $BF_{10} = 2.42\text{x}10^{22}$. Collapsed across study strategy, factual questions about the monetary policy passage (M = 84.26, SE = 1.68) were answered correctly more often than factual questions about the ice age passage (M = 73.93, SE = 1.97), $t(121) = -4.83$, $p < .001$, $d = -0.51$, $BF_{10} = 3515$. In contrast, conceptual questions from the ice age passage (M = 72.46, SE = 2.05) were more often correct than conceptual questions from the monetary policy passage (M = 45.41, SE = 1.73), $t(121) = 10.61$, $p < .001$, $d = 1.29$, $BF_{10} = 9.65\text{x}10^{15}$. The passage x study strategy interaction

was not significant, $F(1, 120) = 0.10$, $p = .75$, $\eta_{2p} = .001$, $BF_{01} = 6.67$, nor was the three-way interaction, $F(1, 120) = 0.62$, $p = .43$, $\eta_{2p} = .01$, $BF_{01} = 4.13$. Therefore, it was dropped from further analyses.

Final test performance

A 2 (strategy: self-test, simultaneous) x 2 (type of question: conceptual, factual) mixed-factor ANOVA was conducted on participants' final test performance, with strategy manipulated between-participants and type of question manipulated within-participants (see Figure 1). Overall, participants answered more factual questions correctly ($M = 79.28$, $SE = 1.48$) than conceptual questions ($M = 58.92$, $SE = 1.41$), $F(1, 120) = 149.10$, $p < .001$, $\eta_{2p} = .55$, $BF_{10} = 5.79 \times 10^{20}$. On average, participants in the self-test ($M = 67.77$, $SE = 1.61$) and simultaneous ($M = 70.44$, $SE = 1.72$) conditions did not differ, $F(1, 120) = 1.28$, $p = .26$, $\eta_{2p} = .01$, $BF_{01} = 3.95$. The type of question x strategy interaction also did not reach conventional significance, $F(1, 120) = 3.18$, $p = .07$, $\eta_{2p} = .03$, $BF_{01} = 1.20$. However, planned comparisons were conducted to compare performance between the two study strategies for factual and conceptual questions separately[3]. A follow-up *t*-test showed that, for factual questions, participants in the simultaneous condition numerically outperformed those in the self-test condition, although this difference was not significant and the Bayes factor was inconclusive, $t(120) = 1.90$, $p = .06$, $d = 0.35$, $BF_{01} =$

[3]Following recommendations from Keppel and Wickens (2004), I did not adjust the alpha level from 0.05 because I planned to run these *t*-test analyses a priori.

1.01. For conceptual questions, those in the simultaneous and self-test conditions did not differ, $t(120) = 0.11$, $p = .91$, $d = 0.02$, $BF_{01} = 5.15$.

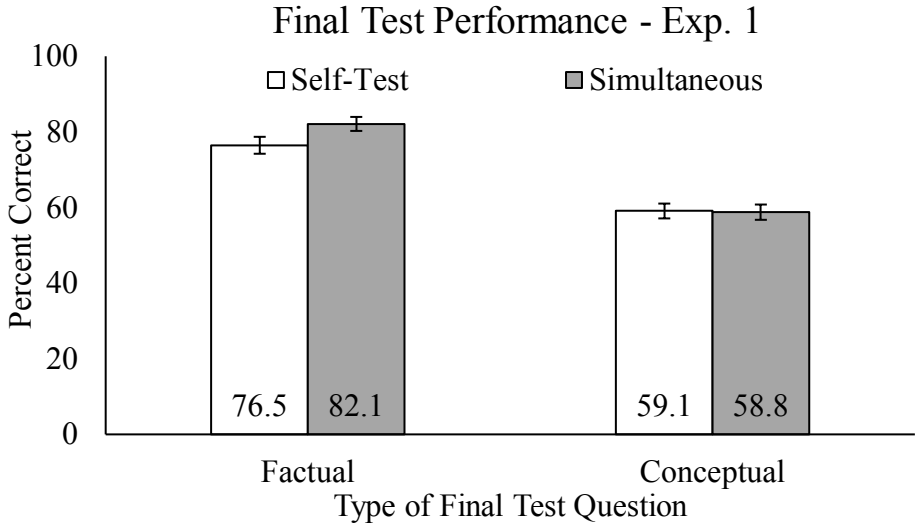

Figure 1. Percent of factual and conceptual questions answered correctly on the final test for those in the self-test and simultaneous conditions. Error bars represent one standard error of the mean.

Analysis of participants' scoring

After participants answered their generated questions, they had an opportunity to score their answers (using the passages) as a means of feedback. The participants' scores were compared to the experimenters' accuracy scores to determine how well participants scored their own questions. Overall, 83.3% of participants' scores agreed with the experimenter scores (IRR: $k = .37$). For questions where the participant and experimenter disagreed, participants scored 52% of their questions as correct when the experimenter marked it as incorrect. Contrary to prior research (e.g., Dunlosky, Rawson, & McDonald, 2002), participants were not overly confident when scoring their accuracy. This may have been because they scored their answers with access to the correct information (within the passages), which reduces overconfidence (Rawson & Dunlosky, 2007).

Analysis of generated questions

Table 1 presents the average number of questions generated, proportions of factual and conceptual questions, proportion of questions that targeted material on the final test, and proportion of questions that participants answered correctly. Participants in the self-test condition, on average, created more questions than participants in the simultaneous condition, $t(120) = 5.33$, $p < .001$, $d = 0.97$, $BF_{10} = 28846.23$. This difference was expected since those in the self-test condition had 7 minutes to create questions and 6 minutes to later answer those questions, whereas those in the simultaneous condition only had 6 minutes to both create and answer questions. Those in the self-test and simultaneous conditions did not differ significantly in the proportion of factual questions, $t(120) = 1.57$, $p = .12$, $d = 0.30$, $BF_{01} = 1.70$, or proportion of questions they created 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 condition were more accurate than those in the self-test condition, $t(120) = 3.39$, $p < .001$, $d = 0.62$, $BF_{10} = 30.33$. Again, this was expected because those in the simultaneous condition answered their questions with access to the passage while those in the self-test condition could only answer their questions from memory.

Table 1. The average number of questions participants generated (No. of questions), proportion of factual questions generated (Prop. Factual), proportion of conceptual questions generated (Prop. Conceptual, calculated as 1 - Prop. Factual), proportion of questions that targeted material on the final test (Prop. On FT), and proportion of generated questions answered correctly (Prop. Correct) for each passage in Experiment 1 and 2.

	Experiment 1					Experiment 2				
Study Strategy	**No. of Questions**	**Prop. Factual**	**Prop. Conceptual**	**Prop. On FT**	**Prop. Correct**	**No. of Questions**	**Prop. Factual**	**Prop. Conceptual**	**Prop. On FT**	**Prop. Correct**
Self-Test	6.95 (2.48)	.58 (.24)	.42 (.24)	.35 (.15)	.84 (.17)	7.05 (2.59)	.56 (.23)	.44 (.23)	.26 (.14)	.83 (.16)
Simultaneous	4.88 (1.67)	.64 (.25)	.36 (.25)	.39 (.12)	.93 (.11)					

Note: Standard deviations in parentheses

Correlations were calculated between the types of questions participants created and their performance on the final test. Because the two study strategy conditions used different procedures, these correlations were run separately for each condition. These analyses are exploratory, as I had no prior predictions on how these factors would impact final test performance. The analyses will inform predictions for Experiment 2, which will serve as confirmatory analyses.

