Journal of Experimental Psychology: Applied
2009, Vol. 15, No. 3, 243–257

1076-898X/09/$12.00 DOI: 10.1037/a0016496

The Pretesting Effect: Do Unsuccessful Retrieval Attempts Enhance Learning?

Lindsey E. Richland
University of California, Irvine

Nate Kornell
Williams College

Liche Sean Kao
University of California, Irvine

Testing previously studied information enhances long-term memory, particularly when the information is successfully retrieved from memory. The authors examined the effect of unsuccessful retrieval attempts on learning. Participants in 5 experiments read an essay about vision. In the test condition, they were asked about embedded concepts before reading the passage; in the extended study condition, they were given a longer time to read the passage. To distinguish the effects of testing from attention direction, the authors emphasized the tested concepts in both conditions, using italics or bolded keywords or, in Experiment 5, by presenting the questions but not asking participants to answer them before reading the passage. Posttest performance was better in the test condition than in the extended study condition in all experiments—a pretesting effect—even though only items that were not successfully retrieved on the pretest were analyzed. The testing effect appears to be attributable, in part, to the role unsuccessful tests play in enhancing future learning.

Keywords: testing, learning, memory, retrieval

Testing has become a central issue in the current U.S. political debate concerning education. To ensure equal access to a high-quality education, operationalized as proficiency on state academic assessments (No Child Left Behind Act, 2001), educational reforms have replaced instruction—sometimes several weeks' worth each year—with standardized testing in an effort to monitor students' knowledge. These tests reduce time spent on curricula, but serve as diagnostic tools and accountability instruments, alerting teachers and administrators to low-performing student populations in need of additional services or reform. The diagnostic function of testing has merit, but there is a second benefit of testing that is often overlooked: Testing enhances memory for the tested material. Taking advantage of the memorial benefits of tests, and integrating testing into the curriculum rather than as an event that *follows* instruction and learning, has the potential to increase the efficiency and utility of school testing practices if this finding were better understood.

A survey of naive undergraduates supports the claim that tests are viewed principally as assessments in the United States. Kornell and Bjork (2007) asked undergraduates whether they tested themselves when they were studying, and if so, why. Whereas most students did report testing themselves (91%), most stated that they did so to "to figure out how well I have learned the information I'm studying." Only 18% described their testing as a learning event (Kornell & Bjork, p. 222).

Tests as Learning Events

Research suggests that testing information that has already been studied not only provides a measure of learners' knowledge, tests also become learning events in their own right. Indeed, testing has often been shown to be more effective than further study in encouraging retention of tested information (e.g., Bjork, 1988; Carrier & Pashler, 1992; Gates, 1917; Glover, 1989; Hogan & Kintsch, 1971; Izawa, 1970; McDaniel, Roediger, & McDermott, 2007; Roediger & Karpicke, 2006a, 2006b; Rothkopf, 1966; Tulving, 1967; Whitten & Bjork, 1977; for a review, see Richland, Bjork, & Linn, 2007). Researchers studying the cognitive underpinnings of testing have argued that testing should be considered a strategy for knowledge acquisition above and beyond its utility as a measure of current knowledge.

Testing as an instrument serving larger instructional goals has traditionally been seen to have a limitation, however: The benefits of testing are most pronounced for test items that were answered correctly (Butler & Roediger, 2007; Karpicke & Roediger, 2007;

Lindsey E. Richland and Liche Sean Kao, Department of Education, University of California, Irvine; Nate Kornell, Department of Psychology, Williams College, Los Angeles.

The Office of Naval Research Grant N000140810186 partially supported the experiments reported herein. This material is also based on work supported by the National Science Foundation under Grant 0757646. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We thank H. L. Roediger for insightful comments on the article, and Keara Osborne for invaluable assistance in running participants. Experiments 1, 2, and 4 were previously published in the proceedings of the Cognitive Science Society (Richland, Kao, & Kornell, 2008).

Correspondence concerning this article should be addressed to Lindsey E. Richland, Department of Education, University of California, 2001 Berkeley Place, Irvine, CA 92697. E-mail: l.e.richland@uci.edu

Leeming, 2002; Roediger & Karpicke, 2006a, 2006b). Generally, items not retrieved correctly when tested see minimal, if any, benefit of testing when compared with being allowed additional study time (for exceptions, see Izawa, 1970; Kane & Anderson, 1978; Kornell, Hays, & Bjork, in press). Unsuccessful tests may even have negative consequences. Proponents of errorless learning (e.g., Guthrie, 1952; Skinner, 1958; Terrace, 1963) suggest that failing to answer a question or answering incorrectly makes future errors more likely. Furthermore, being measured alters knowledge representations, and sometimes questioning can lead to memory distortions (see Davis & Loftus, 2007; Roediger & Marsh, 2005). Thus, testing has the potential to distort knowledge, particularly for any items not recalled correctly.

Providing detailed feedback after a test can ameliorate some of these challenges (e.g., Butler, Karpicke, & Roediger, 2007; Kang, McDermott, & Roediger, 2007; Metcalfe & Kornell, 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005), but this type of feedback is burdensome and often not feasible. This is particularly true in standardized testing, when feedback is rarely individualized by question and is often available to students and teachers only after a substantial delay. Thus, for the lowest performing students, who are No Child Left Behind's foremost priority, testing—in particular, failed tests—may have little value (or worse).

Can Failed Tests Improve Future learning?

The current research posits that the benefits of testing may extend to items that are not answered correctly on the test, and that failure to answer test questions should not be equated with a failure to learn. Rather, five experiments were conducted to evaluate the impact of restructuring the testing environment to actually incur *more* failed tests. Specifically, we evaluated the benefits of testing novel science instructional content *before* learning. Thus, the likelihood of failed tests was high, but we were able to extend our theory of testing to better understand whether trying and failing on test questions actually improved learners' longer term retention of subsequently presented information.

Pretests are regularly used as assessments in pre–posttest design studies with the expectation that they do not affect learning. There were some reasons, however, to expect that pretesting could enhance learning. Many studies have demonstrated benefits of pretraining activities such as advanced organizers (see Huntley & Davies, 1976; Mayer, 1979), outlines (e.g., Snapp & Glover, 1990), and statements to activate learners' prior knowledge schemas (e.g., Bransford & Johnson, 1972). Test questions have also been studied as pretraining activities, beginning with early experiments on the effects of integrating adjunct questions into text passages (e.g., Anderson & Biddle, 1975; Huntley & Davies, 1976; Pressley, Tanenbaum, McDaniel, & Wood, 1990; Rothkopf, 1966; Sagerman & Mayer, 1987). Adjunct questions interwoven into texts, both before and after the target information had been provided, showed improved retention of information asked about in the question and, less reliably, information not asked about (see Anderson & Biddle, 1975; Mayer, 2008; Rickards, 1976).

