Journal of Experimental Psychology:
Learning, Memory, and Cognition
2014, Vol. 40, No. 1, 106–114

© 2013 American Psychological Association
0278-7393/14/$12.00 DOI: 10.1037/a0033699

Attempting to Answer a Meaningful Question Enhances Subsequent Learning Even When Feedback Is Delayed

Nate Kornell
Williams College

Attempting to retrieve information from memory enhances subsequent learning even if the retrieval attempt is unsuccessful. Recent evidence suggests that this benefit materializes only if subsequent study occurs immediately after the retrieval attempt. Previous studies have prompted retrieval using a cue (e.g., *whale–???*) that has no intrinsic answer. Experiment 1 replicated prior word pair studies, but in Experiment 2, when participants learned meaningful trivia questions, testing enhanced learning even when subsequent study was delayed. Even in Experiment 3, when subsequent study was delayed by up to 24 hr, tests enhanced learning on a final test another 24 hr later. These findings may give comfort to educators who worry that asking a question or giving a test, on which students inevitably make mistakes, impairs learning if feedback is not immediate. They also suggest that there is a consensus in the literature thus far: Questions with rich semantic content enhance subsequent learning even when feedback is delayed, but less meaningful questions without an intrinsic answer enhance learning only when feedback is immediate.

Keywords: learning, memory, testing, retrieval, feedback

Retrieving information from memory enhances learning, a finding sometimes referred to as the testing or retrieval effect (Roediger & Butler, 2011; Roediger & Karpicke, 2006). Retrieval attempts are highly beneficial when they are successful (e.g., Karpicke & Roediger, 2008), but this article focuses on the effects of unsuccessful retrieval attempts.

A growing body of evidence suggests that attempting to retrieve information from memory enhances subsequent learning even when the retrieval attempt is unsuccessful (Grimaldi & Karpicke, 2012; Hays, Kornell, & Bjork, 2013; Huelser & Metcalfe, 2012; Izawa, 1970; Knight, Ball, Brewer, DeWitt, & Marsh, 2012; Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009; Vaughn & Rawson, 2012). Most of these studies included a study condition, in which participants were shown a cue and target to study (e.g., *pond–frog*), and a test condition, in which they were tested (e.g., *pond–???*) before being shown the target. Because participants were not allowed to study the pairs before the initial test and correct guesses were excluded from the analyses, these studies isolated the effect of unsuccessful retrieval attempts. Even so, testing enhanced learning in all of these studies more than studying the cue and target together did.

When students are tested, it is not always possible to provide them with immediate corrective feedback. The delay between taking an exam and getting feedback can be on the order of days or even weeks, for example. One goal in the studies presented here was to examine the effect of immediate versus delayed feedback following an unsuccessful attempt to answer a question. Three recent articles have converged on the same result: Unsuccessful attempts to answer a question are valuable only if the correct answer is presented immediately after the retrieval attempt. (As I explain below, Richland et al., 2009, is an important exception.) Hays et al. (2013) and Grimaldi and Karpicke (2012) used similar paradigms, which I replicated in Experiment 1: In the test–study condition, participants were tested and then they studied immediately (i.e., they were shown the cue and then the cue and target together).[1] In the test–delay–study condition, a delay of a few minutes separated the test and subsequent study. In the study-only condition, there was no initial test. The test–delay–study condition and the study-only condition produced roughly equivalent performance on the final test in all of these studies. In other words, tests had no effect on learning when subsequent studying was delayed. A similar study by Vaughn and Rawson (2012) used a variation on this procedure; instead of simply removing tests from the equation in the study-only condition, they replaced tests with additional study time. The result was that the test–delay–study condition produced less learning than did the study-only condition. (The most effective condition in all of these studies was test–study.)

Thus, the consensus from the three recent articles just reviewed (Grimaldi & Karpicke, 2012; Hays et al., 2013; Vaughn & Rawson, 2012) is that when people learn word pairs, guessing before studying is helpful only if the guess is followed by immediate feedback. There is at least one study that produced a different result. Richland et al. (2009) asked participants prequestions (e.g., "What is total colorblindness caused by brain damage called?") before they read a passage about colorblindness. Prequestions answered correctly were excluded from the data analysis. In the

This article was published Online First July 15, 2013.

Shana Carpenter provided valuable comments on drafts of this article. Patricia Jacobs Klein helped design the studies and collect the data. David Morales helped assemble the trivia questions.

Correspondence concerning this article should be addressed to Nate Kornell, Department of Psychology, Williams College, Williamstown, MA 01267. E-mail: nkornell@gmail.com

[1] Previous articles have labeled the conditions differently.

extended study (i.e., control) condition, time spent attempting to answer prequestions was replaced with time spent studying. Key concepts were highlighted in both passages to limit differential attention to pretested information. The prequestion condition produced more learning than did the extended study condition across five experiments. Because participants did not discover the answers to the prequestions until they read the passage, this study demonstrated that attempting to answer a question can enhance learning even when subsequent studying is delayed by a few minutes.

Richland et al.'s (2009) findings diverge from prior studies in at least two important ways. First, delayed feedback enhanced learning. Second, the materials were not arbitrary word pairs. In almost all other prior research, participants have learned related word pairs.[2] A single word (e.g., *whale*) does not have an intrinsic answer—the experimenters selected an answer (e.g., *mammal*) somewhat arbitrarily—and thus getting the answer correct before it is presented is a matter of guessing. In fact, unsuccessful retrieval may not involve retrieval in the typical sense of the word (e.g., Knight et al., 2012; Vaughn & Rawson, 2012). It is more like guessing.