For the self-test condition, both the proportion of generated questions that targeted material on the final test (r = .34, p = .005, BF_{10} = 7.14) and the proportion of questions answered correctly (r = .28, p = .02, BF_{10} = 1.86) were significantly correlated with final test performance (no. of questions: r = .20, p = .11, BF_{01} = 1.79; prop. factual: r = .02, p = .85, BF_{01} = 6.33). Because there were moderate correlations between some of these factors and final test performance, a linear regression analysis was conducted with final test performance for the self-test condition regressed on the number of questions created, proportion of factual questions, proportion of questions that targeted final test material, and accuracy. For each predictor, both the unstandardized (b) and standardized (β) regression coefficients are reported, along with the corresponding standard error (SE), p-value (p), and Bayes factor (BF). In this model, the proportion of created questions that overlapped with final test questions (b = 28.12, β = 0.32, SE = 10.99, p = .013, BF_{10} = 5.29) significantly predicted final test performance while controlling for the other factors, such that a .1 increase in proportion of questions that targeted final test material was associated with a 2.8-percentage point increase in final test performance. The number of questions created (b = 1.47, β = 0.27, SE = .65, p = .03, BF_{10} = 3.13) also predicted final test performance, such that creating 1 additional question was associated with a 1.5-percentage point increase in final test performance. Proportion of factual questions (b = -6.84, β

= -0.06, SE = 8.66, p = .43, BF_{01} = 2.40) and accuracy (b = 14.76, β = 0.17, SE = 12.78, p = .25, BF_{01} = 1.17) did not significantly predict final test performance.

For the simultaneous condition, only the proportion of factual questions that participants created correlated significantly with final test performance (r = -.26, p = .049, BF_{10} = 1.09) (No. questions: r = .02, p = .91, BF_{01} = 6.02; prop. on FT: r = .20, p = .14, BF_{10} = 2.14; accuracy: r = -.17, p = .22, BF_{01} = 2.86). A linear regression analysis was also conducted with final test performance for the simultaneous condition regressed on the number of questions created, proportion of factual questions, proportion of questions that targeted final test material, and accuracy. In this model, the proportion of factual questions (b = -12.76, β = -0.26, SE = 6.84, p = .07, BF_{10} = 1.71) did not reach conventional significance to be considered a predictor of final test performance when controlling for the other factors, and Bayesian evidence was inconclusive. The proportion of questions that overlapped with final test questions (b = 22.19, β = 0.22, SE = 13.83, p = .12, BF_{10} = 1.21), proportion of questions answered correctly (b = -24.76, β = -0.22, SE = 15.27, p = .11, BF_{10} = 1.23), and number of questions created (b = 0.98, β = 0.13, SE = 1.01, p = .34, BF_{01} = ≤ 1.59) also did not significantly predict final test performance.

Global JOLs

A 2 (strategy: self-test, simultaneous) x 2 (passage: ice age, monetary policy) mixed-factor ANOVA was conducted to analyze participants' JOL ratings, with passage manipulated within-participants and strategy manipulated between-participants. Overall, participants predicted that they would perform better on the ice age final test (M = 63.59, SE = 1.60) than the monetary policy final test (M = 55.80, SE = 1.71), F(1, 120) = 33.21, p < .001, η_{2p} = .22, BF_{10} = 1.90x10$_5$. On average, participants' predictions did not differ between the self-test (M = 58.08, SE = 2.07) and simultaneous (M = 61.32, SE = 2.21) conditions, F(1, 120) = 1.14, p = .29, η_{2p} =

.01, BF_{01} = 2.43. However, the passage x strategy interaction was significant, $F(1, 120) = 7.71$, $p = .01$, $\eta_{2p} = .06$, $BF_{10} = 6.13$. Follow-up tests indicated that, for the ice age passage, participants' predictions in the self-test ($M = 63.85$, $SE = 2.30$) and simultaneous ($M = 63.33$, $SE = 2.20$) conditions did not differ, $t(120) = 0.16$, $p = .87$, $d = 0.03$, $BF_{01} = 5.10$. However, for the monetary policy passage, participants in the simultaneous condition ($M = 59.30$, $SE = 2.44$) predicted they would perform significantly better than those in the self-test condition ($M = 52.31$, $SE = 2.39$), $t(120) = 2.04$, $p = .04$, $d = 0.37$, $BF_{10} = 1.25$.

Absolute calibration was also calculated by subtracting participants' actual final test score from their predicted score. Therefore, positive values indicate that participants were overconfident, whereas negative values indicate they were underconfident (i.e., predicted they would perform worse than they did). Participants were, on average, underconfident in how well they would perform on the final tests. A 2 (strategy: self-test, simultaneous) x 2 (passage: ice age, monetary policy) mixed-factor ANOVA indicated that, overall, participants' calibration did not differ between their predictions for the ice age ($M = -9.71$, $SE = 1.89$) and monetary policy passages ($M = -9.10$, $SE = 1.83$), $F(1, 120) = 0.09$, $p = .76$, $\eta_{2p} = .001$, $BF_{01} = 7.19$. Overall, participants' calibration also did not differ between the self-test ($M = -9.69$, $SE = 2.14$) and simultaneous ($M = -9.12$, $SE = 2.29$) conditions, $F(1, 120) = 0.03$, $p = .86$, $\eta_{2p} < .001$, $BF_{01} = 4.59$. However, the passage x strategy interaction was significant, $F(1, 120) = 4.56$, $p = .04$, $\eta_{2p} = .04$, $BF_{10} = 1.55$. Follow-up tests indicated that, for the ice age passage, participants in the self-test condition ($M = -7.85$, $SE = 2.56$) were more accurate than the simultaneous condition ($M = -11.58$, $SE = 2.80$), although this difference was not significant and the Bayes factor supported the null, $t(120) = 0.99$, $p = .33$, $d = 0.18$, $BF_{01} = 3.33$. However, for the monetary policy passage, those in the simultaneous condition ($M = -6.67$, $SE = 2.46$) were more accurate than the self-test

condition (M = -11.54, SE = 2.67), although this difference again was not significant and the Bayes factor favored the null, $t(120) = -1.34$, $p = .18$, $d = -0.24$, $BF_{01} = 2.33$.

Discussion

Contrary to hypotheses, final test performance was largely equivalent between those who answered their generated questions after a delay (self-test condition) and those who answered their questions immediately (simultaneous condition). If anything, those who answered their questions immediately performed slightly (although not significantly) better than those who had a delay between generating and answering questions. Analyses of the questions participants created revealed that, for those in the self-test condition, the number of questions generated and proportion of questions that targeted material on the final test significantly predicted how well they performed on the final test. Specifically, the more questions participants created, the better they performed on the final test. In addition, the more of those questions that specifically targeted final test material, the better they performed on the final test. However, these relationships were still exploratory. I will return to these findings in Experiment 2 and the general discussion. No other variables examined for generated questions predicted final test performance for those in the simultaneous condition.

EXPERIMENT 2

In Experiment 2, self-testing (with a delay between generating and answering questions) was compared to the other two study strategies used in Weinstein et al. (2010). Specifically, one group generated questions and answered them after a delay (self-test), one group answered experimenter-generated questions after a delay (answer), and one group reread the passages (reread) to control for exposure time. Because the self-test condition incorporated both generation and retrieval practice, I expected participants in the self-test condition to outperform those in the answer and reread conditions on a final test. I also expected those in the answer condition to perform better than those in the reread condition due to the benefits of testing.