This basic pattern has held for both direct questions with basic text materials and more complex learning environments with higher level questions. For example, using "deep-level-reasoning questions" to introduce and frame interactions with an automated tutoring system, Autotutor, can greatly affect learning (e.g., Craig, Gholson, Ventura, Graesser, & the Tutoring Research Group, 2000; Craig, Sullins, Witherspoon, & Gholson, 2006; Gholson & Craig, 2006). In some circumstances, integrating these questions into instructional content can make noninteractive instruction as effective as an interactive tutor (VanLehn et al., 2007). Related research demonstrates that training "self-questioning" improves critical thinking and learners' ability to construct knowledge from forthcoming instruction (see King, 1992, 1994).

The early studies on adjunct questions, and the more recent studies with more inferential, higher level questions, did not attempt to contrast failures at the time of testing with successes. Rather, the most common interpretation of the questions' effects on later retention rested on their impact on readers' intentional learning behaviors. Rothkopf (1965, 1966, 1982) coined the term *mathemagenic behaviors* to explain the intentional learning behaviors of readers that are alterable by the instructional activities they encounter. For example, Rothkopf and Bisbicos (1967) found that asking participants questions in which the answers were numbers led to better retention of all numerical information in the text, possibly because participants were able to direct their attention to the type of information that was important to learn given the test they would take.

Direct tests of attention, based on measures of reading time and reaction time to a secondary task, demonstrate that people pay greater attention to reading a text when adjunct questions are interwoven (Reynolds & Anderson, 1982; Reynolds, Standiford, & Anderson, 1979). A practice guide published by the Institute of Education Sciences (an institute within the U.S. Department of Education) reviewed recent research with a similar conclusion, making the instructional recommendation: "We recommend . . . using 'prequestions' to activate prior knowledge and focus students' attention on the material that will be presented in class" (Pashler et al., 2007, p. 30).

In addition to affecting learners' attention and intentional learning behaviors, pretesting may provide a direct impact on memory. The cognitive benefits of testing *after* studying are well established to persist even when there is no opportunity to restudy information (e.g., Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b), which rules out the possibility that those benefits are explainable by attention during text processing. We thus investigated whether there was a similar cognitive benefit for pretesting above and beyond the effect of drawing learners' attention to testable information.

Unlike most previous studies, Pressley et al. (1990) did distinguish the effects of attention direction from the effects of testing itself using pretest questions. Their participants recalled more when they had been asked about the passage before reading it than when they had been presented with the same pretest questions, but had not been asked to try to answer them, before reading the passage (instead, participants were asked whether or not the questions were well written). Because the questions had dichotomous answers, however, participants were frequently able to answer correctly during the pretest.

The current experiments followed a similar study premise, but sought to test more directly whether unsuccessful retrieval attempts enhance retention of tested content beyond directing attention during study. Therefore, in the current study, the prequestions required participants to produce nouns or descriptive statements that they were unlikely to be able to answer on the basis of prior

knowledge (e.g., "What is total colorblindness caused by brain damage called?"). This allowed us to isolate and examine the effects of unsuccessful retrieval attempts.

The questions tested knowledge for exact information presented in the text, rather than knowledge that would require inferential or higher level thinking. Such questions are effective (e.g., Marsh, Roediger, Bjork, & Bjork, 2007; Rickards, 1976; Rickards & Hatcher, 1977–1978; Watts & Anderson, 1971; Yost, Avila, & Vexler, 1977) but could not be used in this study. They would have prevented us from adequately controlling for the fact that test questions draw learners' attention to testable content. Rather, we wanted to be able to create a no-pretest control condition in which we could draw participants' attention to the same information asked about in the test questions.

Typographical cuing (e.g., underlining or bolding; for reviews, see Glynn, Britton, & Tillman, 1985; Waller, 1991) is effective at drawing attention to cued items, sometimes to the exclusion of uncued items (Glynn & DiVesta, 1979). By typographically cuing participants to the aspects of the passages that would be tested, we expected to draw their attention to the key content that needed to be learned. This would allow us to distinguish between the effects of attention direction and any additional benefits of unsuccessful retrieval attempts.

The Present Experiments

We report four experiments that examined the learning effects of pretesting, beyond directing attention to testable information, when the questions were answered incorrectly. Theoretically, we sought to analyze the effects of attempting (but failing) to retrieve or generate test answers from memory, as distinct from participants' use of more directed search strategies while reading the text. In a fifth experiment, we further distinguished between attempting to retrieve answers to test questions and other deep processing of the pretest questions.

In all experiments, participants were asked to read a scientific text about vision in an unstructured reading situation, akin to how a learner might study a textbook. In the first experiment, participants were either tested prior to learning or they were given additional time to study. In Experiments 2 through 5, variations on the same procedure were used to isolate the effect of attempting to derive an answer to a question from the more mundane effect of directing attention by preexposing questions. In Experiment 2, all tested sentences were italicized in the studied text; in Experiment 3, the keyword from each tested sentence was bolded. Experiment 4 used bolded text and assessed the impact of testing versus extended study after a 1-week delay. Experiment 5 sought to differentiate between *reading* potential test questions and attempting to *answer* test questions before studying. Similar to Pressley et al. (1990), we manipulated whether participants memorized the pretest questions versus produced an answer to the same questions.

Experiment 1

We predicted that testing before study would enhance future recall, in spite of learners' failure to provide successful answers to the test questions.

Method

Participants

Participants in this study were 63 undergraduates who were given extra course credit for participating.

Materials

Study materials were selected from Sacks (1995). A two-page text was developed on the basis of an essay about a patient with cerebral achromatopsia (colorblindness caused by brain damage). This text was selected because of its rich scientific content and engaging narrative. The reading level was deemed appropriate for undergraduates, and Sacks's book is assigned in undergraduate coursework. To protect against the possibility that participants had read the passage in coursework, participants were asked whether they had read the passage previously, in which case they would have been excluded. None were excluded for this reason. The length of the story was designed to ensure that participants were not under time pressure and had time to return to sections if they desired to do so.