Guessing becomes less arbitrary when questions have intrinsically correct answers. Instead of being required to guess, the learner can search her memory knowing that a correct answer might actually be stored there. In Experiments 2 and 3, I tested the hypothesis that asking authentic trivia questions would make unsuccessful retrieval beneficial even when subsequent study was delayed. I provide a full discussion of this hypothesis after the data have been presented (see General Discussion), but in brief there are two reasons to suspect it is correct. The first concerns confidence: In prior studies, it was obvious that the questions (e.g., *pond*) did not actually have answers. In Experiments 2 and 3, by contrast, participants were confident that a correct answer existed, and they were probably confident that some of their own incorrect answers were actually correct. The second reason concerns elaboration of a semantic network: Compared with single words, trivia questions activate a richer and more meaningful semantic network.

The studies presented in this article represent an attempt to investigate the effects of delayed feedback on learning of word pairs and more meaningful stimuli. The experiments investigated the effects of two variables: whether the learning materials were word pairs or semantically richer trivia questions, and whether the interval between testing and subsequent study was immediate or delayed.

In Experiment 1, participants were asked to learn word pairs (e.g., *pond–frog*). They studied the pairs (a) immediately after being tested, (b) following a test and a subsequent delay, or (c) without being tested at all. Experiment 2 replicated Experiment 1 but used trivia questions (e.g., "What is the world's tallest grass?"). In Experiment 3, participants were tested on trivia questions three times (following Vaughn & Rawson, 2012) or not at all; subsequent study was delayed by up to 24 hours, and the final retention test occurred 24 hours after studying ended.

The experiments presented in this article are potentially interesting for a number of reasons. First, if unsuccessful tests affect learning of word pairs differently than they affect learning of more meaningful materials, theories that account for the benefits of unsuccessful tests may require revision (e.g., Grimaldi & Karpicke, 2012; Hays et al., 2013). Second, in real life, most questions are more intrinsically meaningful than a word pair. Third, tests that are not followed by immediate feedback are important at a practical level: In classrooms, feedback on exam questions is almost always delayed (Knight et al., 2012), and in general it is commonplace to encounter a question without finding out the answer immediately.

Experiment 1

Participants in Experiment 1 studied and were tested on 30 word pairs. In the test–study condition, they were tested on the pair (e.g., *swim–???*) and then studied the pair immediately (e.g., *swim–float*). In the test–delay–study condition, they were tested on the pair, completed other trials, and then studied the pair. In the study-only condition, they were shown the pair without a prior test.

Method

Participants. Twenty-three participants (16 female, 7 male; median age = 40 years, range = 18–70 years) were paid $1.00 for completing the experiment. All participants were fluent English speakers living in the United States. Participants were recruited online via Amazon's Mechanical Turk, a site whose users have been shown to replicate laboratory findings (Buhrmester, Kwang, & Gosling, 2011; Germine et al., 2012; Mason & Suri, 2012; Sprouse, 2011).

Materials. Fifty-six word pairs (e.g., *star–night, mouse–hole*) were taken from the word pairs used by Kornell et al. (2009). These pairs had a forward association strength of .050 to .054, meaning that when presented with the cue, people produce the target as their first response about 5% of the time (Nelson, McEvoy, & Schreiber, 1998). Although there were 56 pairs, 35 pairs were selected randomly for each participant.

Design and procedure. One independent variable was manipulated within participants. It had three levels: test–study, test–delay–study, and study-only. The experiment took place online. After reading instructions, participants completed three phases: study, distractor, and test.

There were two kinds of trials during the study phase. During test trials, participants were shown a cue (e.g., *frog*) for 8 s and asked to enter the target. During study trials, they were shown the cue and target together (e.g., *frog–pond*) for 5 s. Word pairs were assigned to conditions randomly for each participant.

Unbeknownst to the participants, the study phase was broken into two blocks of trials. During the first block, participants completed 15 trials. They did test trials on the 10 word pairs assigned to the test–delay–study condition. Because the second block was a mixture of presentations and tests and it seemed desirable to give the two blocks similar structures, in the first block five additional pairs were presented using study trials, but these five filler items were not included in the data analysis. During the second block, there were three classes of items. The 10 pairs that had been tested during the first block were presented for delayed study. An average of 4 minutes elapsed between the initial test and feedback for these items. Another 10 pairs were tested for the first time and then

[2] Some studies have also used unrelated word pairs, which do not benefit from guessing (Grimaldi & Karpicke, 2012; Huelser & Metcalfe, 2012). Kornell et al. (2009) also used fabricated trivia questions.

presented for study immediately after they were tested. A third set of 10 pairs was presented for study without having been tested previously. Because these three classes of items were mixed randomly, the average lag between the last presentation of a given item and the final test was the same in the three conditions.

After the study phase, participants were directed to a website where they were supposed to play the video game Tetris for 3 minutes. Unfortunately, that site was not functioning on the day Experiment 1 was conducted, so there was a 3-min retention interval during which the participants were shown a webpage containing an error message.

During the final test, the cues were presented, one by one in random order, and the participants were asked to type in the target they had learned previously. They were given unlimited time to do so.

