Journal of Experimental Psychology: Applied
2010, Vol. 16, No. 3, 308–316

1076-898X/10/$12.00 DOI: 10.1037/a0020992

A Comparison of Study Strategies for Passages: Rereading, Answering Questions, and Generating Questions

Yana Weinstein, Kathleen B. McDermott, and Henry L. Roediger III
Washington University in St. Louis

Students are often encouraged to generate and answer their own questions on to-be-remembered material, because this interactive process is thought to enhance memory. But does this strategy actually work? In three experiments, all participants read the same passage, answered questions, and took a test to get accustomed to the materials in a practice phase. They then read three passages and did one of three tasks on each passage: reread the passage, answered questions set by the experimenter, or generated and answered their own questions. Passages were 575-word (Experiments 1 and 2) or 350-word (Experiment 3) texts on topics such as Venice, the Taj Mahal, and the singer Cesaria Evora. After each task, participants predicted their performance on a later test, which followed the same format as the practice phase test (a short-answer test in Experiments 1 and 2, and a free recall test in Experiment 3). In all experiments, best performance was predicted after generating and answering questions. We show, however, that generating questions led to no improvement over answering comprehension questions, but that both of these tasks were more beneficial than rereading. This was the case on an immediate short-answer test (Experiment 1), a short-answer test taken 2 days after study (Experiment 2), and an immediate free recall test (Experiment 3). Generating questions took at least twice as long as answering questions in all three experiments, so although it is a viable alternative to answering questions in the absence of materials, it is less time-efficient.

Keywords: study strategies, testing effect, metacognition

Supplemental materials: http://dx.doi.org/10.1037/a0020992.supp

Much recent work has focused on optimizing students' study strategies. Some strategies are more successful than others in producing long-term retention. The most consistently effective technique seems to be self-testing (Carpenter, Pashler, & Vul, 2006; Karpicke & Roediger, 2007; McDaniel, Roediger, & McDermott, 2007; see Roediger & Karpicke, 2006a, for a review); taking a test on material sometimes more than doubles retention compared to control conditions involving unrelated tasks or restudy of the material. This effect has been demonstrated with a wide range of materials such as paired associates (Carrier & Pashler, 1992; Karpicke & Roediger, 2008), as well as more complex materials such as passages (Kang, McDermott, & Roediger, 2007; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b) and lectures (Butler & Roediger, 2007). It has also been successfully implemented in a real-world classroom setting (McDaniel et al., 2007). However, despite the demonstrated benefits of self-testing, students do not tend to implement this technique when left to their own devices. Karpicke, Butler, and Roediger (2009) asked college students about their study strategies and found that rereading was reported as a strategy over eight times more often than self-testing (84% and 11% of students surveyed, respectively). Kornell and Son (2009) also found that when self-testing does occur, the motivation is diagnostic purposes rather than cognizance of the direct benefits of testing.

Why does self-testing occur so infrequently? One probable explanation is that students are not aware of the benefits of self-testing. Evidence for this explanation comes from predictions made by students about future memory performance following testing or restudy. Karpicke and Roediger (2008) found that students did not expect any change in retention following retrieval practice, and Agarwal, Karpicke, Kang, Roediger, and McDermott (2008) found that students predicted the same level of performance after testing as after restudy. However, another reason why students may not engage in self-testing is that they may not have access to the resources required to implement this technique. Testing schedules and materials used in laboratory studies are produced by the experimenter, and when these studies are extended to classroom settings, students are provided with practice tests rather than asked to develop their own (e.g., McDaniel et al., 2007). Testing as retrieval practice rather than a diagnostic tool has not yet been accepted in mainstream education, so students may simply not have sets of questions to use for self-testing. In this article, we set out to investigate whether an alternative, more easily implemented technique could yield comparable benefits to taking a practice test. In the absence of practice test questions, could generating and answering one's own practice questions lead to similar benefits for retention?

Yana Weinstein, Kathleen B. McDermott, and Henry L. Roediger III, Department of Psychology, Washington University in St. Louis.

Support for this research was provided by a James S. McDonnell Foundation 21st Century Science Initiative grant: Bridging Brain, Mind and Behavior/Collaborative Award. Thanks to Kristy Duprey for assistance with data collection.

Correspondence concerning this article should be addressed to Yana Weinstein, Department of Psychology, Box 1125, Washington University, One Brookings Drive, St. Louis, MO 63130. E-mail: y.weinstein@wustl.edu

Another reason why students may be reluctant to engage in self-testing is the mental effort involved in retrieving information from memory. Whereas it has been argued that this act of effortful retrieval is crucial to promoting long-term retention (e.g., Gardiner, Craik, & Bleasdale, 1973; Jacoby, 1978), there is also some evidence that self-testing is beneficial to later memory even when the effort during the retrieval process is minimized. Agarwal et al. (2008) showed that the benefits of testing for later retention extend to open-book practice tests—that is, when students practice questions with the material in front of them, but later take a traditional closed-book test. Open-book tests presumably require less retrieval effort and are also preferred by students (Ben-Chaim & Zoller, 1997).

In the present experiments, an additional self-testing technique is introduced. In this technique, students generate and also answer their own questions after reading a passage. Performance on a subsequent closed-book test for passages studied in this manner was compared with two control conditions: one in which participants answered questions set by the experimenter (analogous to an open book test), and one in which they reread the passage. The effects of these three study strategies were examined on a short-answer test taken 30 to 45 min after initial study (Experiment 1), a short-answer test taken 2 days after study (Experiment 2), and a free recall test taken 30 to 45 min after study (Experiment 3).