Some of the text described Sacks's patient suffering from cerebral achromatopsia, as in the following sample:

> I am a rather successful artist just past 65 years of age. On January 2nd of this year I was driving my car and was hit by a small truck on the passenger side of my vehicle. When visiting the emergency room of a local hospital, I was told I had a concussion. . . . I have visited ophthalmologists who know nothing of this color-blind business. I have visited neurologists, to no avail. Under hypnosis I still can't distinguish colors. I have been involved in all kinds of tests. You name it. My brown dog is dark gray. Tomato juice is black. Color TV is a hodgepodge.

Other parts of the text were selected from the more scientific treatment of the disorder, as in the following sample:

> Colorblindness, as ordinarily understood, is something one is born with—a difficulty distinguishing red and green, or other colors, or (extremely rarely) an inability to see any colors at all, due to defects in color responding cells, the cones of the retina. Total colorblindness caused by brain damage, so-called cerebral achromatopsia, though described more than three centuries ago, remains a rare and important condition. It has intrigued neurologists because, like all neural dissolutions and destructions, it can reveal to us the mechanisms of neural construction, specifically, here, how the brain "sees" (or makes) color. (Sacks, 1995, pp. 3–4)

Within the reading packet, 10 sentences were identified as testable items. Test questions were constructed on the basis of these 10 sentences. Two counterbalanced pretests were constructed such that each contained questions about 5 of the selected sentences. Questions were written as fill-in-the-blank or short free-response items (e.g., "What is total color blindness caused by brain damage called?" and "How does Mr. I distinguish red and green traffic lights?"). They addressed facts presented in the text, either general scientific facts or information about the specific patient. See Appendix A for all questions.

A final test included all 10 of the testable items in randomized order. Thus, for participants in the test condition, 5 of the final test questions had been pretested during Time 1 (tested) and 5 had not

been tested previously (untested). Questions from the two pretest versions were always interspersed on the final test. All questions were new for participants in the extended study condition.

Procedure

The experiment was conducted in a group setting. Participants were randomly assigned to an *extended study* condition ($n = 27$) or a *test and study* condition ($n = 36$). We conducted this experiment in a lecture class setting, and assigned participants on the basis of seating. The lecture space had separate seating areas and participants were assigned on the basis of those. This was done to ensure that each section followed the appropriate timing, but we did not have tight control over cell size.

Learning phase. Participants in the test and study condition were given one of the two counterbalanced pretests and allowed 2 min to answer the questions. They were instructed to provide an answer to all five questions, regardless of whether they knew the answer. At the end of 2 min, the pretests were collected, and participants given the text passage and told to study it for 8 min. They were instructed to read the passage through in its entirety at least once.

Participants in the extended study condition were given 10 min to study the passage—the same total time that participants in the test and study condition spent in testing and study of the material. They were given the same reading instructions.

Final test. The text passages were collected after the timed study periods were completed. Participants were then immediately administered the Time 2 test, which consisted of 10 questions. The test was untimed to ensure that time pressure did not affect performance.

Results

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 5% of the questions correctly. Any items answered correctly on the Time 1 pretest were removed from the following analyses of Time 2 test scores on a participant-by-participant basis. Most participants gave an answer for all questions, often providing answers that were incorrect yet appropriate (e.g., writing the name of a scientist in a question referring to Isaac Newton).

An independent samples *t* test examined the effects of testing by comparing mean posttest percentage correct for tested items in the test and study condition with the overall mean score in the extended study condition. As shown in Figure 1, testing resulted in better posttest performance ($M = 75\%$, $SE = 3.2$) than did the provision of extra time to study the same material ($M = 56\%$, $SE = 2.7$), $t(61) = 4.26$, $p < .0001$, $d = 1.1$.

Examining performance within the test and study condition only, tested items ($M = 75\%$, $SE = 3.2$) were recalled on the final test significantly more often than untested items ($M = 50\%$, $SE = 3.4$), $t(35) = 5.03$, $p < .0001$, $d = 1.7$, in spite of the fact that the analyses excluded any items that participants recalled correctly on the pretest. The benefit of testing did not spread to untested items, but neither did it hurt. There was not a significant difference between accuracy on the untested items in the test and study condition and in the extended study condition, $t(61) = 1.3$, $p = .20$.

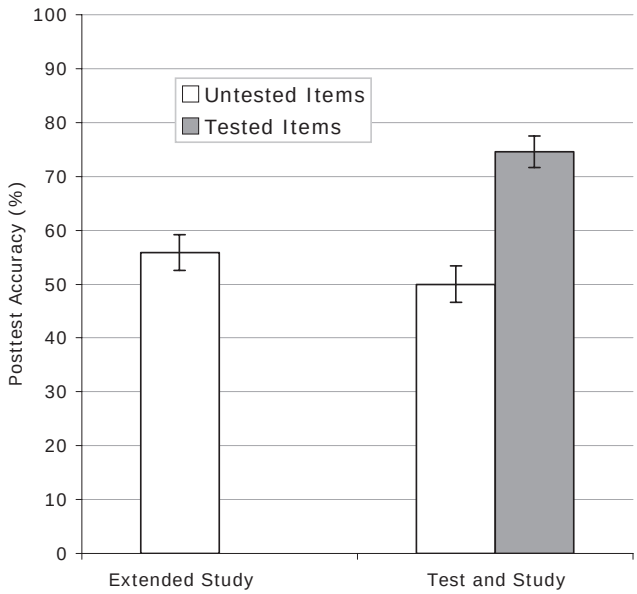

Figure 1. Experiment 1: Performance on a final test across conditions when studying an unmarked text.

Discussion

Experiment 1 revealed that failed tests *can* enhance learning for educational content. Although participants largely failed on the initial test (answering 95% of the questions incorrectly), the effect of those failures was to increase retention of studied content when compared with an extended opportunity to study the materials without being pretested.

The explanation for the benefit of unsuccessful tests is not yet clear. One possibility is that the test directed learners' attention to the key, testable points in the passage. Alternatively, attempting to retrieve an answer to the test problem may have provided an additional benefit above and beyond the impact of attention direction. Experiment 2 used the same procedure as Experiment 1, but all testable sentences were italicized to equalize participants' attention to key concepts in the text. We reasoned that under such conditions, allocation of attention would not differ meaningfully between conditions; therefore, differences in learning would be attributable to the impact of retrieval attempts during the pretest.

Experiment 2

We predicted that pretesting followed by study would enhance future recall more than the provision of extended time to study an instructional text, even when differences in attention direction were minimized by italicizing key sentences in the text in both conditions.