Methods

Participants

Participants were 228 undergraduate students (73 in the self-test condition, 78 in the answer condition, 77 in the reread condition) from Colorado State University who completed the experiment in exchange for course credit. Twenty-six participants were removed from analyses because of technical difficulties or not completing the experiment ($n = 2$) or having already seen the passages ($n = 24$). Therefore, data from 202 participants (68 in the self-test condition, 66 in the answer condition, 68 in the reread condition) 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 alpha of .05, power of .80, and a two-tailed test. Participants (79 men, 118 women, 2 non-binary, 1 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

Participants studied the same two passages used in Experiment 1. The final tests were also the same as Experiment 1, including five factual and five conceptual questions for each passage. All experiment activities again took place in Qualtrics. I developed eight new short-answer questions to use as the experimenter-generated questions for the answer condition (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 targeted information that was on the final test (On FT), and two of each type did not appear on the final test. Based on this design, 40% (four of the ten questions) of participants' final test questions were targeted on the initial test in the answer condition, which is close to the proportion of questions targeting final test material that participants generated in Experiment 1. The eight questions were chosen based on participant performance in a pilot experiment and agreement between scorers for whether questions were conceptual or factual and whether they targeted information on the final test.

Procedure

Appendix C depicts the procedure used in Experiment 2. After providing consent, participants were randomly assigned to either the self-test, answer, or reread condition. Participants in the self-test and answer conditions answered questions without access to the passage, although they had an opportunity to receive feedback. First, all participants were given 5 minutes to read one of the two passages. Then, participants in the self-test condition were given 7 minutes 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 minutes. All participants then followed this same procedure for the second passage. Order of passages was again counterbalanced across participants.

Next, both the self-test and answer conditions answered questions over the first passage for 6 minutes, which served as an initial test. Participants in the self-test condition answered the questions they generated and participants in the answer condition answered the eight experimenter-generated 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 6 minutes to score their answers using the procedure described in Experiment 1. Participants in the reread condition reviewed the passage again for these 6 minutes. Next, participants predicted how many of the ten final test questions they would answer correctly on a final test for the first passage they studied. Participants then completed these procedures again for the second passage. Participants followed the same procedure as Experiment 1 for the 3-minute distractor phase and the two final tests. Lastly, participants were asked if they had seen or read either of the study passages before.

Participant removal and analysis plan

Participants' data were removed if they did not generate any questions in the self-test condition, if they responded they had seen the study passages before, or if they did not complete the entire experiment. Data were analyzed using the same methods as Experiment 1.

Scoring

Questions generated by participants in the self-test condition were scored as conceptual or factual (C/F) and for whether the question targeted information on the final test or not (On FT/Not on FT). Participants' answers to their own questions were also scored by researchers as correct or incorrect (accuracy; 0 = incorrect, 1 = correct). In addition, answers that participants gave to questions in the answer condition were scored as correct or incorrect. Inter-rater

reliabilities were lower than adequate (C/F: k = .39, On FT/Not on FT: k = .69, accuracy: k = .56 in the self-test condition, k = .77 in the answer condition). Therefore, all questions were scored by two scorers and controversies were settled by a third scorer.

Results

Order effects

A 2 (order: ice age first, monetary policy first) x 3 (strategy: self-test, answer, reread) x 2 (type of question: conceptual, factual) mixed-factor ANOVA was conducted to analyze the effect of the order in which participants read the two passages. Order and strategy were manipulated between-participants, whereas type of question was manipulated within-participants. Overall, those who studied the ice age passage first (M = 67.62, SE = 1.40) and those who studied the monetary policy passage first (M = 68.37, SE = 1.40) did not differ in final test performance, $F(1, 196) = 0.14$, $p = .71$, $\eta_{2p} = .001$, $BF_{01} = 7.30$. Order also did not interact significantly with strategy, $F(1, 196) = 0.11$, $p = .89$, $\eta_{2p} = .001$, $BF_{01} = 13.89$, or type of question, $F(1, 196) = 0.43$, $p = .51$, $\eta_{2p} = .002$, $BF_{01} = 5.49$. The three-way interaction was not significant, $F(2, 196) = 0.37$, $p = .69$, $\eta_{2p} = .004$, $BF_{01} = 2.42$. Because order did not have a significant effect on final test performance, it was dropped from further analyses.

Performance by passage

A 3 (strategy: self-test, answer, reread) x 2 (passage: ice age, monetary policy) x 2 (type of question: factual, conceptual) mixed-factor ANOVA was also conducted to examine differences in final test performance by passage. Passage and type of question were manipulated within-participants, while strategy was manipulated between-participants. Overall, participants performed better on questions about the ice age passage (M = 72.78, SE = 1.33) than the monetary policy passage (M = 63.24, SE = 1.10), $F(1, 199) = 43.28$, $p < .001$, $\eta_{2p} = .18$, $BF_{10} =$

612,515.08. The three-way interaction was not significant, $F(1, 199) = 0.73$, $p = .48$, $\eta_{2p} = .01$, $BF_{01} = 10.64$. However, there was a significant passage x type of question interaction, $F(1, 188) = 154.47$, $p < .001$, $\eta_{2p} = .44$, $BF_{10} = 3.23\text{x}10^{59}$. Collapsed across strategy, factual monetary policy questions ($M = 80.40$, $SE = 1.42$) were answered correctly more often than factual ice age questions ($M = 74.36$, $SE = 1.60$), $t(201) = -3.44$, $p = .001$, $d = -0.28$, $BF_{10} = 22.96$. In contrast, conceptual ice age questions ($M = 71.19$, $SE = 1.70$) were more often correct than conceptual monetary policy questions ($M = 46.04$, $SE = 1.45$), $t(201) = 12.04$, $p < .001$, $d = 1.12$, $BF_{10} = 2.21\text{x}10^{22}$. The passage x study strategy interaction was also significant, although the Bayes factor provided positive evidence of no interaction, $F(1, 199) = 3.44$, $p = .03$, $\eta_{2p} = .03$, $BF_{01} = 3.02$. Collapsed across type of question, ice age final test scores differed based on the study strategy used, $F(1, 199) = 6.75$, $p = .001$, $BF_{10} = 18.00$, but monetary policy test scores did not, $F(1, 199) = 1.24$, $p = .29$, $BF_{01} = 6.58$. For the ice age final test, those in the answer ($M = 73.18$, $SE = 2.29$) and reread ($M = 78.53$, $SE = 2.31$) conditions performed significantly better than those in the self-test ($M = 66.62$, $SE = 2.33$) condition, $t(132) = 2.01$, $p = .046$, $d = 0.35$, $BF_{10} = 1.15$ and $t(134) = 3.63$, $p < .001$, $d = 0.62$, $BF_{10} = 63.08$, respectively. The answer and reread conditions did not differ significantly, $t(132) = 1.65$, $p = .10$, $d = 0.28$, $BF_{01} = 1.58$.

Final test performance

A 3 (strategy: self-test, answer, reread) x 2 (type of question: conceptual, factual) mixed-factor ANOVA was conducted on participants' final test performance, with strategy manipulated between-participants and type of question manipulated within-participants (see Figure 2). Overall, participants answered more factual questions correctly ($M = 77.41$, $SE = 1.20$) than conceptual questions ($M = 58.61$, $SE = 1.18$), $F(1, 199) = 194.93$, $p < .001$, $\eta_{2p} = .50$, $BF_{10} =$

3.23x10$_{28}$. The main effect of study strategy was also significant, $F(1, 199) = 5.10$, $p = .01$, $\eta_{2p} = .05$, $BF_{10} = 2.22$.

The type of question x strategy interaction did not reach conventional significance, $F(1, 199) = 2.64$, $p = .07$, $\eta_{2p} = .03$, $BF_{01} = 1.96$. 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 condition outperformed those in the self-test condition, $t(132) = 3.13$, $p = .002$, $d = 0.54$, $BF_{10} = 14.37$ and $t(134) = 3.00$, $p = .003$, $d = 0.51$, $BF_{10} = 10.17$, respectively. Factual test performance for those in the answer and reread conditions did not differ, $t(132) = 0.16$, $p = .87$, $d = 0.03$, $BF_{01} = 5.35$. 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, $t(132) = 1.37$, $p = .17$, $d = 0.24$, $BF_{01} = 2.31$. Those in the reread condition numerically outperformed those in the self-test condition, although this difference was not significant and the Bayes factor slightly favored the null, $t(134) = 1.75$, $p = .08$, $d = 0.30$, $BF_{01} = 1.36$.