Results

In this study and those that follow, responses that scored 75 or greater using the similar_text function in the programming language PHP, including answers that contained minor spelling errors, were considered correct. Participants guessed the correct answer on 4.0% of trials during the study phase. These items were excluded from further analysis, which created a small selection bias that favored nontested items.

As Table 1 shows, cued recall on the final test was highest in the test–study condition, followed by the study-only condition, and was lowest in the test–delay–study condition. There was a main effect of condition on final test recall accuracy, $F(2, 44) = 6.01$, $p = .005$, $\eta_p^2 = .21$. Planned comparisons showed that the test–study condition outperformed the study-only condition, $t(22) = 2.58$, $p = .017$, $d = .37$, but the test–delay–study condition did not differ significantly from the study-only condition, $t(22) = -0.94$, $p = .356$, $d = .16$.

Discussion

Guessing prior to studying enhanced learning, compared to studying the answer without guessing. However, guessing was valuable only when followed by immediate studying. When the study opportunity was delayed, final test performance looked as though the test had never happened. These results replicate the findings of Grimaldi and Karpicke (2012) and Hays et al. (2013).

Experiment 2

Although the word pairs in Experiment 1 were related (e.g., *antler–horn*), they were somewhat artificial in two important ways. First, as the participants surely realized, there is no intrinsically correct answer to a cue like *antler.* Second, a single word is not rich in meaningfulness or semantic complexity. A trivia question, which is essentially a fact with one element removed to turn it into a question, is more meaningful. (Of course, educational materials such as an essay, lecture, or book make simple facts seem relatively impoverished.) Experiment 2 was virtually identical to Experiment 1 except that the materials were trivia questions, not word pairs (see the Appendix). Unlike the trivia questions used by Kornell et al. (2009), the questions were not fictional. They were selected such that the answers were memorable but largely unknown to participants at the outset of the study.

Table 1
Mean (and Standard Deviation) Percentage of Questions Answered Correctly on the Final Test in Experiments 1 and 2

Condition	Experiment 1	Experiment 2
Study-only	63 (27)	78 (20)
Test–delay–study	59 (24)	85 (16)
Test–study	73 (28)	89 (19)

Method

Participants. Thirty-one participants (20 female, 11 male; median age = 26 years, range = 19–50 years), who were recruited via Amazon's Mechanical Turk, were paid $1.00 for completing the experiment. All participants were fluent English speakers living in the United States.

Materials. Thirty-five trivia questions were assembled for the experiment (see the Appendix, which includes the 35 questions from Experiment 2 as well as 13 questions used in Experiment 3). The ultimate selection of questions depended on the judgment of the experimenter, but three criteria were used to select questions. First, they were real questions and their answers were correct. Second, most people do not know the answers. Third, the answers seemed to be learnable. We avoided using a person's name as an answer unless the person was widely known (e.g., Lincoln).

Design and procedure. With the exception of the materials, there were two differences between Experiments 1 and 2. First, because it takes more time to read a trivia question than it does to read a single word, the duration of study and test trials were increased from 8 and 5 seconds in Experiment 1 to 12 and 8 seconds in Experiment 2. As a result, the delay between the initial test and study, in the test–delay–study condition, increased to 6 minutes. Second, instead of being directed to a broken website during the distractor task, participants played the video game Asteroids for 3 minutes.

Results

During the study phase, when participants attempted to answer the questions without having studied them first, they gave correct responses on 2.6% of the trials. These items were excluded from further analysis.

As Table 1 shows, cued-recall on the final test was highest in the test–study condition, followed by the test–delay–study condition, and was lowest in the study-only condition. There was a main effect of condition on final test recall accuracy, $F(2, 60) = 6.52$, $p = .003$, $\eta_p^2 = .18$. Planned comparisons showed that the test–study condition outperformed the study-only condition, $t(30) = 4.15$, $p = .0003$, $d = .56$. Unlike in Experiment 1, the test–delay–study condition also outperformed the study-only condition. Although the effect was weak, it was significant, $t(30) = 2.06$, $p = .049$, $d = .36$.

Discussion

Like in Experiment 1, the benefit of trying to guess the answer to a trivia question was significant when the subsequent studying happened immediately. Unlike in Experiment 1, the benefit was also significant when the subsequent studying was delayed. These results diverge from prior studies that used word pairs as stimuli (Grimaldi & Karpicke, 2012; Hays et al., 2013; Vaughn & Rawson, 2012). They are more similar to the benefits of pretesting Richland et al. (2009) obtained using delayed feedback in a deeply semantic task. Further discussion of these findings is reserved for the General Discussion.

Experiment 3a

Experiment 2 suggested that guessing followed by delayed feedback can be beneficial when people learn facts even if it is not beneficial when they learn paired associates. The purpose in Experiment 3a was to test this proposition using longer and more realistic retention intervals. The experimental design included two retention intervals: the interval between initial guessing and the opportunity to study the answer and the interval between studying the answer and taking the final test. Both intervals were increased to 24 hours (see Figure 1). In prior studies (including Experiment 1), lengthening the first interval to even a few minutes had a lethal effect on learning. Increasing this interval to 24 hours allowed a test of the hypothesis that feedback can be truly beneficial even after a long delay. It also seemed realistic to increase this interval because it is not uncommon to ponder a question (sometimes in a classroom or on a test) and then find out the answer after a significant delay. The second interval was increased so that the study would measure relatively long-term learning. Another change intended to increase realism was that participants were allowed to decide how long they spent attempting to guess the answer, how long they spent studying, and how long they spent on final test questions.