Method

Participants

The participants were 61 undergraduates (mean age = 21 years, 44 women and 17 men) who were given extra course credit for participating. Participants were sampled from an upper division

psychology course on human stress. Data from 2 participants were excluded from analyses because of a failure to respond to final test questions.

Materials

The study materials were the same text and testable sentences as used in Experiment 1. The key difference was that within the reading packets, the 10 testable sentences were italicized. Italicizing was considered a way to ensure that all participants were equally alerted to what was deemed to be important information in the same way that many textbooks emphasize key elements of a chapter. Participants in both conditions read the same italicized text. For example, see the following text paragraph:

> The history of our knowledge about the brain's ability to represent color has followed a complex and zigzag course. *Newton, in his famous prism experiment in 1666, showed that white light was composite—could be decomposed into, and recomposed by, all the colors of the spectrum.* The rays that were bent most ("the most refrangible") were seen as violet, the least refrangible as red, with the rest of the spectrum in between. (Sacks, 1995, p. 18)

Procedure

The procedure was exactly the same as the procedure in Experiment 1. Participants were tested in a group setting and were randomly assigned to the extended study condition ($n = 26$) or to the test and study condition ($n = 33$). Participants were not given any specific instruction regarding the text italics.

Results

In the test and study condition, participants answered 22% of questions on the initial pretest correctly. Correct answers were distributed across test problems. The population of participants in this experiment seems to have had a higher level of relevant background knowledge on pretest items than in Experiment 1, perhaps because they were sampled from a higher level psychology course, but as in Experiment 1, any items answered correctly at Time 1 were removed from the following analyses on a participant-by-participant basis. If anything, this led to inflation in participants' scores in the untested conditions, counter to our hypothesis.

The data revealed benefits for testing over the provision of extra time for studying the same material. As Figure 2 shows, recall of tested and italicized items in the test and study condition ($M = 71\%$, $SE = 5.6$) was significantly greater than recall of italicized-only items in the extended study condition ($M = 54\%$, $SE = 3.7$), $t(57) = 2.3$, $p < .022$, $d = 0.61$.

Examining performance within the test and study condition, tested items were recalled on the final test ($M = 71\%$, $SE = 5.6$) significantly more often than untested, italicized-only items ($M = 53\%$, $SE = 4.3$), $t(32) = 3.27$, $p < .003$, $d = 0.63$, in spite of the fact that the analyses excluded items that participants recalled correctly on the pretest. Testing did not appear to negatively affect the untested items; there was not a significant difference between accuracy on the italicized-only items in the test and study condition and the italicized-only items in the extended study condition, $t(57) = 0.15$, $p = .88$.

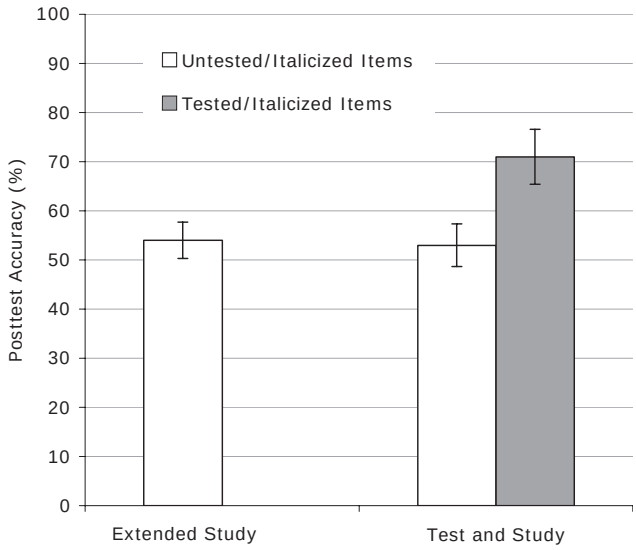

Figure 2. Experiment 2: Performance on a final test across conditions when studying text with italicized key sentences.

Discussion

The results of Experiment 2 replicated the results from Experiment 1, and again suggest that the testing effect can and should be extended to failed tests. Testing items created more potent learning opportunities than extended study of the same items, even when the key information in both conditions was italicized, equalizing attention direction. Thus, testing appears to provide a unique benefit above and beyond directing learners' attention to content that has a high probability of being tested later.

In textbooks, italicized sentences are less common than bolded keywords, which are ubiquitous. It remains possible that participants in Experiment 2 were unfamiliar with the meaning of italics within text, and thus differences in attention were not minimized. To rule out that possibility, Experiment 3 used the same procedure as Experiment 2, but bolded keywords were used instead of italicized sentences because we expected that bolding might act as a stronger (and more realistic) attention prompt. Experiment 3 thus examined the impact of testing when compared with extended opportunities to study text in which the key test items were bolded.

Experiment 3

We predicted, similar to Experiment 2, that testing before reading would enhance future recall above and beyond the impact of extended study time. Instead of presenting key sentences in italics, keywords were presented in bold.

Method

Participants

Participants in this study were 64 undergraduates (44 women, 17 men, 3 unstated) who were given extra credit in their courses for participating. Participants' average age was 22 years.

Materials

The test materials were exactly the same as those used in Experiments 1 and 2. The study materials were exactly the same as those used in Experiments 1 and 2, with the exception of the treatment of the 10 testable sentences. Within the reading packet, one word was bolded from each of the sentences that had been italicized and tested in Experiment 2. The bolded word was the answer to the fill-in-the-blank or short-answer questions used in the tests. An example of a paragraph with bolded words follows:

> Colorblindness, as ordinarily understood, is something one is born with—a difficulty distinguishing red and green, or other colors, or (extremely rarely) an inability to see any colors at all, due to defects in color responding cells, the **cones** of the retina. Total color blindness caused by brain damage, so-called **cerebral achromatopsia**, though described more than three centuries ago, remains a rare and important condition. (Sacks, 1995, pp. 3–4)

Procedure

The procedure was exactly the same as the procedure in Experiments 1 and 2. The experiment was conduced in a group setting. Participants were randomly assigned to the extended study condition ($n = 33$) or the test and study condition ($n = 31$). No specific instructions were given regarding the bolded text.

Results

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 21% of the questions correctly. Two pretest items about vision were answered correctly at unexpectedly high rates, something that had not occurred in the previous experiments, so these questions were removed from all further analyses of posttest data for this experiment in both conditions. Excluding those questions led to a pretest average performance level of 11%. Any other items answered correctly at Time 1 were removed from the following analyses on a participant-by-participant basis.