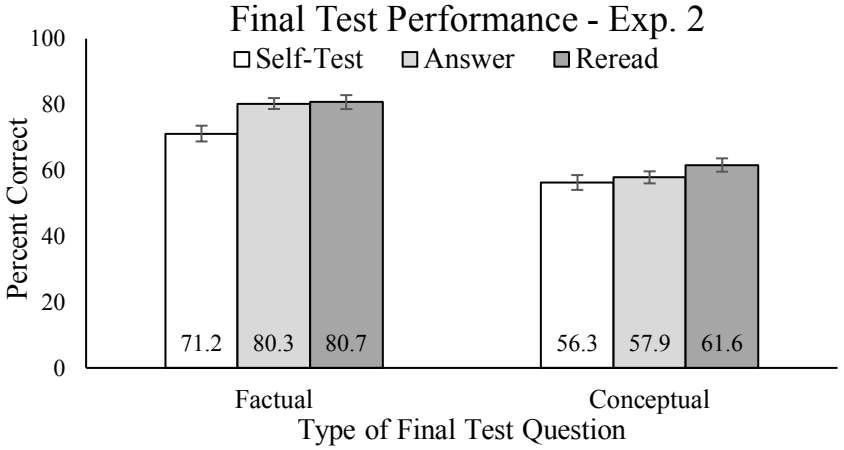

Figure 2. Percent of factual and conceptual questions answered correctly on the final test for those in the self-test, answer, and reread conditions. Error bars represent one standard error of the mean.

Analysis of participants' scoring

Participants' scores in the self-test and answer conditions were compared to the experimenters' accuracy scores to determine how well participants scored their own questions. A programming error resulted in participants not being shown their answer for one of the eight monetary policy questions while they scored their answers. Therefore, accuracy for this question was dropped from scoring analyses. Overall, scores overlapped for 83.3% of the questions in the self-test condition and 78.2% of questions in the answer condition (IRR: $k = .52$ for self-test condition, $k = .51$ for answer condition). For the self-test condition, a majority (76.9%) of contradicting questions were scored as correct by the experimenter and incorrect by the participant. For the answer condition, participants scored 53.7% of contradicting questions as incorrect while the experimenter marked it as correct.

Analysis of generated questions

Table 1 presents the average number of questions created, proportion of factual and conceptual questions, proportion of questions that targeted final test material, and proportion of questions that participants answered correctly in the self-test condition from Experiment 2. Participants in the answer condition answered 67.4% of their practice test questions correctly. An independent-samples *t*-test indicated that those in the self-test condition were significantly more accurate than those in the answer condition, $t(132) = 5.07$, $p < .001$, $d = 0.88$, $BF_{10} = 10855$.

Again, correlations were conducted between the types of questions participants in the self-test condition created and their performance on the final test. Based on the results from Experiment 1, I hypothesized that the number of questions generated and the proportion of questions that targeted final test material would correlate with final test performance. Following these hypotheses, in Experiment 2, both the proportion of generated questions that targeted final

test material (r = .49, p < .001, BF_{10} = 1,110.33) and proportion of questions answered correctly (r = .37, p = .002, BF_{10} = 18.04) significantly correlated with final test performance (no. of questions: r = .20, p = .11, BF_{01} = 1.92; prop. factual: r = -.18, p = .13, BF_{01} = 6.33). A linear regression analysis was conducted with final test performance for the self-test condition regressed on the number of questions created, proportion of factual questions, proportion of questions that targeted final test material, and accuracy. Similar to Experiment 1, the proportion of generated questions that targeted final test material (b = 43.27, β = 0.41, SE = 12.31, p < .001, BF_{10} = 47.62) significantly predicted final test performance while controlling for the other factors, such that a .1 increase was associated with a 4.3-percentage point increase in final test performance. The number of questions created also significantly predicted final test performance (b = 1.42, β = 0.25, SE = 0.58, p = .02, BF_{10} = 3.82), such that creating one additional question was associated with a 1.4-percentage point increase in final test performance. Proportion of factual questions (b = -9.06, β = -0.14, SE = 6.53, p = .17, BF_{01} = 1.44) and proportion of questions answered correctly (b = 17.45, β = 0.19, SE = 10.92, p = .12, BF_{01} = 1.09) did not significantly predict final test performance. For the answer condition, a correlation between final test performance and the proportion of questions participants answered correctly was calculated. This correlation suggested that the more experimenter-generated questions participants answered correctly, the better they performed on the final test (r = .63, p < .001).

Global JOLs

A 3 (strategy: self-test, answer, reread) x 2 (passage: ice age, monetary policy) mixed-factor ANOVA was conducted to analyze participants' JOL ratings, with passage manipulated within-participants and strategy between-participants. Overall, participants predicted they would perform better on the ice age questions (M = 66.58, SE = 1.32) than the monetary policy

questions (M = 62.49, SE = 1.29), $F(1, 199) = 13.14$, $p < .001$, $\eta_{2p} = .06$, $BF_{10} = 49.69$. On average, participants' predictions differed based on study strategy, $F(2, 199) = 3.73$, $p = .03$, $\eta_{2p} = .04$, $BF_{10} = 1.95$. The passage x strategy interaction did not reach conventional significance, $F(2, 199) = 2.57$, $p = .08$, $\eta_{2p} = .03$, $BF_{01} = 1.98$. However, because the interaction was close to significant, follow-up tests were conducted separately for the ice age and monetary policy. For the ice age passage, participants' predictions in the self-test (M = 64.12, SE = 2.30) and answer (M = 67.76, SE = 2.40) conditions did not differ, $t(132) = 0.49$, $p = .62$, $d = 0.09$, $BF_{01} = 4.83$. Ice age predictions for those in the reread (M = 69.85, SE = 2.15) condition tended to be higher than the self-test condition, although this difference was not significant, $t(134) = 1.82$, $p = .07$, $d = 0.31$, $BF_{01} = 1.21$. Predictions did not differ between the answer and reread conditions for the ice age final test, $t(132) = 1.27$, $p = .21$, $d = 0.22$, $BF_{01} = 2.59$. For the monetary policy passage, participants in the answer (M = 64.39, SE = 2.45) and reread conditions (M = 66.47, SE = 1.95) predicted they would perform significantly better than those in the self-test condition (M = 56.62, SE = 2.30), $t(132) = 2.31$, $p = .02$, $d = 0.40$, $BF_{10} = 2.06$ and $t(134) = 3.27$, $p = .001$, $d = 0.56$, $BF_{10} = 21.19$, respectively. The answer and reread conditions did not differ in their predictions for the monetary policy final test, $t(132) = 0.66$, $p = .51$, $d = 0.12$, $BF_{01} = 4.42$.