The main aim of this article was to determine whether generating questions could provide benefits equal to those of self-testing using prepared materials, which may not be available to all students. However, there is some evidence to suggest that this condition might produce benefits over and above self-testing using materials prepared by a third party. Searching the text for information to generate questions could provide benefits via three processes: generation, elaboration, and synthesis. First, the generation effect reveals that self-generated information is remembered better than information that is passively encoded (Slamecka & Graf, 1978), at least in a mixed-list design (McDaniel & Bugg, 2008). Of course, the answer-questions condition also involves generation, but it could be that generating both the questions and the answers will produce additional benefits to retention. Second, one account of the testing effect ascribes the benefits of testing to the increased elaboration that results from having to retrieve an answer (Carpenter, 2009; Carpenter & DeLosh, 2006). In our task, selecting information for the questions could produce greater elaboration leading to better retention of the material. Third, preparing to relay information to a peer has been found to improve retention because it promotes synthesis of the material (Nestojko, Bui, Kornell, & Bjork, 2009). In our task, scanning the text with a view to finding appropriate material for questions could produce a similar effect.

An important motivation for testing this technique is that it is often recommended by educators, who may believe that generating questions promotes deeper engagement with and understanding of the material, thus promoting retention (see, for instance, Robinson [1970], who proposed a study technique called SQ3R: survey, question, read, recite, review). Up until now, this recommendation has seldom been tested, apart from in large-scale studies involving lengthy training procedures to get students accustomed to the technique (e.g., Martin, 1985; see McDaniel, Howard, & Einstein, 2009). Our primary goal was to empirically test the effectiveness of the advice in a simple paradigm.

In addition to this primary goal, the article also addressed two additional questions: metacognition and efficiency of the study strategies. First, if generating and answering one's own questions is beneficial to later memory, this is only helpful insofar as students are aware of the benefits and choose to use the technique. We thus collected participants' predictions of how they would perform in each of the three conditions (rereading the passage, answering questions set by the experimenter, and generating and answering their own questions). These judgments allowed us to determine whether participants could differentiate among the three tasks when predicting their performance on the final test and whether predictions followed the same pattern as performance. Second, as time is a limited resource, it was also important to determine how long each task would take. To measure the efficiency of each of the three tasks, we let participants spend as much time as they needed on each task. Any task that could potentially aid later test performance also takes time, and there is trade-off between these benefits and the time taken to achieve them. In the event that differences in performance are found among tasks, these must be qualified by their efficiency—that is, improvements in performance as a function of additional time taken.

Experiment 1

Method

Participants. Twenty-nine participants volunteered for the experiment and were reimbursed $10 for 1 hour of their time. Participants were recruited over the summer from the Psychology Department Subject Pool at Washington University in St Louis. Participants were thus current students, recent graduates, and members of the local community. The age range was 18 to 32, with a mean age of 21.4 years (SD = 2.8 years). Participants were predominantly female (23 women and 6 men). The sample was ethnically diverse, with 12 Caucasians, 19 Asian or Pacific Islanders, and 7 African Americans. Twenty-three participants were current undergraduates: 6 had completed 1 year of college, 7 had completed 2 years, and 10 had completed 3 years. Of the remaining participants, three held a bachelor's degree, two held a master's degree, and one had completed grade school. Three of the participants were not native English speakers, but had spoken English for 8, 9, or 16 years each.

Materials. Four passages were created by adapting Wikipedia pages on Salvador Dalí, the KGB, Venice, and the Taj Mahal (see Supplemental Materials for the passages reproduced in full; these materials were designed and previously used by Butler, Marsh, & Roediger, 2005, but are reproduced here for the first time). Passages were approximately 575 words long and were divided into four paragraphs. For the final test, two questions per paragraph were devised (see Supplemental Materials). All questions could be answered by a single word or short phrase.

In addition to the eight final test questions, eight comprehension questions were also devised per passage (see Supplemental Materials). The comprehension questions were used in the encoding task in the answer questions condition, whereas the final test questions were given to all participants at the end of the experiment to test their memory of the passages. To equate the quality and content of comprehension questions set by the experimenter with those generated by participants, we conducted a pilot study. A

group of 26 participants from the same pool as those in the current study were first shown the practice passage and comprehension questions written by the experimenter and asked to generate eight questions of a similar type for one of the three passages. They were instructed to generate two questions for each of the four paragraphs. From this bank of questions, we picked the two most frequently generated questions for each of the four paragraphs of each passage (eight questions per passage in total), with the constraint that four questions per passage probed information that was later tested, and the other four questions probed information that was not tested. As a result, the final test for each passage consisted of four questions that tested information from the comprehension questions, and four questions that tested new information. Note that test questions that overlapped with comprehension questions did not necessarily consist of the same wording, but tested the same material. For instance, one of the comprehension questions for the Dalí passage was: "What painting movement was Dalí a part of?" In the final test, the same information was probed with the following question: "Salvador Dalí created some of the most widely recognized images to come out of what artistic movement?"

Design. We used a within-subjects design with study condition (reread/answer questions/generate questions) as the only manipulated variable. The order of conditions and the assignment of passages to conditions were randomly determined by the program for each participant.