As shown in Figure 3, tested and bolded items in the test and study condition were recalled significantly more often on the final test ($M = 82\%$, $SE = 3.8$) than were bolded-only items in the extended study condition ($M = 64\%$, $SE = 4.0$), $t(62) = 3.3$, $p < .002$, $d = 0.84$, revealing a benefit for testing over extra time spent studying the same material. Even when keywords were bolded in both conditions, pretesting led to higher retention of bolded and tested items than did extended study.

Within the test and study condition, there was a numerical advantage for tested and bolded items ($M = 82\%$, $SE = 3.8$) over items that were bolded but not tested ($M = 77\%$, $SE = 3.0$), but unlike in Experiments 1 and 2, the difference was not significant, $t(30) = 1.4$, $p = .17$. This lack of difference may indicate that even untested items benefited from testing. Indeed, untested items in the test and study condition were recalled at a higher rate than items in the extended study condition, a difference that approached significance, $t(62) = 1.9$, $p = .062$, $d = 0.48$. Although this finding was not reliable across all studies reported herein, it is consistent with the early arguments that testing before learning affects readers' intentional learning practices. At minimum, these data suggest that testing did not hurt recall of untested items when keywords were bolded.

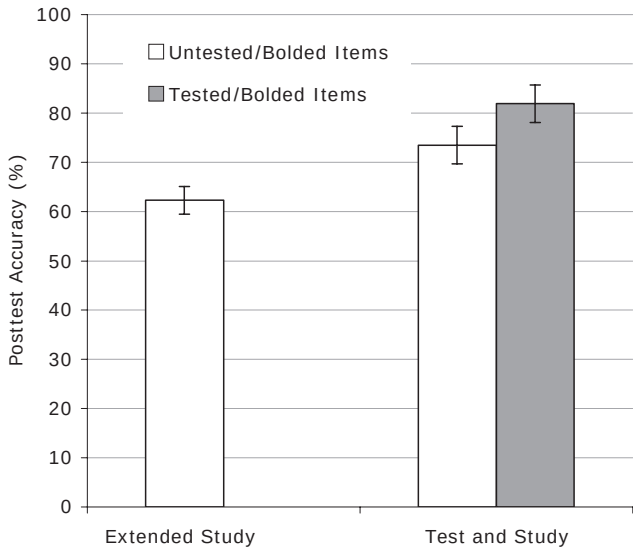

Figure 3. Experiment 3: Performance on a final test across conditions when studying text with bolded keywords.

Discussion

Experiment 3 demonstrated that unsuccessful tests can enhance learning for new educational content, replicating and extending the findings from Experiments 1 and 2. Testing items before learning was a more potent learning opportunity than the provision of extended study time, even when keywords were bolded in the text and only items that participants failed to answer on the initial test were included in the analyses. Once again, these results suggest that testing provides a unique benefit above and beyond serving to direct learners' attention to materials that might be tested at a later point. The results of Experiment 3 also suggest that testing some items may additionally benefit learning for untested items.

Experiment 4

In the first three experiments, the effects of pretesting were measured on an immediate test. Previous research has shown, in the context of successful tests, that the size of the testing effect grows as the delay between study and a final memory test increases because tested items are forgotten more slowly than items that have not been tested (Hogan & Kintsch, 1971; Roediger & Karpicke, 2006b). In Experiment 4, to investigate the effect of delaying the final test for items that have been tested unsuccessfully, we examined learning after a 1-week delay. Doing so was also a way to connect the findings with the goals of education, which involve improving long-term learning. There was also a second change to Experiment 4. To better distinguish between the effects of bolding and testing, we manipulated bolding within subjects. Testing versus extended study remained a between-subjects manipulation.

We predicted that the results would be similar to the results of the previous experiments—that is, that final test performance at a

delay would be higher for items that were pretested in the test and study condition than for items that were bolded in the extended study condition, even if the retrieval attempts on the pretest were unsuccessful. We also predicted that bolding would benefit retention relative to nonbolded items in the extended study condition, but that testing would be more advantageous than bolding.

Method

Participants

Participants in this study were 158 undergraduates (137 women, 15 men, 6 not stated; mean age = 20 years) who were given extra course credit for participating.

Materials

The study materials were exactly the same as those used in Experiment 3 with one exception: Rather than emphasizing all 10 key concepts, as in Experiments 2 and 3, only 5 items were bolded. In the test and study condition, the 5 items tested on the Time 1 pretest were the same items that were bolded. In the extended study condition, 5 corresponding items were selected to be bolded. These were matched to the items tested in the counterbalanced test conditions. Thus, for a given participant, of the 10 items tested on the posttest, 5 had been emphasized during initial study (by being bolded in the extended study condition, or by being tested and bolded in the test and study condition), and 5 items had not. Tested and bolded items were counterbalanced across participants. This manipulation allowed us to make separate estimates of the effects of bolding and the effects of testing.

In addition, two questions that had received relatively high accuracy rates on the pretest were rewritten. See Appendix B for replacement questions.

Procedure

The learning phase of the experiment was identical to the learning phase of Experiments 1–3, except that, to control the timing of the final test, participants were tested individually. After completing the first session, participants were asked to return 1 week later at the same time of day. When they returned, the final test was administered. The test procedure was the same as the tests in the previous experiments. Participants were randomly assigned to the extended study condition ($n = 79$) or the test and study condition ($n = 79$).

Results

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 10% of the questions correctly. Items answered correctly on the pretest were removed from the analyses on a participant-by-participant basis.

The data were analyzed differently from those in Experiments 1–3 because posttest performance for both conditions could be separated into bolded and nonbolded items. Thus, there was a within-subjects manipulation of bolding and a between-subjects manipulation of testing. Because testing and bolding were manipulated together in the test condition (items were tested and bolded or untested and unbolded), this is not a full factorial design and effects were analyzed using one-tailed *t* tests. The effects of testing were examined by holding bolding constant between the testing and extended study conditions (bolded and tested vs. bolded). The effects of bolding were studied in the extended study condition (bolded vs. unbolded).