In terms of absolute calibration, participants were well-calibrated and again slightly underconfident in how well they would perform on the final tests. A 3 (strategy: self-test, answer, reread) x 2 (passage: ice age, monetary policy) mixed-factor ANOVA was conducted to analyze differences in metacognitive calibration. On average, participants were better calibrated for the monetary policy passage (M = -0.74, SE = 1.46) than the ice age passage (M = -6.20, SE = 1.48), $F(1, 199) = 11.13$, $p = .001$, $\eta_{2p} = .05$, $BF_{10} = 10{,}491.62$. Overall, participants' calibration did not differ between the self-test (M = -3.38, SE = 2.11), answer (M = -4.02, SE = 2.14), and

reread conditions (M = -3.02, SE = 2.11) conditions, $F(2, 199) = 0.06$, $p = .95$, $\eta_{2p} = .001$, $BF_{01} = 4.50$. However, the passage x strategy interaction was significant, although the Bayes factor favored the null, $F(2, 199) = 5.56$, $p = .004$, $\eta_{2p} = .05$, $BF_{01} = 2.70$. Follow-up tests indicated that, for the ice age passage, participants in the answer (M = -7.42, SE = 2.29) and reread conditions (M = -8.68, SE = 2.67) were more underconfident than the self-test condition (M = -2.50, SE = 2.70), although these differences were not significant, $t(132) = 1.39$, $p = .17$, $d = 0.24$, $BF_{01} = 2.26$ and $t(134) = 1.63$, $p = .11$, $d = 0.28$, $BF_{01} = 1.64$, respectively. Calibration for the ice age passage did not differ between the answer and reread conditions, $t(132) = 0.36$, $p = .72$, $d = 0.06$, $BF_{01} = 5.10$. For the monetary policy passage, those in the self-test (M = -4.26, SE = 2.46) condition were more underconfident than the answer (M = -0.61, SE = 2.68) and reread conditions (M = 2.65, SE = 2.47), although the difference between self-test and answer conditions was not significant, $t(132) = 1.01$, $p = .32$, $d = 0.17$, $BF_{01} = 3.41$ and $t(134) = 1.92$, $p = .049$, $d = 0.34$, $BF_{10} = 1.09$. Calibration for the monetary policy passage did not differ between the answer and reread conditions, $t(132) = 0.89$, $p = .37$, $d = 0.27$, $BF_{01} = 3.76$.

Discussion

Contrary to hypotheses, final test performance was poorest for participants who answered their own questions (self-test condition), relative to those who answered experimenter-created questions (answer condition) and those in the reread condition. Similar to Experiment 1, analyses of the questions generated by participants in the self-test condition revealed that the more questions participants generated and the more those questions overlapped with final test material, the better they performed on the final test. A significant correlation between accuracy when answering experimenter-generated questions and final test performance also suggested that the

better participants perform on a provided practice test, the better they perform on a final test, even when the final test is composed of different questions.

GENERAL DISCUSSION

The current study sought to understand whether students still benefitted from testing if they created their own test questions. Experiment 1 added to the current literature on self-testing by exploring whether participants benefitted more from self-testing when they delayed answering their questions compared to generating questions and answers simultaneously (e.g., Weinstein et al., 2010). If students delay answering their questions, they can practice retrieval and have a larger delay between study and a first test, both of which have been shown to improve memory performance (Jacoby, 1978; Rawson, Vaughn, & Carpenter, 2015; Rowland, 2014). Experiment 2 then compared a self-testing strategy to answering experimenter-generated questions (typical in testing effect studies) and restudying. Because those who self-test both participated in generating questions and retrieving the answers to those questions, their performance was predicted to be better than the other two strategies. However, results did not support predictions for either experiment.

Contrary to predictions, Bayesian analyses provided positive evidence that final test performance did not differ when participants answered their generated questions after a delay compared to generating and answering questions simultaneously in Experiment 1. Indeed, participants who generated and answered questions simultaneously performed slightly (although not significantly) better than the comparison condition on factual final test questions. Thus, adding a retrieval opportunity by inserting a delay between generating one's own test questions and attempting to answer those questions did not appear to benefit learning. In Experiment 2, self-testing again did not lead to learning benefits and resulted in poorer test performance compared to answering experimenter-generated questions and rereading. This suggests that

students might not be able to benefit from testing themselves if they must create their own testing materials. No noteworthy differences in metacognitive predictions (i.e., global JOLs) were found between conditions in either experiment.

These findings contrast with other studies evaluating self-testing (e.g., Foos, Mora, & Tkacz, 1994; Weinstein et al., 2010). The discrepancies may reflect differences in the delays between the study activities and the final test; Weinstein et al. (2010) and Foos et al. (1994) both administered their final test two days after initial study. Nevertheless, several studies found benefits of self-testing even within a final test administered immediately after study (Bugg & McDaniel, 2012; Denner & Rickards, 1987; Weinstein et al., 2010 Experiment 1). Contradictory findings could also be explained by prior studies providing guidance to participants on what types of questions they should generate (Bugg & McDaniel, 2012; Denner & Rickards, 1987; Weinstein et al., 2010), whereas the current study did not provide any guidance. Foos and colleagues (1994) still detected benefits of self-testing (compared to having participants answer pre-made questions) without providing any guidance in question generation. However, Foos et al. (1994) permitted participants to keep their materials for further study before the final test, making it difficult to determine how participants used those materials within the two days before the final test. Overall, this may suggest that students need training in what types of questions to create before they can reap benefits from self-testing.

Furthermore, the current results do not necessarily disagree with the results found by Weinstein and colleagues (2010), whose procedures were most similar to those of the current study. To elaborate, Weinstein and colleagues (2010, Experiment 1) found that self-testing was just as beneficial as answering experimenter-generated questions, and both conditions outperformed a restudy condition. However, Weinstein and colleagues (2010) allowed

participants to spend as much time as they wanted completing these study activities (i.e., study was self-paced), and they found that those in the self-testing condition spent at least twice as long as the other two conditions. Therefore, the benefits of self-testing may have been reduced in the current study because all study activities were confined to the same amount of time.

Another key finding of this study is that self-testing was more beneficial when participants generated more questions and a higher proportion of those questions targeted final test material. Importantly, these factors were not a significant predictor of final test performance when participants generated questions and answers at the same time, although the Bayes factors were inconclusive (Experiment 1 simultaneous condition). This may suggest that it is important to retrieve final test material and not simply review it. However, these results are correlational and require experimental evidence to fully support this conclusion. Future research could explore how telling participants how many questions to generate and providing participants with some knowledge of the final test influences the effects of self-testing.

In both experiments, the proportion of generated questions participants answered correctly (i.e., accuracy) was significantly correlated with final test performance but did not predict performance when entered into a linear regression with the other factors. It could be that there are some interactions between accuracy and the other factors that reduce the effects of accuracy when controlling for the other factors. It could also be that there was not enough variability in accuracy to detect a linear relationship between accuracy and final test performance since participants answered almost 85% of their generated questions correctly. Some evidence from the current study might suggest that other components of the questions that participants generated are also associated with final test performance (e.g., proportion of factual questions if questions and answers are generated simultaneously). However, evidence from the current study

was not strong enough to make recommendations about these components. Inter-rater reliabilities were also low, particularly for scoring generated questions as factual or conceptual. This may have reduced the ability to detect the effects of some of these factors, and future research is needed to develop more reliable scoring methods.

Implications for Self-Testing Research

In the current study, it was hypothesized that self-testing would lead to the greatest learning benefits because self-testing allowed both generation (i.e., writing questions) and retrieval practice (i.e., answering questions from memory). It is unclear why these generation and retrieval opportunities did not benefit later memory for those in the self-test conditions. As alluded to previously, it is possible that self-testing did not benefit learning because the delay between the study activities and final test was not long enough for testing effects to appear (Roediger & Karpicke, 2006b; Rowland, 2014). Indeed, no testing benefits were observed in the current study, even when participants answered experimenter-generated questions in Experiment 2. Therefore, self-testing should be explored with a longer delay before any strong conclusions can be made. It is also important to note that the detriments of self-testing were stronger for factual final test questions than conceptual questions in the current study. Thus, self-testing might have different benefits depending on the type of test questions.