Method

Participants. Seventy-six participants (51 female, 25 male; median age = 30 years, range = 18–67 years) were paid $3.00 ($1.00 after each session) for completing the experiment. All participants were fluent English speakers living in the United States. Participants were recruited online via Amazon's Mechanical Turk. Ninety-two participants completed Session 1, 83 of these completed Session 2, and 76 of these completed Session 3.

Materials. Forty-eight trivia questions, including the 35 questions from Experiment 2, were used in Experiment 3 (see the Appendix). The criteria for selecting questions were the same as in Experiment 2.

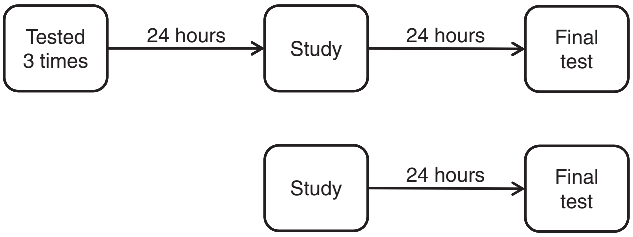

Figure 1. Procedure in Experiment 3a in the test–delay–restudy condition (top) and study-only condition (bottom). Experiment 3b was identical except that in the test–delay–restudy condition, the first 24-hr delay was eliminated.

Design and procedure. One independent variable with two levels was manipulated within participants. Half of the items were tested in Session 1; the other half did not appear in Session 1. All items were studied in Session 2 and tested in Session 3. Items were assigned to conditions randomly for each participant.

During Session 1, 24 trivia questions were presented one at a time. Participants were asked to type in the answer and press *return.* Trial timing was under the participant's control; when *return* was pressed, the next question appeared. Participants were asked each question three times with other questions in between (see Vaughn & Rawson, 2012). Each question was asked once in each of three question blocks, and the order of questions was randomized anew for each participant during each of the three question blocks.

During Session 2, participants were shown 48 questions accompanied by their answers. Half of the questions had been asked during Session 1 and half had not. The questions were presented one at a time; participants could study each question as long as they wanted to and then press *return* to move on to the next question/answer pair. The order of the questions was randomized anew for each participant.

During Session 3, participants were tested on the 48 questions they had studied during Session 2. Again, the timing was under their control; when they pressed *return,* they were shown the correct answer for 1 s and then the next question appeared on the screen. The order of the questions was randomized anew for each participant.

Results

Questions that were answered correctly once or more during Session 1 were excluded from further analysis. These questions accounted for 14.2% of all Session 1 questions.[3] Again, excluding them creates an item selection bias in favor of nontested items.

As Table 2 shows, the percentage of questions answered correctly on the final test was greater in the test–delay–study condition than the study-only condition, $t(75) = 8.75$, $p < .0001$, $d = .86$. The benefit of testing occurred even though, during Session 2, participants spent significantly more seconds studying in the study-only condition ($M = 5.13$, $SD = 2.87$) than in the test–delay–study condition ($M = 4.36$, $SD = 2.49$), $t(75) = 6.19$, $p < .0001$, $d = .29$.

One might expect that giving an incorrect answer, without being corrected for 24 hours, would have a detrimental effect on learning. To test this possibility, I divided each participant's questions into those he or she answered at least once during Session 1 (i.e., commission errors) and those he or she never answered during Session 1 (i.e., omission errors). As in the other analyses, questions

[3] In Experiment 2, participants answered these questions correctly on only 2.4% of trials, whereas the percentages were 14.2% and 14.0% in Experiments 3a and 3b, respectively. Three factors probably influenced this difference. First, Experiment 3 was conducted before Experiment 2, and items that were frequently answered correctly in Experiment 3 were removed from the question pool for Experiment 2. Second, participants were allowed to respond three times during Experiment 3, rather than once. Third, response time was limited to 12 s Experiment 2, but it was unlimited in Experiment 3.

Table 2
Mean (and Standard Deviation) Percentage of Questions Answered Correctly on the Final Test in Experiments 3a and 3b

Condition	Experiment 3a	Experiment 3b
Study-only	62 (20)	49 (27)
Test–delay–study	78 (18)	62 (27)

answered correctly one or more times were excluded. Among the 52 participants who had at least one commission error and one omission error, 59% of questions were categorized as commission errors during Session 1 (across all participants, 70% of questions were answered). Final test accuracy following an omission error (M = 74.4%, SD = 25.9) and a commission error (M = 76.4%, SD = 19.6) did not differ, $t(51) = 0.52$, $p = .603$, $d = .07$. Thus, there was no indication that final test accuracy was worse following Session 1 commission errors than following Session 1 omission errors. This finding corresponds to Kang et al.'s (2011) finding that forcing people to guess the answer to a trivia question, versus letting them decide not to answer, did not affect learning.

Experiment 3b

Fifty-two participants (30 female, 20 male, 1 other, 1 did not report gender; median age = 30 years, range = 18–57) completed a conceptual replication of Experiment 3a. Sixty-eight participants completed Session 1, but of these 16 did not complete Session 2. The method was identical except that it took two sessions instead of three. The study phase began immediately after the initial test ended, instead of 24 hours later. The retention interval between the study phase and final test remained 24 hours. Experiment 3b's raison d'être was to serve as a basis for comparison with Experiment 3a.