Procedure. Participants were tested individually with the experimenter present in the room during the session. Participants were told that they would study some passages for a later test. They were also told that for each passage, they would either be reading the text twice, or answering questions after initial reading, or generating questions after initial reading. Participants were not told how many passages to expect, or the order of conditions, to avoid anticipation of the third condition once two passages had been studied. Instructions were presented on the computer, and each passage was handed to participants printed on a single sheet of paper.

Participants initially took part in a practice phase so that they could familiarize themselves with the format of the passages, comprehension questions, and test. They were handed the practice passage and read it at their own pace. Once they were done reading the practice passage, the experimenter handed them a sheet of paper with eight comprehension questions, which participants answered while keeping the practice passage in front of them. Again, there were no time constraints on this task. Following completion of the questions, the passage and comprehension questions were removed and participants took the practice test, which consisted of eight short-answer questions (including four questions on material tested in the comprehension questions, and four questions on untested material). These questions appeared on the screen, and participants typed their responses. Following completion of the practice phase, participants were reminded that the comprehension questions they had answered should serve as an example of the types of questions they would be expected to generate later on in the experiment, and the main experiment began. Note that the task participants performed in the practice phase was equivalent to that performed in the answer questions condition in the main part of the experiment.

Upon completion of the practice phase, participants were handed their first passage to read at their own pace. Which passage they got was determined randomly by the program. At this stage, participants did not know which task they would be performing on the passage they were currently reading. After reading the passage, participants pressed a key to continue and were then instructed to do one of the three tasks: read the passage again at their own pace (reread condition); answer eight comprehension questions with the passage in front of them (answer questions condition); or generate eight comprehension questions and answers with the passage in front of them (generate questions condition). There were no time constraints on any of these tasks, and time taken was measured. For the answer questions condition, participants were handed a sheet with eight comprehension questions and blanks to fill in their responses. Questions were arranged in the order information appeared in the passage. For the generate questions condition, participants were handed a sheet of paper with eight long blanks for questions and eight shorter blanks for responses. Participants were instructed to generate two questions for each of the four paragraphs in the passage, and fill in the answers. Once participants completed the appropriate task for one passage, they were asked to estimate how much of the information from that passage they thought they would remember at the end of the experiment. Responses were given as a number from 0 ("*you don't think you're going to remember anything at all*") to 100 ("*you think you're going to remember the passage perfectly*"). The whole process was then repeated for the other two passages. Every participant read one passage twice, answered comprehension questions on another passage, and generated comprehension questions and answers on a third passage.

Following a 15-min retention interval during which participants played Tetris, they were tested on the material from each passage. A total of 24 questions were answered, eight from each of the three passages. Questions were blocked by passage, and the order of blocks and questions within the blocks was randomized.

Results and Discussion

Below we present the results as a function of condition for predictions (how much of the information participants thought they would remember on the test); performance (what proportion of test questions participants answered correctly); and time on task (both to read the passage initially, and then to complete the task appropriate to each condition). All three dependent measures were subjected to within-subjects analyses of variance (ANOVAs), with study condition (reread/answer questions/generate questions) as the within-subjects variable. All reported effects were significant at $p < .05$. Follow-up *t*-tests were only performed for significant main effects.

Predictions. Participants predicted how much information they would remember from each passage on a scale from 0 to 100 after completing the appropriate task (rereading the passage, answering questions, or generating questions). Predictions are presented in the left panel of Figure 1. Participants felt that they would do better on the test after having generated their own comprehension questions and answers ($M_{prediction} = 72.4$; $SD = 17.3$), than after having read a passage twice ($M_{prediction} = 63.0$; $SD = 21.2$) or answered comprehension questions set by the experimenter ($M_{prediction} = 63.4$; $SD = 21.8$). There was a significant difference

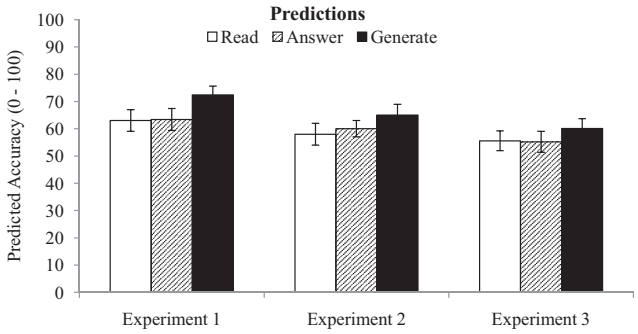

Figure 1. Mean predicted accuracy by study condition in Experiments 1–3. Error bars represent *SEM*. Predictions were made on a scale from 0 = "*you don't think you're going to remember anything at all*" to 100 = "*you think you're going to remember the passage perfectly.*"

in predictions among study conditions, $F(2, 56) = 5.37$, $\eta_p^2 = .16$. In particular, predictions for the generate condition were significantly higher than predictions made in the reread condition, $t(28) = 3.16$, $d = 0.59$ and predictions made in the answer questions condition, $t(28) = 2.41$, $d = 0.45$. Predictions made in the reread and answer questions conditions did not differ ($p = .91$).

Performance. Performance was measured in terms of the proportion of questions participants answered correctly on the final test (out of a total of eight per passage). Each answer was scored as either correct or incorrect (there were no half-points awarded) by a research assistant who was blind to the experimental conditions and hypotheses. Questions varied somewhat in the amount of information required. For questions that required only one word, a correct point was awarded when that word was included in the answer. Points were not deducted for incorrect spellings. For instance, a question with a one-word answer is "Between the 9th and 12th centuries, Venice flourished as the result of trade between Western Europe and what empire?" and the correct answer is "Byzantine." Participants would get a point for answers such as "Bysantine," "Byzantin," and "Byzantine Empire," but not for the answer "Roman." For questions that required a short sentence, a correct point was awarded when the answer was judged to contain at least two-thirds of the correct information. An example of such a question is "For what action did Salvador Dalí praise Francisco Franco?" and the correct response is "Signing death warrants for political prisoners." Participants would get a point for answers such as "Signing death orders for political prisoners," "Killing political dissidents," and "His death decrees to political prisoners," but not "Overthrowing the dictatorship" or "His repression policies" (these are examples of actual responses given by participants).