It is also possible that self-testing benefits did not accrue because the delay between answering questions and attempting to retrieve the answers was not long enough to allow memory consolidation, as retrieval practice becomes more beneficial with longer delays between initial study and a practice test (Jacoby, 1978; Rawson, Vaughn, & Carpenter, 2015). However, Jacoby (1978) found larger testing benefits with only a delay of 2 minutes between study and a practice test, although Jacoby (1978) used related word pairs rather than passages.

A longer delay may be particularly important during self-testing because participants most likely had the correct answer in mind when they were generating test questions using the passage. Furthermore, participants generated fewer questions (7 on average) for which they must remember the answers compared to typical memory experiments (e.g., Jacoby, 1978) where participants must often remember 20 or more items. Indeed, although accuracy was significantly lower when participants had to answer their questions from memory than when they had access to the passage (Experiment 1), accuracy was still high during self-testing (83-84%). Thus, it may take more time for participants to begin to forget the answers to their generated questions than the 12 minutes allotted in the current study, and testing that requires more effort to recall answers affords larger benefits (Halamish & Bjork, 2011). This explanation fits well with theories of testing that argue that testing encourages one to elaborate upon other material related to the correct answers, which can later serve as mediators to return to the correct answers (elaborative retrieval hypothesis, Carpenter, 2009, 2011; Pyc & Rawson, 2010). Specifically, if participants in the self-test condition can easily access the correct answer to their generated questions, they would spend little time elaborating upon other material from the passage to reach their answer. Thus, information on the final test that was not directly targeted in generated questions would not be strengthened by the self-testing opportunity, and participants would have few mediators in memory to bolster retrieval (see Karpicke, Lehman, & Aue, 2011; Rickard & Pan, 2018, for other perspectives on the benefits of testing).

In the current experiments, self-testing sometimes appeared to harm memory performance compared to the other study strategies. Although I have no evidence to verify either account, there are two memory phenomena that could explain why self-testing might be detrimental for memory. First, those who self-tested may have suffered from retrieval-induced

forgetting, a phenomenon where retrieving a subset of information reduces one's ability to remember other information (for a review, see Murayama, Miyatsu, Buchli, & Storm, 2014). In the present study, retrieving information to answer their own questions might have interfered with participants' access to the answers to final test questions in the self-testing conditions. This explanation has some limitations in the current circumstances, though. Specifically, it is unclear why answering one's own questions would interfere more than answering provided questions (as in the answer condition), particularly when the overlap between practice test and final test questions was similar in both conditions. In addition, retrieval-induced forgetting is typically found between two pieces of information that are associated with the same memory cue (e.g., *cat-dog* would interfere with retrieving *cat-kitten*). However, in the present experiments, the memory cues provided (i.e., the specific question prompts) were not associated with multiple possible answers. Furthermore, having participants integrate competing information appears to block retrieval-induced forgetting (Anderson & McCulloch, 1999; Chan, 2009). With the highly-interrelated material within each passage used in the present study, it is unclear whether retrieval-induced forgetting would occur. Another possible explanation is that the self-testing procedures required participants to divide their attention between creating quality test questions and learning the material. Subsequently, this led to reduced learning in the self-testing condition since dividing attention while encoding seems particularly harmful (Anderson, Craik, & Naveh-Benjamin, 1998; Craik, Govoni, Naveh-Benjamin, & Anderson, 1996). This explanation may explain why self-testing was harmful in the current study, especially because participants were under time pressure to create their questions. Nevertheless, future research is needed to directly examine these accounts using self-testing procedures.

The current study also found that self-testing was more beneficial when participants generated more questions and a higher proportion of questions targeted final test material. Consequently, self-testing might not have benefitted memory because participants, on average, did not generate enough questions that overlapped with final test questions (average proportion overlap was only around 0.35). Self-testing benefits may appear if participants generated questions that more closely aligned with final test questions, consistent with prior research suggesting that the benefits of testing may be confined to material that appears on the test (Hinze, Wiley, & Pellegrino, 2013; Pan & Rickard, 2018; Pilotti et al., 2009; but see Butler, 2010; Chan, McDermott, & Roediger, 2006). For example, Pan and Rickard (2018) found in a meta-analysis on testing that, while testing benefits could transfer in some circumstances, a benefit of testing was not found for studied materials that were not initially tested (d = 0.16 [-0.10, 0.43]). Future research could further explore this possibility by manipulating participants' exposure to final test material.

Practical Implications

The present study suggests that students may not be able to benefit from testing if they are not given testing materials (and thus must self-test), at least compared to other study methods. Therefore, instructors and researchers may need to exercise caution when recommending that students test themselves. One possible solution that could allow students to benefit from self-testing is to provide information about what material will be covered on the final test, allowing them to generate more questions relevant to final test material. This could take the form of study guides highlighting key concepts or instruction on how to identify important information. However, experimental evidence is needed to verify this suggestion. Future self-testing research is also needed in more educationally relevant situations beyond the

use of short passages. It is possible that participants could more easily identify important material from a semester's worth of lectures than a one-page passage. Furthermore, motivation to do well on course exams might encourage students to generate better questions and test themselves more diligently. Nevertheless, given the current evidence, instructors are well-advised to provide testing materials so that students can test themselves effectively. Until evidence is available on how self-testing without instruction can benefit learning, it would be a better use of students' time to use free recall methods to test themselves, which have been supported (Roediger & Karpicke, 2006b; Rowland, 2014) rather than writing their own test questions.

REFERENCES

Agarwal, P. K., D'Antonio, L., Roediger III, H. L., McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131-139.

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & 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

Agarwal, P. K., & Roediger, H. L. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*(8), 836–852. https://doi.org/10.1080/09658211.2011.613840

Anderson, N. D., Craik, F. I. M., & Naveh-Benjamin, M. (1998). The attentional demands of encoding and retrieval in younger and older adults: 1. Evidence from divided attention. *Psychology and Aging, 13*(3), 405-423.

Anderson, M. C., & McCulloch, K. C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(3), 608-629.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201-210.

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.

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

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

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

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(6), 1547.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279-283.

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

Chan, J. C. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language, 61*(2), 153-170.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553

Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General, 125*(2), 159-180.

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

Dunlosky, J., & Metcalfe, J. (2008). *Metacognition*. Sage Publications.

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.

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. *The Quarterly Journal of Experimental Psychology: Section A, 55*(2), 505-524.

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,* 175-191.

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

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 801.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69,* 151-164.

IBM Corp. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

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*, 639-667.

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. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of learning and motivation* (Vol. 61, pp. 237-284). Academic Press.

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

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.

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.

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

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger, H. L., & 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. *The Psychology of Learning and Motivation: Advances in Research and Theory, 45,* 175-214.

Murayama, K., Miyatsu, T., Buchli, D., & Storm, B. C. (2014). Forgetting as a consequence of retrieval: A meta-analytic review of retrieval-induced forgetting. *Psychological Bulletin, 140*(5), 1383-1409.

Nelson, T. O. (1996). Consciousness and metacognition. *American Psychologist*, *51*(2), 102-116.

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

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

Pilotti, M., Chodorow, M., & Petrov, R. (2009). The usefulness of retrieval practice and review-only practice for answering conceptually related test questions. *Journal of General Psychology, 136,* 179-204.

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

Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, *19*(4-5), 559-579.

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.

Rhodes, M. G. (2019). Metacognition. *Teaching of Psychology*, *46*(2), 168-175.

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.

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, *25*(3), 847-869.

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

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

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

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239. https://doi.org/10.1037/a0017678

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/10.2307/1170607

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.

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

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

Spitzer, H. F. (1939). Studies in retention. *The Journal of Educational Psychology, 30*(9), 641-656.

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

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638.