Results

The results agree with those of Experiment 3a. Correct answers during the study phase, which accounted for 14.0% of items, were excluded from further analysis. Accuracy on the final test was higher in the test–delay–study condition than in the study-only condition ($p < .0001$, $d = .47$). Participants spent less time studying if they had been tested previously ($M = 2.97$, $SD = 1.95$) than if they had not ($M = 3.45$, $SD = 2.00$), $t(51) = 5.78$, $p < .0001$, $d = .24$. Final test accuracy did not differ following commissions errors (M = 63.8%, SD = 30.3) versus omission errors (M = 66.1%, SD = 31.2), $t(33) = 0.51$, $d = .07$.

One might expect, based on Experiments 1 and 2 and prior studies, that the value of initial testing would be greater when the study phase followed testing immediately rather than after a delay. This hypothesis was not supported. An analysis of variance that compared Experiments 3a and 3b showed that the interaction between testing and delay was not significant, $F(1, 126) = 1.57$, $p = .212$, $\eta_p^2 = .01$. If anything, the results were opposite the prediction (see Table 2): The advantage of prior testing was slightly larger when subsequent study was delayed by 24 hours (a 16 percentage point difference in Experiment 3a) than when it was not (a 13 percentage point difference in Experiment 3b). Table 2 shows that performance on the final test was lower overall in Experiment 3b than Experiment 3a, $F(1, 126) = 13.66$, $p = .0003$, $\eta_p^2 = .10$, but the reason for this difference may be that participants in Experiment 3b spent less time studying overall (possibly due to fatigue at the end of the long first session), $F(1, 126) = 12.92$, $p = .0005$, $\eta_p^2 = .09$.

Discussion of Experiments 3a and 3b

Being asked a question enhanced subsequent learning of the answer even when the answer was not presented until up to 24 hours later. This finding cannot be explained by differential time spent during the study phase, because participants spent more time studying previously untested items than tested items. Contrary to the hypothesis that making an error and having it stand uncorrected would impair learning, final recall did not differ depending on whether participants made errors of commission or omission during their initial test (cf. Kang et al., 2011). Furthermore, the initial test was no less valuable when subsequent study was delayed by 24 hours (in Experiment 3a) than when it was delayed by a few minutes (in Experiment 3b).

Participants who completed Session 1 but did not complete the subsequent session(s) accounted for 17% and 24% of the participants in Experiment 3a and Experiment 3b, respectively. If participants who did not complete the study were demonstrably different from participants who did complete the study, this might limit the generalizability of the findings. Therefore, I analyzed Session 1 test performance for participants who finished all sessions compared to those who did not. The percentage of questions answered correctly at least once during the initial test, in Experiment 3a, was 14.2% for participants who completed all three sessions and 14.7% for participants who did not. In Experiment 3b the respective percentages were 14% and 12.3%. The difference was not significant in Experiment 3a, $t(91) = 0.15$, $p = .884$, $d = .03$, or in Experiment 3b, $t(66) = -0.47$, $p = .638$, $d = .12$. Thus, there is no obvious reason to suspect that attrition had an undue impact on the findings of Experiment 3.

General Discussion

Three experiments suggested that being asked a question and failing to give a correct answer can have a positive impact on learning even if one does not find out the correct answer until after a substantial delay. The findings point to features of the learning materials as crucial in determining the effect of delayed feedback following a failure to answer correctly.

In Experiment 1, unsuccessfully attempting to answer a question had no measurable effect on learning unless the attempt was followed by an immediate presentation of the correct answer. As in prior studies showing the same effect (Grimaldi & Karpicke, 2012; Hays et al., 2013; Vaughn & Rawson, 2012), the learning materials were related word pairs. In Experiment 2, however, when word pairs were replaced with trivia questions, attempting to answer enhanced learning even when the subsequent study opportunity was delayed. In Experiment 3, participants attempted to answer three times, waited either a few minutes or 24 hours before being told the correct answer, and then took a final test another 24 hours after studying.

In Experiments 1 and 2 and in prior research (Grimaldi & Karpicke, 2012; Hays et al., 2013; Vaughn & Rawson, 2012), increasing the test–study interval (i.e., the interval between attempting to answer and being told the correct answer) diminished or eliminated the benefit of the initial test. Based on these findings, which showed that a test–study interval of a few minutes impaired learning, one would predict that a test–study interval that was orders of magnitude longer would have a serious detrimental effect on learning. Thus, it is ironic that the largest benefit of initial testing occurred in Experiment 3a, in which the test–study interval was 24 hours. This benefit was slightly, though not significantly, larger than the benefit when the interval was only a few minutes in Experiment 3b. Next I turn to why testing affected word pairs and trivia questions differently.

Elaborative Retrieval

Being asked to think of a response when presented with a cue activates the semantic network surrounding the cue. In Experiment 1, a cue like *pond* presumably activated concepts like "lake," "water," "lily," "duck," and "frog." The trivia questions in Experiments 2 and 3, by contrast, probably activated larger and more meaningfully interconnected networks. For example, a question like "Who was *Time Magazine*'s 'Man of the Year' in 1938?" presumably activated concepts related to contemporary political figures such as Roosevelt, Churchill, and the correct answer, Hitler (all answers given by at least one participant), as well as how these leaders related to the larger historical context, including the buildup to World War II and the Great Depression in the United States. The question also prompted responses that were less political, such as Bogart, Astaire, Einstein, and Superman (not to mention more dubious answers such as President Johnson, Hugh Hefner, and Babe Ruth). This web of activation is richer than the network activated by *pond*.