Performance did not follow the same pattern as predictions, and is presented in the left panel of Figure 2. Participants performed equally well after answering ($M_{proportion\ correct} = .72$; $SD = .21$) and generating ($M_{proportion\ correct} = .72$; $SD = .23$) questions, but worse after rereading the passage ($M_{proportion\ correct} = .57$; $SD = .23$). Study condition had a significant effect on performance $F(2, 56) = 10.04$, $\eta_p^2 = .26$, but contrary to participants' predictions, performance did not differ between the answer questions and generate questions conditions (in fact, performance was numerically identical). Instead, the effect was driven by a significant difference in performance between the reread condition and the answer questions condition, $t(28) = 5.15$, $d = 0.96$, as well as between the reread condition and the generate questions condition, $t(28) = 4.95$, $d = 0.75$. That is, there was no additional benefit of generating one's own questions over answering questions provided, although both tasks led to better performance than simply rereading the passage.

We also looked at performance on only those questions that were not probed during the study task—that is, questions that tested information that did not feature in the comprehension questions participants answered or generated. Because of the nature of the materials, this number was fixed at 4 questions per passage in the answer questions condition. In the generate questions condition, however, the number of test questions that referred to material that participants had not included in their generated questions ranged from 3 to 8 ($M_{untested} = 5.3$, $SD = 1.1$). Whereas performance on the untested questions was numerically higher when participants generated questions ($M_{performance} = .64$; $SD = .31$) than when participants answered questions ($M_{performance} = .53$; $SD = .36$), this difference did not reach significance ($p = .17$).

To check for order effects, we examined whether performance in the answer and generate questions conditions differed depending on which condition came first. Because of the random nature of condition order, 11 participants answered questions before generating questions, and 18 participants generated questions before answering them (although recall that all participants took part in a practice phase in which they answered example questions, so all had been exposed to questions written by the experimenter). A 2 × 2 mixed ANOVA on performance with study condition (answer/generate) as the within-subjects variable and order as the between-subjects variable produced no significant effects ($ps > .09$).

Time on task. Two sets of data were analyzed in relation to time on task. First, the time taken to read the initial passage was compared between conditions. No differences were expected here because participants in fact did not know which task was coming up while reading the passage. As predicted, no such differences were found ($p = .80$); the mean time spent initially reading each passage was 181 seconds across all conditions. More importantly, we also looked at the time spent on each of the three tasks

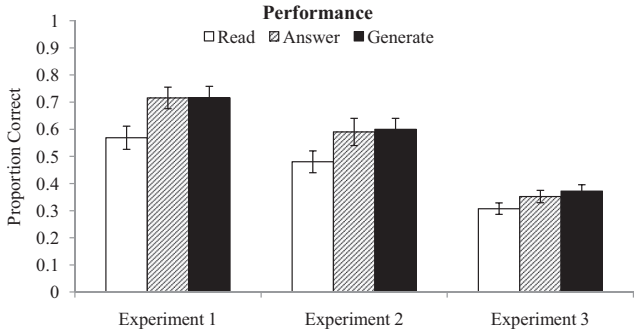

Figure 2. Mean accuracy by study condition in Experiments 1–3. Error bars represent *SEM*. Accuracy is calculated in terms of the number of questions answered correctly in Experiments 1 and 2, and in terms of the number of idea units correctly recalled in Experiment 3.

(rereading, answering questions, and generating questions) after the initial reading of the passage. The time spent on each of those three tasks is presented in the left panel of Figure 3. Time on task differed by study condition $F(2, 48) = 111.25$, $\eta_p^2 = .82$. Answering questions took on average 23 seconds longer than rereading the passage, although this difference was not significant, $t(24) = 1.74$, $p = .09$.[1] However, generating questions took more than three times longer than either rereading the passage, $t(24) = 11.45$, $d = 2.29$ ($M_{difference} = 279$ s, $SD = 120$), or answering questions, $t(24) = 11.16$, $d = 2.23$ ($M_{difference} = 256$ s, $SD = 115$). Thus, although the two question conditions produced comparable recall, the condition in which the questions were provided to students was much more efficient.

Experiment 2

In Experiment 1, we showed that answering questions, whether they were set by the experimenter or self-generated, improved performance relative to rereading. Generating and answering one's own questions did not produce any additional benefit to memory over and above answering questions set by the experimenter. Although some studies have found that the benefits of testing sometimes appear immediately (e.g., Carpenter, 2009), others have shown no effect of previous testing on an immediate test in contrast to large testing effects on a delayed test (e.g., Roediger & Karpicke, 2006b). In Experiment 2, we delayed the final test by two days instead of only by 15 minutes to determine whether any additional benefits of generating and answering one's own questions over and above answering someone else's questions emerge after a longer retention interval.