The results are displayed in Figure 4. There was a significant pretesting effect: Bolded and tested items in the test and study condition were recalled better ($M = 55\%$, $SE = 2.0$) than bolded-only items presented for longer study time in the extended study condition ($M = 45\%$, $SE = 3.0$), $t(156) = 2.8$, $p < .0025$, $d = 0.45$. When the test and study condition was examined separately, tested and bolded items were recalled at a distinctly higher rate ($M = 55\%$, $SE = 0.30$) than untested and unbolded items ($M = 42\%$, $SE = 3.0$), $t(78) = 3.3$, $p < .001$, $d = 0.75$. There was also a smaller difference between bolded items and unbolded when the extended study condition was examined separately, $t(78) = 1.9$, $p < .04$, $d = 0.43$ ($M = 45\%$, $SE = 2.2$, and $M = 39\%$, $SE = 2.6$, respectively), revealing that bolding was an effective typographical tool for directing attention. There were no significant differences between the test and study condition and the extended study condition on untested and unbolded items, $t(156) = 0.58$, $p = .28$ ($M = 42\%$, $SE = 3.0$, and $M = 39\%$, $SE = 3.0$, respectively).

Discussion

Experiment 4 demonstrated that failed tests can affect learning for educational content even after a 1-week delay, extending the findings from Experiments 1–3. Once again, the results suggest that testing provides a unique benefit above and beyond serving to direct learners' attention to materials that might be tested at a later point. Indeed, directing attention by bolding items provided a minimal benefit in the extended study condition, whereas bolding

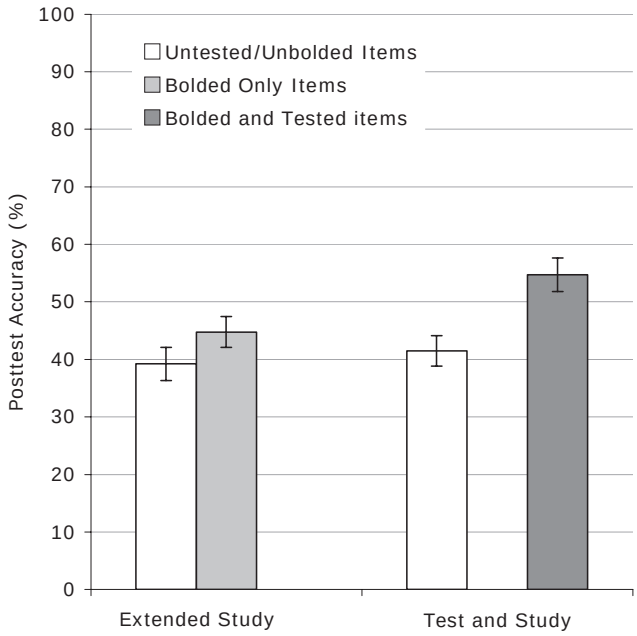

Figure 4. Experiment 4: One-week delayed performance on a final test across conditions when keyword bolding was manipulated within subjects.

accompanied by testing significantly enhanced learning in the test and study condition.

Although previous results have shown that testing effects sometime increase as the delay between study and final test increases (e.g., Roediger & Karpicke, 2006a), the sizes of the effects in Experiment 4 were comparable to the sizes of the effects in Experiment 3. This finding suggests that, unlike successful tests, unsuccessful tests may not slow the rate of forgetting, although further evidence would be needed to support that hypothesis.

Experiment 5

In Experiment 5, we investigated why unsuccessful tests enhance learning. Specifically, we sought to determine whether the pretesting effects that we had identified could be attributed to attempting (albeit unsuccessfully) to *answer* test questions versus simply seeing potential test questions before beginning to study. Providing test questions, even if participants do not have to answer them, could have similar results to pretesting, and thus explain the pretesting effects in Experiments 1–4 without reference to the direct benefits of tests. Two explanations for those results could include that (a) test questions may provide an organizational framework that indirectly affects retention by guiding future learning, and (b) allowing participants to read test questions may induce deep processing more effectively than does merely reading the passage.

Providing potential posttest questions to readers before reading the passage may serve as a guide for readers' interactions with the forthcoming text, either as an organizational framework to better structure causal structure and knowledge interpretations (e.g., Craig et al., 2000; Pashler et al., 2007) or by affecting learners' looking behaviors (Rothkopf, 1965, 1982). If this is the case, simply reading the questions before studying may be as effective as attempting to answer them—perhaps more so if answering questions incorrectly leads to retention for those incorrect answers (Marsh et al., 2007; Roediger & Marsh, 2005).

Alternatively, the levels of processing theory have been posed to explain the benefits of testing in general and may apply to both successful and unsuccessful tests. For example, Kane and Anderson (1978) found that asking participants to fill in the last word in a sentence helped them remember the correct last word, even when most of the words the participants filled in were incorrect. These authors hypothesized that testing resulted in a deeper level of processing than simply reading, which served to organize sentence information in participants' minds and make it recallable on a later test.

Ghatala (1981) provided further support for the notion that testing benefits learners mainly because it induces deep processing. Participants were asked to fill in the missing last word of a sentence; unlike Kane and Anderson's (1978) materials, the missing word was obvious and participants usually retrieved it successfully. Ghatala compared the retrieval condition to a condition in which participants did not retrieve, but were asked to do a task that induced deep processing. Ghatala found that "the operations involved in generating information from semantic memory have no special mnemonic value beyond inducing optimal processing of the material" (p. 443). For questions in which the missing word was obvious, retrieving the key word was no better than other deep processing of the sentence.

It is interesting that a follow-up study indicated that testing might produce an additional benefit over deep processing when the missing word was not obvious (Ghatala, 1983). Ghatala interpreted these combined results as suggesting that attempting to retrieve did not by itself provide a direct mnemonic advantage, but could indirectly improve learning by strengthening memory for the sentence's organizational structure. This echoes the benefits of providing "prequestions" or other advanced learning techniques to organize forthcoming knowledge.

For a different interpretation, one may consider the Pressley et al. (1990) findings, reviewed above, which showed benefits of pretesting greater than the benefits of viewing test questions before learning. When taken together with the Ghatala (1983) data, these findings suggest that failed tests may affect retention both directly and indirectly.

Experiment 5 investigated whether the benefits of unsuccessful tests result from the active attempt to recall key information from memory versus simply more active processing of the test sentence as an organizational structure. We compared the two conditions from the previous four experiments (i.e., extended study and test and study), as well as a third condition. In the third condition—the question learning condition—participants were asked, before reading the passage, to memorize the test questions without attempting to answer them (similar to the procedure used by Pressley et al., 1990). We expected that trying to memorize the questions would induce a relatively deep level of processing as participants focused on integrating the semantic structure of test sentences. Thus, any advantage of processing the test questions as organizational structures should be comparable across the two prequestion conditions.