A rich web of activation enhances elaborative retrieval, at least in part, by increasing the effectiveness and number of mediators: Concepts that are triggered by the cue and themselves trigger the target (e.g., Carpenter, 2011; Pyc & Rawson, 2010). Mediators can be incorrect answers that the participant gave as a response, or they can be concepts that were activated (perhaps unconsciously) by the cue—in either case they could help elicit the target (e.g., being asked about *Time*'s Man of the Year could elicit Churchill, which could, in turn, elicit Hitler). Mediators for a pair like *pond–frog* might be more difficult to generate and less memorable once generated.

According to elaborative retrieval explanations of testing effects, retrieval is beneficial, at least in part, because searching memory for an answer leads to sustained activation of the semantic network surrounding the cue (Carpenter, 2009; Carpenter & DeLosh, 2006; Grimaldi & Karpicke, 2012; Kornell et al., 2009). Retrieval is thought to enhance learning because this activation is more sustained and far-reaching than the activation created when one is shown a question and answer at the same time that. Trivia questions may have led to more robust testing effects because elaborative retrieval worked on a relatively large, rich network.

When a person is asked a question (or shown a single cue word) and cannot answer it correctly, the semantic network associated with the question is primed and this priming enhances learning during immediate feedback. Some authors have argued that delayed feedback does not benefit from a prior test, because priming fades over time and is no longer active when delayed feedback occurs (Grimaldi & Karpicke, 2012; Hays et al., 2013). This explanation fits with Experiments 1 and 2, where delay seemed to diminish learning, but in Experiment 3 there was no apparent decrement as a function of delay. Two factors may have conspired to prevent delay effects in Experiment 3. First, a question like "What was the name of the dog from *The Grinch Who Stole Christmas*" may be memorable enough that even after a delay, its level of activation did not return to baseline. Instead, participants may have formed a long-term memory of the question that remained available after 24 hours. They may also have formed a long-term memory of incorrect responses (e.g., Rex) that could then serve as mediators. Second, the timing of delayed feedback in Experiments 1 and 2 was roughly the same as the timing of immediate feedback in Experiment 3 (all were a few minutes). If short-term priming had largely faded after a few minutes, delays of a few minutes and 24 hours would not differ with respect to short-term priming.

In summary, the foregoing explanation is speculative, but it suggests that short-term priming did not play a role in the benefit of testing in Experiment 3 (although it may have affected Experiments 1 and 2). Instead, participants may have encoded the questions they were being asked in long-term memory, which made the semantic network surrounding the questions more active when the question was presented again after a delay. This network being more active would have conferred the benefits of elaborative retrieval. Participants may have encoded incorrect answers as well, which, by serving as mediators, would have allowed participants to encode the correct answers to the questions more effectively once the answers were presented. If true, this explanation also suggests that the mechanisms underlying short-term and long-term feedback may be different: Both may rely on elaborative retrieval, but with one based on short-term priming and the other based on long-term memory of a question and one or more incorrect answers.

Thus, some combination of enhanced elaborative retrieval, enhanced mediator effectiveness, and enhanced cue memorability may explain why the effect of retrieval was larger with trivia questions than it was with word pairs. In either case, the findings fit well with the previous literature. In prior studies that used word pairs, guessing did not enhance learning when feedback was delayed. However, Richland et al. (2009) found the opposite: Pretesting participants on a set of questions—even if they did not answer correctly—enhanced subsequent learning. Perhaps not coincidentally, Richland et al.'s questions (e.g., "What is total color blindness caused by brain damage called?") resembled the questions used in Experiments 2 and 3. Thus, the current findings make sense of a seeming inconsistency in the prior literature and suggest that there is a consensus, at least thus far: Questions with rich semantic content enhance subsequent learning even when feedback is delayed, but questions that consist of a single cue word (e.g., *whale–???*) do not.

Error Correction and Confidence

There is a long-standing belief in the danger of making errors, dating back to the idea of errorless learning in conditioning (e.g., Terrace, 1963). Errorless learning is beneficial for memory rehabilitation in some populations (e.g., Clare & Jones, 2008), but its effects are more elusive in normal populations (e.g., Metcalfe &

Kornell, 2007). If making errors impaired learning in the present studies, one might expect commission errors to be particularly harmful—especially because participants often made the same commission error three times during the study phase, which probably strengthened their "knowledge" of this incorrect answer. Yet, there was no measurable decrement in performance following commission errors as compared to omission errors. Based on this finding, it appears that, at least in normal populations, learning a wrong answer does not necessarily make it more difficult to learn the correct answer. Previous research on the hypercorrection effect (e.g., Butterfield & Metcalfe, 2001, 2006; Fazio & Marsh, 2010) has shown that when people are more certain of an incorrect response, they find it easier, not harder, to correct their error once the correct answer is identified. Thus, in Experiment 3, it is possible that making repeated commission errors did, in fact, strengthen confidence in incorrect responses, but that doing so led to hypercorrection and thus enhanced subsequent learning of the correct answer.