Method

Participants. Twenty-four participants volunteered for the experiment and were reimbursed $10 for 1 hour of their time. As in Experiment 1, participants were recruited over the summer from the Psychology Department Subject Pool at Washington University in St Louis. Participants were thus current students, recent graduates, and members of the local community. The age range was 18 to 28, with a mean age of 21.0 years ($SD = 2.3$ years). There were roughly equal numbers of males and females (11 women and 13 men). There were 12 Caucasians, 8 Asian or Pacific Islanders, 2 African Americans, 1 Hispanic, and 1 person who ticked "Other." Eighteen participants were current undergraduates: 5 had completed 1 year of college, 6 had completed 2 years, and 7 had completed 3 years. Two participants were current graduate students who had completed 1 and 2 years of graduate school, respectively. Of the remaining participants, two held a bachelor's degree and two held a master's degree. Four of the participants were not native English speakers, but had spoken English for 8, 13, 16, or 16 years each.

Design and procedure. The materials and procedure for this experiment were identical to those in Experiment 1, except that the study and test phases were separated by a retention interval of two days. Participants in this experiment also took a free recall test prior to a short answer test in the same format as that of Experiment 1. However, because of a program malfunction, these free recall data were lost.

Results and Discussion

The presentation of our results is identical to that of Experiment 1: predictions (how much of the information participants thought they would remember on the test); performance (what proportion of questions participants answered correctly on the test); and time on task (both to read the passage initially, and then to complete the task appropriate to each condition) are presented as a function of study condition. All three dependent measures were subjected to ANOVAs with study condition (reread/answer questions/generate questions) as the within-subjects variable.

Predictions. Participants predicted how much information they would remember from each passage on a scale from 0 to 100 after completing the appropriate task (rereading the passage, answering questions, or generating questions). Predictions are presented in the middle panel of Figure 1. As in Experiment 1, participants predicted that they would do better on the test after generating questions than after answering questions or rereading the passage. The overall ANOVA showed a significant effect of study condition on predictions $F(2, 46) = 3.36$, $p = .04$, $\eta_p^2 = .13$. However, only the comparison between the generate questions and reread conditions reached significance, $t(23) = 2.57$; $d = 0.52$ ($p = .13$ for the comparison between the generate questions and answer questions condition). Similarly to Koriat, Bjork, Sheffer, and Bar (2004), who showed that participants in between-subjects designs predict the same performance regardless of retention interval, participants in this experiment did not make lower predictions than those tested on the same day in Experiment 1: a cross-experiment comparison yielded a main effect of study condition, $F(2, 102) = 8.14$, $\eta_p^2 = .14$, but no effect of retention interval, on predictions ($p = .23$).

Performance. Performance was scored in the same manner as for Experiment 1, and replicated the pattern seen there, as shown in the middle panel of Figure 2. Participants performed equally

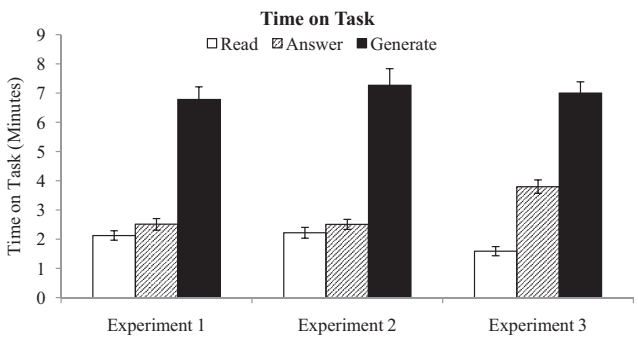

Figure 3. Mean time taken in minutes to complete each task by study condition in Experiments 1–3. Error bars represent *SEM*. Time shown does not include initial reading time, which did not vary by study condition.

[1] Because the recording of time on task was contingent upon participants pressing a key on the computer, occasionally participants failed to adhere to this instruction and the time taken to complete that particular task was not recorded. This resulted in loss of 6.9% of the data for initial reading time and 4.6% of the data for time on study task in Experiment 1; 5.6% of the data in Experiment 2; and 5.6% of the data in Experiment 3. Degrees of freedom in the analyses are adjusted for these data losses.

well after answering and generating questions, but worse after rereading the passage. Study condition had a significant effect on performance $F(2, 46) = 2.86$, $\eta_p^2 = .18$. As in Experiment 1, performance did not differ between the answer questions and generate questions conditions, but rereading the passage produced worse performance than both the answer questions condition, $t(23) = 2.51$, $d = 0.51$, and the generate questions condition, $t(23) = 3.63$, $d = 0.74$. As in Experiment 1, there was no additional benefit of generating questions over answering questions, although both tasks led to better performance than simply rereading the passage. Performance across the three conditions was 10% worse than in Experiment 1, as a result of the 2-day retention interval. A cross-experiment comparison yielded a main effect of study condition, $F(2, 102) = 15.00$, $\eta_p^2 = .22$, and a main effect of retention interval, $F(1, 51) = 4.97$, $\eta_p^2 = .09$, on performance, but no interaction between the two ($p = .85$).

Looking only at test questions that were not probed in the answered and generated questions (the number of which ranged from 2 to 7; $M = 4.7$, $SD = 1.2$), as in Experiment 1, performance was numerically higher in the generate questions condition ($M_{accuracy} = .45$; $SD = .28$) than in the answer questions condition ($M_{accuracy} = .42$; $SD = .27$), but this difference did not approach significance ($p = .66$). Eleven participants answered questions before generating questions, and 13 participants generated questions before answering. A 2 × 2 mixed ANOVA on performance with study condition (answer/generate) as the within-subjects variable and order as the between-subjects variable produced no significant effects ($ps > .38$).