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., & 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

APPENDICES

Appendix A. Passages used for study material

MONETARY POLICY

The U.S. is the largest economy in the world. Therefore, the U.S. dollar is considered a stable value. Many factors affect the strength of the dollar relative to other currencies, including the trade deficit or surplus, the size of the Government deficit, interest rates, and the strength of the U.S. economy. The strength of the dollar is also affected by the monetary policy imposed by the Federal Reserve System.

The Fed, as it is called, is the central bank of the U.S. The Fed's duties include conducting the nation's monetary policy by influencing money and credit conditions in the economy in pursuit of full employment, stable prices, and promoting the stability of the financial system. The Fed conducts monetary policy using three major tools. It buys and sells U.S. Treasury and federal agency securities in the open market; it sets the discount rate, which is the interest rate that banks pay the Fed to borrow money; and it sets reserve requirements, which is the amount of funds that banks must hold in reserve against deposits made by their customers. Monetary policy can affect short-term interest rates, foreign exchange rates, long-term interest rates, the amount of money and credit, and, ultimately, a range of economic variables, including employment, output, and prices of goods and services.

Monetary policy works by affecting the amount of money circulating in the economy. The Fed can change the amount of money that banks are holding in reserves by buying or selling existing U.S. Treasury bonds. The Fed sells bonds, which decreases banks' reserves and their ability to make loans. As banks lend less and the money supply decreases, interest rates increase. The Fed buys bonds, which decreases the reserve requirement and increases banks' ability to make loans. As banks lend more and the money supply increases, interest rates decrease.

Lower interest rates mean that consumers pay less when they charge purchases. They may be more willing to spend. They may even buy expensive goods, like cars and refrigerators, to take advantage of lower interest rates. As the demand for more goods increases, either businesses will increase production to satisfy the demand or prices of goods will increase.

Lower interest rates may encourage businesses to expand to meet the increasing consumer demand. They may run extra shifts or build new factories. This may create new jobs. As workers who were previously unemployed return to the workforce, they will eventually spend their paychecks. This too will increase the demand for goods. Again, either businesses will increase production or prices of goods will increase.

Sometimes consumer spending is so great that production can't keep up with demand. The excessive demand for goods can lead to inflation. Inflation can also occur as a result of increasing the amount of money circulating in the economy. Inflation means dollars are worth less. The Fed will try to keep inflation in check.

Inflation may undermine the strength of the economy. Inflation increases the difficulty of forecasting prices and costs of doing business, so it discourages businesses from planning and investing. People also may be uncertain and reluctant to spend. Both of these factors could reduce the long-term level of economic growth. Inflation also increases the cost of carrying out transactions. Inflation in U.S. increases cost of U.S. goods; therefore, imports increase and exports decrease.

ICE AGE

An ice age is a period of time—usually millions or tens of millions of years—when vast glaciers cover as much as a third of the Earth's land surface. Average global temperatures can drop by as many as 12 degrees Celsius overall. The latest Ice Age began about 2.5 million years ago, and ended approximately 15,000 years ago. Average global temperatures decreased by approximately 8 degrees Celsius. Sea-level was lowered substantially due to the amount of water that was frozen in the glaciers. Ice core analysis indicated there were reduced amounts of carbon dioxide in the atmosphere. Giant ice sheets that originated at the North Pole advanced and retreated many times in North America and Europe. The movement of the glaciers coincided with cycles of warm and cold periods in the Earth's temperature. Throughout history, cycles of changes in global temperatures usually occur every 100,000 years or so. Each cycle consists of a long, generally cold period during which the entire Earth cools, followed by a relatively short warm period during which Earth warms up rapidly.

We are now in a warming period that has lasted more than 10,000 years, which is longer than many of the previous warming intervals. Warm temperatures over the last century have been attributed to the increased man-made release of carbon dioxide. CO2 prevents long-wave radiation from escaping from the Earth into space. The more CO2 there is in the atmosphere, the more long-wave radiation is kept from leaving the Earth. The more radiation that is trapped, the hotter the Earth becomes. This trapping of radiation works like a gardener's greenhouse, and this phenomenon is commonly known as the 'Greenhouse Effect'.

Carbon dioxide (CO2) is a common gas that is contained in the Earth's atmosphere. CO2 is released whenever organic matter decays, and when carbohydrates are broken down by plants and animals in the process of respiration. The burning of fossil fuels also releases large amounts of CO2.
Carbon dioxide can be removed from the atmosphere also. CO2 can be combined with other minerals in the ground and buried, or absorbed into the oceans or trapped in ice and snow. Green plants also absorb carbon dioxide from the atmosphere, and through the process of photosynthesis, form carbohydrates.

The release and storage of CO2 is a natural process and proceeds in a circular fashion. For example, plants convert CO2 from the atmosphere into carbohydrates, which they use to grow. When the plant dies, the carbohydrates that the plant made are converted back into CO2 through the decaying process. At any time it is possible for there to be more CO2 being stored than released, and also vice-versa. Thus, the amount of CO2 in the atmosphere can fluctuate.

The amount of radiation that the Earth receives from the sun can also fluctuate. Fluctuations in solar radiation can change average global temperatures by up to 4-6 degrees Celsius. The amount of solar radiation that the sun emits can vary. For example, an increase in the amount of sunspots on the Sun's surface has been correlated with an increase in the amount of energy that is output by the Sun. The amount of solar radiation energy that actually reaches the Earth is influenced by the distance the Sun's rays must travel to reach the Earth, and also the angle at which the Sun's rays strike the surface of the Earth. The farther that light rays travel, the less energy will be contained in the Sun's rays. Cyclical changes in the shape of the Earth's orbit around the Sun influence how far the Sun's rays have to travel. When the Earth's orbit is extremely oval-shaped, the distance from the Earth to the Sun can vary greatly. The farther the Earth from the Sun, the less solar radiation reaches the Earth.

Other cyclical changes in the tilt of the Earth's axis vary the angle at which light energy strikes the surface of the Earth in a given region. If the Sun's rays strike the Earth at a great angle, for example as it does at the North Pole, solar energy is reflected off of the Earth, rather than being absorbed into it. When light strikes a region at a great angle, not very much of the solar radiation is absorbed by the Earth. When a region receives less solar radiation, there is less energy to warm that area. Less heat energy leads to cooler temperatures. Cooler temperatures can cause more snow and ice to form. Snow and ice can reflect what little solar energy reaches the surface of the Earth back into space. The formation of snow and ice can also steal large amounts of CO2 from the atmosphere and trap it in a frozen, solid form.

Through the course of millions of years, the surface of the earth also changes. Continents collide and split apart, mountains are uplifted and eroded, volcanoes erupt, and ocean basins open and close. These changes alter the size and elevation of the continental land masses. Different elevations of land masses support different types of climates. Land masses at high elevations usually support colder climates. For example, the tops of mountains high above sea-level are usually covered in snow and plains at sea-level are usually warm. These events also release minerals in the Earth's crust. These minerals are often carried by rivers to the sea, where they can be absorbed into the atmosphere. In this way, CO2 and other compounds can be released from their solid mineral forms and introduced into the atmosphere.

Changes in geography also affect the ocean by the opening and closing of gateways that carry currents. A change in ocean currents affects how water flows from one area of the Earth to another. A majority of Earth's heat energy is transferred around the globe by the ocean currents. More heat energy is stored in the oceans than in the atmosphere. Surface ocean currents assist in the transfer of heat from low to high latitudes.

The Earth might be due for another Ice Age. However, not all scientists are convinced that there will be one. Some believe that the man-made release of CO2 into the atmosphere might prevent the Earth from cooling sufficiently. On the other hand, some scientists believe that the recent global warming might actually increase the magnitude of the cooling period.