The hypercorrection effect may provide another possible explanation of the difference between word pairs and trivia questions in the present studies. It is likely that when participants were presented with a single word they had little confidence in their responses, because guessing does not tend to instill a lot of confidence. Because trivia questions have actual answers and because participants made many plausible responses during the initial test (including the 14% of their responses that were correct in Experiment 3), it is likely that participants were relatively more confident in their answers to trivia questions than to single words by the time they were presented with the actual answer. This confidence may have had a positive impact on learning via the hypercorrection effect: High-confidence errors are more likely to be corrected. A limitation of this explanation is that it does not, by itself, explain the positive effect of omission errors on subsequent learning.[4]

Practical Implications

There are times when asking a question is possible but providing an answer immediately afterward is not. The classic example is when someone takes a test. Many educators worry that when a student makes an error on a test, and it is not corrected immediately, his learning suffers (Pashler et al., 2007). The present results suggest the opposite: Taking a test and getting an answer incorrect enhances subsequent learning, even if the learner is not told the correct answer until a substantial amount of time later. (The findings reported here do not speak to situations in which students are never told the correct answer.) Classroom tests are not the only situation in which questions are posed long before the answer becomes available. Sometimes the answer is not yet known (e.g., when a news report poses a question such as "who will win the election?" or when a scientist sets out to answer a novel research question), and sometimes giving the answer away would spoil the narrative (e.g., in a mystery novel).[5] In all of these situations, asking the question probably has a positive effect on subsequent learning. Moreover, it appears that delayed feedback was just as effective as immediate feedback, based on Experiment 3 (see Butler & Roediger, 2008; Kulik & Kulik, 1988; Metcalfe & Kornell, 2007).

The present studies suggest that attempted retrieval enhanced subsequent learning, but it remains possible that exposure to the question would have enhanced learning even without a retrieval attempt. Perhaps future research can address this possibility by comparing a retrieval condition to a condition in which participants are shown the questions but are not asked to answer them. Experiment 5 by Richland et al. (2009) attempted such a comparison and showed that participants benefited when they were asked to memorize a set of questions instead of answer them. It is possible that their participants attempted to answer at least some questions, even if only covertly.

The fact that the testing was manipulated within participants is another possible limitation of the present studies. Performance on untested items might have suffered because they were studied in the presence of previously tested items. The fact that participants spent more time studying untested items than tested items might mitigate this concern, but it remains possible that performance on untested items would have been higher if tested items had not been mixed in.

The present studies emphatically do not suggest that taking a test without feedback is more effective than studying would have been. To make this distinction clear, I conducted an additional study using the same procedure as Experiment 3b, with one change: In the test–delay–restudy condition, participants were tested three times and then studied once, as in Experiment 3b, but in the study-only condition they studied each item four times (rather than once, as in the study-only condition in Experiment 3). Final test performance was higher in the study-only condition (M = 77.8%, SD = 19.3) than the test–delay–study condition (M = 61.0%, SD = 22.0), $t(65) = 8.19$, $p < .0001$, $d = .81$. These findings fit with Vaughn and Rawson's (2012) finding that guessing three times and then studying was less effective than studying four times. However, these results diverge from the Richland et al.'s (2009) finding that pretesting was more effective than spending additional time studying.

From a practical perspective, the results suggest that asking someone a meaningful question that he or she cannot answer enhances subsequent learning, even if the correct answer is not provided until after a substantial delay. These results give comfort to educators who face situations that require tests without immediate feedback. However, they do not suggest that a test with delayed feedback is more effective than two opportunities to study or a test with immediate feedback.

[4] Previous studies have shown that learning is largely unaffected by forcing people to guess versus letting them leave answers blank (Butler & Roediger, 2008; Kang et al., 2011; Metcalfe & Kornell, 2007). This result is consistent with the finding that there was no difference between commission errors and omission errors in Experiment 3. It is also interesting to note, however, that people are probably no more confident in guesses they did not want to produce than they are in not answering at all. In other words, forced guesses may be offered with so little confidence that they suffer from a hypocorrection effect.

[5] The findings from the experiments presented in this paper do not speak to a related situation in which a question is posed and then the learner is told an incorrect answer, which is corrected later (e.g., Grimaldi & Karpicke, 2012).

References

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6,* 3–5. doi:10.1177/1745691610393980

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36,* 604–616. doi:10.3758/MC.36.3.604

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1491–1494. doi:10.1037/0278-7393.27.6.1491

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1,* 69–84. doi:10.1007/s11409-006-6894-z

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. doi:10.1037/a0017021

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,* 1547–1552. doi:10.1037/a0024140

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. doi:10.3758/BF03193405

Clare, L., & Jones, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology Review, 18,* 1–23. doi:10.1007/s11065-008-9051-4

Fazio, L. K., & Marsh, E. J. (2010). Correcting false memories. *Psychological Science, 21,* 801–803. doi:10.1177/0956797610371341

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19,* 847–857. doi:10.3758/s13423-012-0296-9

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory & Cognition, 40,* 505–513. doi:10.3758/s13421-011-0174-0

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 290–296. doi:10.1037/a0028468

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition, 40,* 514–527. doi:10.3758/s13421-011-0167-z

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344. doi:10.1037/h0028541

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology, 103,* 48–59. doi:10.1037/a0021977

Karpicke, J. D., & Roediger, H. L. (2008, February 15). The critical importance of retrieval for learning. *Science, 319,* 966–968. doi:10.1126/science.1152408

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language, 66,* 731–746. doi:10.1016/j.jml.2011.12.008

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 989–998. doi:10.1037/a0015729

Kulik, J. A., & Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58,* 79–97. doi:10.3102/00346543058001079

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44,* 1–23. doi:10.3758/s13428-011-0124-6

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review, 14,* 225–229. doi:10.3758/BF03194056

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. Available from http://w3.usf.edu/FreeAssociation/

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M. A., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007–2004). Washington, DC: National Center for Education Research.