This procedure led to two conflicting predictions. On the basis of Ghatala's (1981, 1983) results and interpretation, we anticipated that testing and question learning might have equivalent effects because both induce deep processing and support learning for test sentences as organizational structures. On the basis of the hypothesis that attempting to retrieve is more effective than deep processing alone, however, and in agreement with Pressley et al. (1990), we predicted that pretesting would lead to higher retention rates than would attempting to memorize and reproduce test questions without answering them.

Method

Participants

The participants in Experiment 5 were 76 undergraduates (64 women, 12 men), with an average age of 20 years, who were given extra credit in their courses for participating. An additional 3 participants were excluded from the analyses for failing to complete the posttest, and 2 were excluded for prior knowledge of the tested passage.

Materials

The study materials were similar to those used in Experiment 4; for any given participant, half of the key concepts in the text were emphasized. We returned to italicizing multiple words (as in Experiment 2) rather than bolding single words (as in Experiments 3 and 4) because doing so provides more information to the learner about exactly what information to focus on. Italicizing might also

be an effective way to focus participants' attention because of its novelty. Instead of italicizing whole sentences, as in Experiment 2, we italicized key phrases within the sentences to make the direction of attention more precise (e.g., "*Total color blindness caused by brain damage, so-called cerebral achromatopsia*, though described more than three centuries ago, remains a rare and important condition").

The test questions used in the pre- and posttests were modified versions of the tests used in Experiments 3 and 4. Because participants would be memorizing the questions, we wanted to minimize difference in form, length, and number of untested facts included in the question text. All questions were rewritten to fill-in-the-blank format and longer questions were simplified and shortened. See Appendix C for modified questions.

Procedure

The procedure was the same as Experiments 1, 2, and 3, except that there was a third between-participants condition, in which a new set of instructions was given during the Time 1 pretest. Participants were tested individually and were randomly assigned to one of three conditions: extended study ($n = 26$), test and study ($n = 24$), and question learning and study ($n = 26$).

The procedures for the extended study and test and study conditions were the same as Experiments 1–4. In the question learning and study condition, participants were given the same counterbalanced Time 1 tests as were used in the test and study condition. Instead of being asked to answer the questions, however, participants were asked to memorize the test questions because, they were told, they would be testing another person on the questions. They were asked to pay careful attention to where the blank fell in the question. This instruction was intended to support the students in learning the question without filling in an answer. After 2 min of studying the questions, participants were given a blank sheet of paper and asked to write down the questions. They were again cautioned to make sure to leave a blank in the correct spot. This procedure step provided an additional level of processing the question.

Results

In the test and study condition, on the initial test that preceded the presentation of the passage, participants answered 6% of the questions correctly. All items answered correctly on the initial test were removed from analyses of posttest performance.

The results are shown in Figure 5. One-tailed planned comparisons first examined the hypothesis that participants in the test and study condition would outperform participants in the question learning and study and extended study conditions on ability to answer posttest questions on italicized keywords. The test and study condition outperformed the question learning and study condition, $t(48) = 2.04$, $p < .02$, $d = 0.59$, which in turn outperformed the extended study condition, $t(50) = 2.02$, $p < .02$, $d = 0.57$ (test and study: $M = 90\%$, $SE = 3.8$; question learning: $M = 78\%$, $SE = 4.2$; extended study: $M = 63\%$, $SE = 5.9$). As expected, the largest difference was between the test and study condition and the extended study condition, $t(48) = 3.7$, $p < .001$, $d = 1.1$.

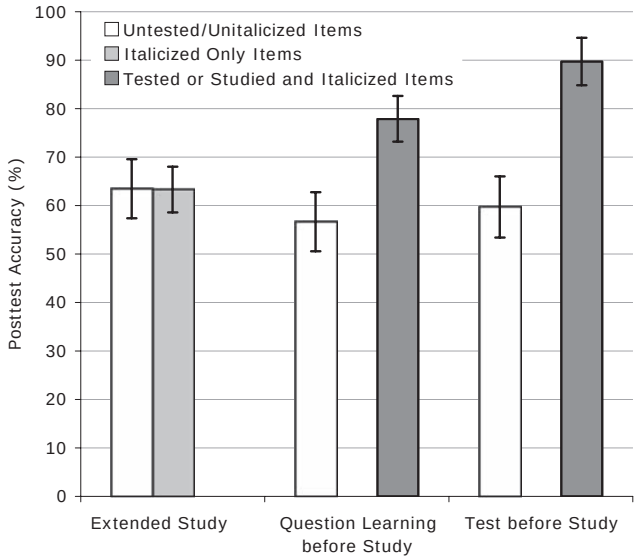

Figure 5. Experiment 5: Performance on a final test following varied pretest learning activities and study.

To examine any related differences on untested items, a one-way analysis of variance compared performance across conditions on items that were neither tested nor italicized. There were no differences between conditions on these items, $F(1, 73) = 0.31$, $MSE = 0.096$, $p = .74$.

Discussion

The results of Experiment 5 extend findings from the previous experiments demonstrating the benefit of attempting to retrieve a response to a question, even when the attempt is unsuccessful. Attempting to answer a prequestion was significantly more effective than reading the same question and attempting to memorize it without making an attempt to retrieve the answer. It is even possible that the benefits of the question learning and study condition were attributable to participants attempting to answer some of the questions despite being asked not to—in essence, to the benefit of testing. These data suggest that the benefits of testing extend beyond the benefits of engaging in deep processing. Most important, the data support the hypothesis that unsuccessful tests are useful because of their role as tests, apart from the role prequestions may play in encouraging deep processing or supporting organization of forthcoming knowledge.

General Discussion

Previous research has demonstrated the memory benefits of successfully answering test questions. The five experiments reported herein provide evidence for the power of tests as learning events even when the tests are unsuccessful. Participants benefited from being tested before studying a passage—a pretesting effect—although they did not answer the test questions correctly on the initial test, as compared with being allowed additional study time. Furthermore, the benefits of pretests persisted after a 1-week delay. Tests can direct participants' attention to the important information, but such attention direction cannot explain the current

findings because the important information was highlighted in all conditions using italics or bolding in Experiments 2–5. Moreover, participants in Experiment 5 learned more after unsuccessfully attempting to answer test questions than they did after attempting to memorize, but not answer, the same questions. These data imply that testing has advantages that exceed the benefits prequestions may have in supporting the organization of knowledge structures or in giving rise to deep processing (see also Kornell et al., in press).