Time on task. The mean time spent initially reading each passage was 182 seconds (no differences among condition, $p = .80$). The time spent on each of the three study tasks is presented in the middle panel of Figure 3. Time on task differed by study condition $F(2, 40) = 88.06$, $\eta_p^2 = .82$. Answering questions took on average 17 seconds longer than rereading the passage, and this difference was not significant, ($p = .17$). However, generating questions took more than three times longer than either rereading the passage $t(21) = 9.67$, $d = 2.11$ ($M_{difference} = 303$ s, $SD = 120$), or answering questions $t(22) = 10.70$, $d = 2.23$ ($M_{difference} = 286$ s; $SD = 71$).

Experiment 3

In Experiments 1 and 2 we showed that generating answers to questions—whether self- or other-generated—improved performance on a cued-recall test relative to rereading, both after a 15-min and 2-day retention interval. We also showed that generating questions did not lead to any benefits over and above answering questions set by someone else. However, performance on the final test may in part have been affected by the overlap between questions that participants interacted with in the task, and the questions on the final cued-recall test. More specifically, whereas the overlap between the questions set by the experimenter and the final test questions was controlled, this level of control was not possible in the generate questions condition. The questions set by the experimenter were taken from the same pool as those generated by participants (from a pilot study, see Experiment 1). However, the generated questions overlapped on average less with the final test questions than did the questions set by the experimenter, and were, by design, more variable. One way to get around this issue is to use a criterial test that does not involve cues that could overlap differentially with the encoding tasks. In Experiment 3 we gave participants a free recall test in which they were simply asked to recall as much information as possible from each passage. This design avoided issues of overlap between the questions in the task and the test questions.

Method

Participants. Thirty-three participants volunteered for the experiment and were reimbursed $10 for 1 hour of their time. As in Experiments 1 and 2, participants were recruited over the summer from the Psychology Department Subject Pool at Washington University in St Louis. Participants were thus current students, recent graduates, and members of the local community. The age range was 18 to 31, with a mean age of 21.0 years ($SD = 2.5$ years). There were more women than men in the sample (21 women and 12 men). There were 16 Caucasians, 12 Asian or Pacific Islanders, 3 African Americans, and 2 people who ticked "Other." Twenty-six participants were current undergraduates: 10 had completed 1 year of college, 4 had completed 2 years, 10 had completed 3 years, and 2 had completed 4 years. Three participants were current graduate students who had completed 1 and 2, and 7 years of graduate school respectively. Of the remaining participants, three held a bachelor's degree and one held a master's degree. Five of the participants were not native English speakers, but had spoken English for 6, 11, 12, 14, or 16 years each.

Materials. Four shorter passages were created for this experiment. Passages were created by adapting Wikipedia pages on the film director Pedro Almodovar (this was used as the practice passage), the singer Cesaria Evora, the archipelago Svalbard, and the TV show Top Gear (see Supplemental Materials). Passages were approximately 350 words long, and contained approximately 40 idea units each, as defined by two raters. An idea unit was identified as one self-contained fact, and there could be multiple idea units in each sentence. An example of a sentence containing two units is: "Top Gear is an award-winning BBC TV series about motor vehicles, primarily cars." The two idea units in this sentence are: "Top Gear is an award-winning BBC TV series" and "Top Gear is about motor vehicles, primarily cars." Passages were split into four paragraphs, and two comprehension questions per paragraph were devised by the experimenter (see Supplemental Materials).

Design and procedure. As in Experiments 1 and 2, we used a within-subjects design with study condition (reread/answer questions/generate questions) as the only manipulated variable. The assignment of passages to conditions was counterbalanced, whereas the order of conditions was randomly determined by the program for each participant.

The procedure was identical to that of Experiment 1, except that at test participants had 5 minutes per passage to recall as much information as they could without any cues. Responses were typed by participants. Instructions stated that the order in which the information was recalled did not matter, and that participants should try their best to recall as much content as they could. This free recall test was also given for the practice passage, so participants had experience of the type of test they would be getting.

Results and Discussion

Below we present the results as a function of condition for predictions (how much of the information participants thought they would remember on the test); performance (what proportion of idea units participants recalled on the free recall test); and time on task (both to read the passage initially, and then to complete the task appropriate to each condition). All three dependent measures were subjected to within-subjects ANOVAs with study condition (reread/answer questions/generate questions) as the within-subjects variable.

Predictions. Participants predicted how much information they would remember from each passage on a scale from 0 to 100 after completing the appropriate task (rereading the passage, answering questions, or generating questions). Predictions are presented in the right panel of Figure 1, and produced a pattern similar to those of Experiments 1 and 2. Predictions were numerically higher for the generate questions condition compared with the other two conditions, but the main effect of task was not significant ($p = .10$).

Performance. Scoring was done by two raters blind to the conditions. Participants received one point for each idea unit for which they correctly recalled 2/3 of the information. For instance, for the idea unit "She became an international star at the age of 47," "She became a star at the age of 47" would get one point. Scores were averaged across those given by the two raters. The two raters' scores were highly correlated, $r = .88$. Performance on the final test in terms of the proportion of idea units recalled in each condition is presented in the right panel of Figure 2. These scores are much lower than the short answer test results from previous experiments because the present experiment involved a free recall test. As in previous experiments, there was a significant difference in performance between study conditions, $F(2, 64) = 4.72$, $\eta_p^2 = .13$. Performance did not differ between the answer questions and generate questions conditions ($p = .36$), but rereading the passage resulted in fewer idea units being recalled than both generating and answering questions, $t(32) = 3.18$, $d = 0.87$, and answering questions set by the experimenter, $t(23) = 1.95$, $d = 0.33$, $p = .06$, although the latter difference did not reach significance. As in Experiments 1 and 2, there was no additional benefit of generating questions over answering questions, but both tasks lead to better performance than simply rereading the passage.