Pyc, M. A., & Rawson, K. A. (2010, October 15). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330,* 335. doi:10.1126/science.1191465

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15,* 243–257. doi:10.1037/a0016496

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. doi:10.1016/j.tics.2010.09.003

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

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods, 43,* 155–167. doi:10.3758/s13428-010-0039-7

Terrace, H. S. (1963). Errorless transfers of a discrimination across two continua. *Journal of the Experimental Analysis of Behavior, 6,* 223–232. doi:10.1901/jeab.1963.6-223

Vaughn, K. E., & Rawson, K. A. (2012). When is guessing incorrectly better than studying for enhancing memory? *Psychonomic Bulletin & Review, 19,* 899–905. doi:10.3758/s13423-012-0276-0

(*Appendix follows*)

Appendix

Trivia Questions Used in Experiments 2 and 3

Question	Answer	Study
What U.S. state has the highest percentage of people who walk to work?	Alaska	2 and 3
What was used to power the engines of the starship *Enterprise* in the *Star Trek* television series?	Antimatter	2 and 3
Dr. John S. Pemberton invented Coca-Cola in 1886 in what city?	Atlanta	3
What is the present-day name of the land that Columbus called "San Salvador" in 1492?	Bahamas	2 and 3
What is the world's tallest grass?	Bamboo	2 and 3
What sport is the most common cause of eye injuries in the United States?	Baseball	3
What was the first trademarked product?	Beer	2 and 3
What is the name of the dog on the Cracker Jack box?	Bingo	2 and 3
What is the name of the country in which the game of dominos was invented?	China	2 and 3
What is the oldest inhabited city in the world?	Damascus	2 and 3
What was the name of the cat Alice left behind when she fell down the rabbit hole in *Alice's Adventures in Wonderland*?	Dinah	2 and 3
Who earned infamy for noting: "A billion dollars isn't worth what it used to be"?	Getty	3
What do you call a village without a church?	Hamlet	2 and 3
What was the first song to be sung in outer space?	Happy Birthday	3
What kind of poison did Socrates take at his execution?	Hemlock	3
Who was *Time Magazine*'s "Man of the Year" in 1938?	Hitler	2 and 3
The world's first skyscraper was an office for what?	Home Insurance	3
What city has the most Rolls-Royce per capita?	Hong Kong	3
What was the first of H. J. Heinz's "57 varieties"?	Horseradish	2 and 3
What nation consumes the most Coca-Cola per person?	Iceland	2 and 3
What is the last name of the first U.S. president born outside the 13 original states?	Lincoln	2 and 3
What nation produces two thirds of the world's vanilla?	Madagascar	2 and 3
What is the last name of the shortest American president?	Madison	2 and 3
What was the name of the dog from *The Grinch Who Stole Christmas*?	Max	2 and 3
What was the first capital of ancient Egypt?	Memphis	2 and 3
What is the only metal that is liquid at room temperature?	Mercury	3
What is the most common mammal in the United States?	Mouse	2 and 3
What planet has surface winds that have been measured at 1,500 mph—the strongest in the solar system?	Neptune	2 and 3
What bird's eye is bigger than its brain?	Ostrich	3
What is a group of owls called?	Parliament	2 and 3
What is the name of the constellation that looks like a flying horse?	Pegasus	3
What was the first U.S. consumer product sold in the Soviet Union?	Pepsi	2 and 3
What country has the world's highest railway?	Peru	2 and 3
In what book did the name Wendy first appear in print?	Peter Pan	3
In what city was the first U.S. zoo built?	Philadelphia	2 and 3
What is the only English word with a completely different meaning when the first letter is capitalized?	Polish	3
What is the longest English word without the normal vowels, *a, e, i, o,* or *u*?	Rhythms	2 and 3
What was the first city in the world to have a population of more than 1 million?	Rome	2 and 3
In what California city did the last Pony Express ride end?	Sacramento	2 and 3
What is the name of the brightest star in the sky excluding the sun?	Sirius	2 and 3
What trade was Greek philosopher Socrates trained for?	Stonecutting	2 and 3
What was the name of the horse Teddy Roosevelt rode in the Battle of San Juan Hill during the Spanish–American War?	Texas	2 and 3
In what country is Angel Falls located?	Venezuela	2 and 3
What is the only planet in the solar system to rotate clockwise?	Venus	2 and 3
In which state were the first peanuts in the United States grown?	Virginia	2 and 3
What was the name of the dog that was with Rip Van Winkle when he fell asleep for 22 years?	Wolf	2 and 3
What was the first state to allow women to vote?	Wyoming	2 and 3
What company was the first to offer a mouse on a commercially available computer?	Xerox	3

Received February 11, 2013
Revision received May 21, 2013
Accepted May 21, 2013 ■