Appendix B. Final test questions

MONETARY POLICY

F - Factual Question
C - Conceptual Question

1. (F) Which country has the world's largest economy?
    a. China
    b. United Arab Emirates
    c. Japan
    d. United States

2. (F) What is the Fed?
    a. the central bank of the U.S.
    b. the Department of the Treasury
    c. the Department of Commerce
    d. the Securities and Exchange Commission

3. (F) Which of the following does monetary policy affect?
    a. the amount of tariffs on foreign goods
    b. the amount of unemployment compensation available to citizens
    c. the amount of money circulating in the economy
    d. the amount of money printed by the U.S. Treasury

4. (F) What does inflation in the U.S. tend to result in?
    a. decrease in U.S. exports
    b. decrease in imports of foreign goods
    c. increase in consumer spending
    d. increase in the stability of the U.S. dollar

5. (C) Which of the following is a cause of inflation?
    a. long-term interest rates rise above short-term interest rates
    b. production can't keep up with consumer demand
    c. production costs rise faster than the demand for goods
    d. price of stocks rise faster than earnings

6. (F) If interest rates are lowered, consumers are more likely to
    a. buy more cars.
    b. buy more food.
    c. save more.
    d. travel less.

7. (C) Which of the following is **NOT** a likely result of lower interest rates?
    e. prices of goods will decrease
    f. consumers are willing to spend more
    g. consumers will buy more expensive goods
    h. businesses will decrease production

8. (C) Which is likely to occur when the Fed increases the reserve requirement?
    a. consumer spending will increase
    b. interest rates will increase
    c. local banks will increase lending
    d. the economy will grow

9. (C) What might the Fed do if it wants to affect the economy in a way that is similar to that of lowering income taxes?
    a. decrease loans to consumers and businesses
    b. decrease the reserve requirement
    c. increase the discount rate
    d. decrease the money supply

10. (C) Unemployment will tend to decrease when
    a. interest rates decrease.
    b. consumer demand decreases.
    c. business investment decreases.
    d. the money supply decreases.

ICE AGE

F - Factual Question
C - Conceptual Question

1. (F) How much of the earth is covered by glaciers during an ice age?
   A. less than 10 percent
   B. about a third
   C. over half
   D. almost all

2. (F) How much do average global temperatures lower during an ice age?
   A. 4-6 degrees Celsius
   B. 8-12 degrees Celsius
   C. 20-25 degrees Celsius
   D. 40-50 degrees Celsius

3. (F) How long has the current warming period lasted?
   A. 2.5 million years
   B. 100,000 years
   C. 50,000 years
   D. 10,000 years

4. (C) What is the greenhouse effect?
   A. the absorption of CO2 by growing plants
   B. the trapping of radiation due to CO2
   C. the increase in heat of the earth due to sunspots
   D. the increase in burning of fossil fuels

5. (F) What is **NOT** true of CO2?
   A. it is a common gas in the earth's atmosphere
   B. it is released by decaying plants
   C. it is released by burning fossil fuels
   D. it cannot be removed from the atmosphere

6. (F) What is true about ice ages?
   A. Regional temperatures within an Ice Age do not fluctuate.
   B. **Sea levels are lower during an Ice Age.**
   C. All regions of the Earth are covered with ice.
   D. Ice ages occur because the temperature of the core of the Earth cools.

7. (C) What is true of the oceans?
   A. Changes in ocean currents could cause glaciers to form or retreat.
   B. Ocean currents follow the same path around the globe as they did 2.5 million years ago.
   C. More heat energy is stored in the atmosphere than the oceans.
   D. The oceans keep a constant temperature.

8. (C) Higher levels of CO2 in the atmosphere lead to
   A. higher sea levels.
   B. the creation of mountain ranges.
   C. the formation of more ice and snow.
   D. changes in the earth's surface.

9. (C) What can cause less solar radiation to reach earth?
   A. when the Earth's orbit is closer to the Sun
   B. sunspots
   C. the formation of more mountain ranges
   D. the Earth's tilt

10. (C) What is true of earth's temperature?
    A. it goes through long warming cycles followed by short cooling cycles
    B. temperature changes are random and unpredictable
    C. the temperature increases with the amount of long-wave radiation in our atmosphere
    D. deforestation lowers the earth's average temperature

Appendix C. Diagrams of procedure

Figure C1. Diagram of Experiment 1 procedure for the self-test (answer generated questions after a delay) and the simultaneous (generate and answer questions simultaneosuly) conditions. White boxes represent tasks for Passage 1. Gray boxes represent tasks for Passage 2.

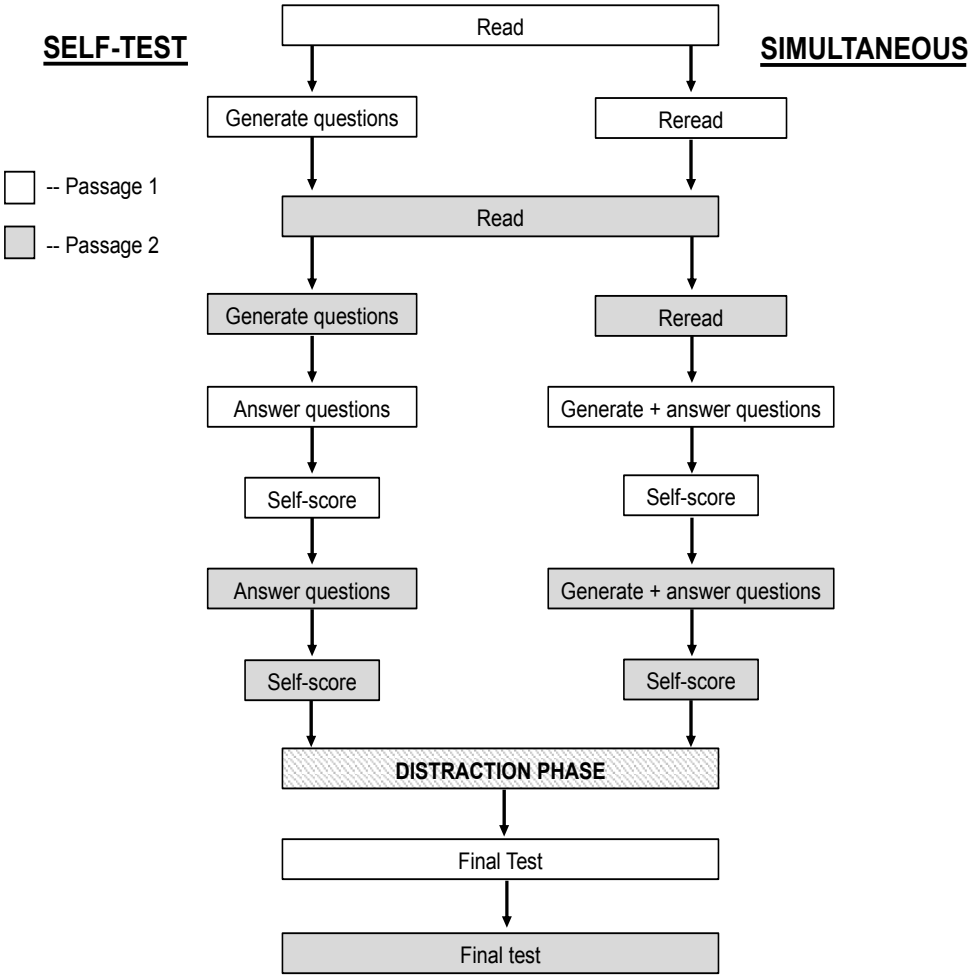

*Figure C2***.** Diagram of Experiment 2 procedure for the self-test (generate questions and answer them after a delay), answer (answer experimenter-generated questions) and reread (restudy material, controlling for total exposure time) conditions. White boxes represent tasks for Passage 1. Gray boxes represent tasks for Passage 2.