Looking only at idea units that were not probed in the answered and generated questions conditions, performance was identical across the two conditions ($M_{accuracy} = .39$; $SD = .15$). Fourteen participants answered questions before generating questions, and 19 participants generated questions before answering them. A 2 × 2 mixed ANOVA on performance with study condition (answer/generate) as the within-subjects variable and order as the between-subjects variable produced no significant effects ($ps > .24$).

Time on task. The mean time spent initially reading each passage was 141 seconds (no differences between condition, $p = .95$). The time spent on each of the three study tasks is presented in the right panel of Figure 3. Time on task differed by study condition $F(2, 50) = 140.35$, $\eta_p^2 = .85$; in particular, answering questions took double the amount of time that it took to reread the passage $t(26) = 8.23$, $d = 1.58$, and generating questions took double the amount of time that it took to answer questions $t(28) = 11.37$, $d = 2.11$.

General Discussion

The main aim of the present research was to test the effectiveness of an often-recommended study technique designed to aid retention of information presented in a passage. The technique is requiring students to create their own comprehension questions while reading the passage in front of them, and then answering these questions. This study technique was pitted against two alternative techniques: rereading the passage with no other task, and answering questions prepared by the experimenter. Rereading is a technique that is commonly employed by students preparing for a test (Karpicke et al., 2009), even though experimental findings demonstrate that additional readings of a text sometimes do not produce much improvement in performance (Callender & McDaniel, 2009a). Answering comprehension questions provided by the experimenter in preparation for a test has been shown to yield performance superior to that after rereading (Agarwal et al., 2008; Nungester & Duchastel, 1982). The advantage of the study technique we tested is that it does not require any additional material other than the passage itself. Thus, if this technique produces comparable performance to one in which students answer questions prepared by someone else, it could be useful for situations in which such materials are not available. Indeed, in three experiments, we found that generating and answering one's own questions in preparation for a memory test produced performance comparable to answering the experimenter's questions, and always represented a significant improvement in performance over rereading. However, we did not find generating one's own questions to be any more beneficial to retention than answering questions set by the experimenter. This pattern of data was found on an immediate cued-recall test (Experiment 1), a delayed cued-recall test (Experiment 2), and an immediate free recall test (Experiment 3).

It is important to consider the role of individual differences in the effectiveness of a task that puts the onus on the student to generate their own material for study. One issue that arises is that higher ability students may be more adept at selecting information that will later be tested. Some relevant preliminary work came to our attention after the experiments presented here were completed. Callender and McDaniel (2009b) had low- and high-ability readers highlight key information in a passage or generate questions about that information. One week later, participants returned and restudied the highlighted information or answered the questions they had generated, and then took a cued-recall test. In addition to performing quantitatively better on all tasks, high-ability readers were also qualitatively different from low-ability readers in that they selected more important information for highlighting or generating questions. Low ability readers appeared to benefit from the generation task only insofar as their generated questions overlapped with test questions. We did not design our study in a way that would permit an analysis of individual differences, but we were able to conduct rudimentary post hoc analyses to see whether overall performance mediated the differences between conditions we report. For each experiment we calculated mean differences between each pair of conditions (read-generate, read-answer, and answer-generate), and correlated these difference scores with overall performance. None of the correlations reached significance in any of the three experiments, suggesting that the pattern of results we report was not driven by a subset of the sample. In addition, in Experiment 3 we eliminated the issue of overlap between the questions answered in

the task and the test questions by giving participants a free recall test in which all information recalled was scored equally (i.e., we did not assign differential value to central and peripheral information). Thus, the choice of comprehension questions was less likely to affect performance on the criterial test, but generating questions still did not produce a significant advantage over answering questions. There are two caveats to this conclusion. First, all participants were given a practice phase and thus had an idea of the types of questions they were expected to generate. In the absence of such information, stronger individual differences and/or differences between conditions may have emerged. Second, we only examined the effectiveness of self-testing as a technique for memorizing information. In other domains such as those requiring original thought or creativity it may be that self-generated testing could lead to bigger improvements than experimenter-led testing because the ability to generate appropriate questions for self-testing is an inherent component of the task. We would, however, expect our findings to replicate in other situations involving factual information.

Two other issues were of import in the present article: that of performance predictions, and that of time on task. First, in order for the technique to be adapted by students, they need to be mindful of its advantages over the popular method of rereading, as they tend not to be when it comes to self-testing (e.g., Dunlosky & Nelson, 1992; Roediger & Karpicke, 2006b). We replicated Agarwal et al.'s (2008) finding that students do not predict any improvement on a later closed-book test from answering questions (i.e., taking an open-book self-test) as compared with rereading the text. However, participants did predict better performance in the condition in which they generated and then answered their own questions. This could have both positive and negative consequences. On one hand, this task appears to promote better metacognition than answering preset questions, in that participants recognize its superiority over rereading. So, whereas they do not seem to be aware of the benefits of answering questions generated by someone else, they do seem to recognize the advantage of generating their own questions.

A likely reason for the difference in predictions between the generate questions condition and the other two tasks is the large difference in time taken to complete the tasks, with much greater time used for question generation. There were no time limits set for the completion of the encoding tasks, so that the time taken on each task in the absence of control could be measured. Whereas rereading and answering comprehension questions took roughly the same amount of time, generating and answering one's own questions was a far more time-consuming endeavor. Generating and answering questions turned out to be a time-consuming task both because of the additional effort required to search out appropriate information to write questions about and because of the amount of additional writing involved as compared to answering preset cued recall questions. It is possible that this difference in time investment rather than an awareness of the benefits of the task drove students' predictions of later recall. The discrepancy in the amount of time taken to complete each task raises issues with regards to the efficiency of the tasks. Whereas the answer questions and generate questions conditions produced equivalent performance, the former was far more time-efficient. Clearly, if testing materials are available, students should opt to use them to maximize the efficiency of their time spent studying. However, in the absence of such materials, generating and answering questions is a viable alternative to rereading that appears to be just as beneficial to later retention even though much more time consuming.

References

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,* 861–876.

Ben-Chaim, D., & Zoller, U. (1997). Examination-type preferences of secondary school students and their teachers in the science disciplines. *Instructional Science, 25,* 347–367.

Butler, A. C., Marsh, E. J., & Roediger, III, H. L. (2005, May). *Distractor items on multiple-choice tests: Helpful or harmful?* Poster presented at the annual meeting of the Midwestern Psychological Society, Chicago.

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

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

Callender, A. A., & McDaniel, M. A. (2009b, November). *Self-generated Questions and the testing effect.* Poster presented at the annual meeting of the Psychonomic Society, Boston.

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,* 1563–1569.

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

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13,* 826–830.

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 632–642.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20,* 374–380.

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216.

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,* 649–667.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *The European Journal of Cognitive Psychology, 19,* 528–558.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17,* 471–479.

Karpicke, J. D., & Roediger, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 704–719.

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

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133,* 643–656.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory, 17,* 493–501.

Martin, M. A. (1985). Students' applications of self-questioning study techniques: An investigation of their efficacy. *Reading Psychology, 6,* 69–83.

McDaniel, M. A., & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychological Bulletin & Review, 15,* 237–255.

McDaniel, M. A., Howard, D., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20,* 516–522.

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

Nestojko, J. F., Bui, D. C., Kornell, N., & Bjork, E. L. (2009, November). *Preparing to teach improves the processing and retention of information.* Poster presented at the annual meeting of the Psychonomic Society, Boston.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74,* 18–22.

Robinson, F. P. (1970). *Effective Study* (4th ed.). New York: Harper & Row.

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

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

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory, 4,* 592–604.

Received February 25, 2010
Revision received June 10, 2010
Accepted July 2, 2010 ■

New APA Editors Appointed, 2012–2017

The Publications and Communications Board of the American Psychological Association announces the appointment of 9 new editors for 6-year terms beginning in 2012. As of January 1, 2011, manuscripts should be directed as follows:

- *Emotion* (http://www.apa.org/pubs/journals/emo), **David DeSteno, PhD,** Department of Psychology, Northeastern University, Boston, MA 02115
- *Experimental and Clinical Psychopharmacology* (http://www.apa.org/pubs/journals/pha), **Suzette M. Evans, PhD**, Columbia University and the New York State Psychiatric Institute, New York, NY 10032
- *Journal of Abnormal Psychology* (http://www.apa.org/pubs/journals/abn), **Sherryl H. Goodman, PhD**, Department of Psychology, Emory University, Atlanta, GA 30322
- *Journal of Comparative Psychology* (http://www.apa.org/pubs/journals/com), **Josep Call, PhD,** Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany
- *Journal of Counseling Psychology* (http://www.apa.org/pubs/journals/cou), **Terence J. G. Tracey, PhD,** Counseling and Counseling Psychology Programs, Arizona State University, Tempe, AZ 85823
- *Journal of Personality and Social Psychology: Attitudes and Social Cognition* (http://www.apa.org/pubs/journals/psp), **Eliot R. Smith, PhD,** Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405
- *Journal of Experimental Psychology: General (*http://www.apa.org/pubs/journals/xge), **Isabel Gauthier, PhD,** Department of Psychology, Vanderbilt University, Nashville, TN 37240
- *Journal of Experimental Psychology: Human Perception and Performance* (http://www.apa.org/pubs/journals/xhp), **James T. Enns, PhD**, Department of Psychology, University of British Columbia, Vancouver, BC V6T 1Z4
- *Rehabilitation Psychology* (http://www.apa.org/pubs/journals/rep), **Stephen T. Wegener, PhD, ABPP,** School of Medicine Department of Physical Medicine and Rehabilitation, Johns Hopkins University, Baltimore, MD 21287

Electronic manuscript submission: As of January 1, 2011, manuscripts should be submitted electronically to the new editors via the journal's Manuscript Submission Portal (see the website listed above with each journal title).

Manuscript submission patterns make the precise date of completion of the 2011 volumes uncertain. Current editors, Elizabeth A. Phelps, PhD, Nancy K. Mello, PhD, David Watson, PhD, Gordon M. Burghardt, PhD, Brent S. Mallinckrodt, PhD, Charles M. Judd, PhD, Fernanda Ferreira, PhD, Glyn W. Humphreys, PhD, and Timothy R. Elliott, PhD will receive and consider new manuscripts through December 31, 2010. Should 2011 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 2012 volumes.