There was little cost to testing; it did not require the provision of additional time on task, and nontested items were not adversely affected when other items were tested in the test condition. The effects on untested items varied between experiments, showing a positive effect of testing in Experiment 3 but no significant differences across the other studies. These data do not seem to reveal a systematic pattern, but the important point is that, on the basis of the present findings, pretesting did not seriously impair retention of untested items, as it has been posited to do previously (see Frase, 1968, 1970; Pashler et al., 2007). We tentatively conclude that pretesting can be employed without significant risk to untested items.

Theoretical Implications of Unsuccessful Tests

The present findings suggest that the testing effect—that is, the finding that more learning occurs during testing than when information is presented without a test—is not solely a result of the benefits of successful attempts to retrieve information from memory. Successful tests may play a powerful role in enhancing memory, but attempting to retrieve information, by itself, enhances future learning.

From a cognitive perspective, there are a number of reasons why unsuccessful tests might enhance future learning. One reason is that retrieval strengthens retrieval routes between the question and the correct answer (e.g., Bjork, 1975, 1988; McDaniel & Masson, 1985). Participants frequently generated appropriate but incorrect answers, which might seem more likely to strengthen dead ends than retrieval routes; however, the function of a failed retrieval attempt may be to weaken or suppress errors, rather than to strengthen them (e.g., Carrier & Pashler, 1992). Alternatively, retrieving appropriate content potentially could have strengthened retrieval pathways to related content, identifying a need for additional information, thus strengthening the route even before the content was provided.

A second potential reason for the benefits of unsuccessful tests is that they can encourage deep processing of the question in a way that merely reading the question does not (Bjork, 1975; Carpenter & DeLosh, 2005; Ghatala, 1981, 1983; Kane & Anderson, 1978). To retrieve an appropriate answer to a question, a learner may attempt to imagine or creatively search for potential solutions. For example, even if a learner cannot think of the correct answer to a question such as "How does Mr. I distinguish red and green traffic lights?" the question may prompt the learner to picture a traffic light, think about approaching a traffic light while driving, consider what color blindness is like, what sorts of mishaps one might encounter, and so on. Even when these thoughts do not produce a correct answer, they may create a fertile ground for later encoding of the answer when it is eventually provided, and therefore may produce benefits similar to the effects of deep processing of the answer (e.g., Craik & Lockhart, 1972). Under some circumstances, an unsuccessful retrieval attempt might, by this logic, even result in more learning than a fast, successful retrieval attempt.

We examined the effects of deep processing of the question, in Experiment 5, by presenting participants with pretest questions, but asking them either to try to answer the question or to try to memorize the question. Both instructions were designed to induce deep processing of the semantic meaning of the question. Ghatala (1981) posited that, at least with respect to successful tests, the testing effect was attributable to the benefits of deep processing and internalizing the organizational structure of sentences. The present results suggest that testing was more beneficial than deep processing of the sentence. If participants in the test and study condition in Experiment 5 engaged in the type of thought processes described above (e.g., thought about approaching a traffic light while driving or other information outside of the strict confines of the question), however, testing may have resulted in deeper, more complex levels of processing than question memorization.

Thus, the nature of the processing learners perform during a prelearning activity may be more crucial than the amount of processing performed. Carpenter and DeLosh (2005) found, for tests administered after learning and before a final knowledge assessment, that the degree of elaborative processing required during testing was predictive of final test performance, regardless of the final test format. Free-response tests, which require the most elaborative processing, led to the highest overall retention, whereas recognition and cued recall produced smaller benefits. Similarly, in Experiment 5, attempting to retrieve an answer to the pretest questions could have produced qualitatively more elaborative processing than attempting to learn the test question as an organizational structure, even if the *amount* of processing in the two conditions was similar.

Applications to Educational Practice

Even if tests are not answered successfully, they have the potential to improve future learning, as measured by both immediate and delayed performance measures. This finding suggests that using tests as learning events in educational settings could have lasting benefits for learners' content acquisition, and that tests should be considered a potent learning opportunity, rather than simply as an assessment measure.

According to Bransford and Schwartz (1999), the quality and cognitive impact of a learning event can be measured, in part, by the impact of the learning event on future knowledge acquisition. Bransford and Schwartz emphasize the importance of "preparation for future learning" (p. 8) as a measure of transfer. The current experiments show that one way to prepare learners for future knowledge acquisition is to ask them to answer test questions before studying, even if they are unsuccessful in their attempts.

Although feedback on tests is known to aid learning (e.g., Kang et al., 2007), our data suggest that instruction following testing need not be individualized to learner errors. Rather, instruction that appropriately draws attention to key content may build on the previous cognitive acts performed when attempting to answer a test question. This implies that standardized tests, or other test situations where it is difficult to provide timely item-by-item feedback, could still provide learning benefits for successful and

unsuccessful test takers, as long as those test takers are given an opportunity to learn the information on which they were previously tested. Such a goal, however, would rely on close alignment between the test and subsequent learning opportunities. Standardized tests are not yet so closely aligned with curriculum, although such alignment could potentially enhance the usefulness of both the testing and the instruction that followed it.

At a practical level, pretests could be relatively feasible to implement in classrooms. Although research is necessary to clarify whether the same benefits apply across diverse texts and question types, many teachers already do some pretesting, and teachers could fairly easily make use of test bank or end-of-chapter questions that textbooks almost universally provide. Textbook questions often correspond to bolded keywords in the text, leading to similarities to the current experimental manipulations. As one usage that would be quite similar to the current experiments, pretests might help teachers ensure that students memorized key basic facts for a unit, freeing them to spend more subsequent time on more conceptual and inferential reasoning.

The reconceptualization of pretests as learning events might even aid teachers in optimizing formative assessments, that is, informal assessments that are integrated into daily classroom practice. Although formative assessments are used widely to measure learning, sometimes at the beginning of a unit before instruction, they are less often directed toward directly improving student learning (Black & William, 1998). These embedded assessments are usually intended to allow teachers to better modulate their instruction to meet students' knowledge levels, but they might also serve as potent learning tools in their own right. One could imagine direct empirical tests of this speculation that would mirror the current experiments, but in a dynamic, interactive classroom setting.

Profiting From Standardized Testing

Standardized testing is a more formal mode of assessment that plays an increasingly large role in classroom time. On the basis of the current No Child Left Behind Act, students are tested on their attainment of established curriculum standards for 2 weeks or more each year in many states, and that number is increasing as more districts seek to align with political pressures for assessment. Teachers and administrators alike describe these testing days as outside of instruction and as reducing an already affected curriculum schedule. Failed tests are viewed as indicating a lack of student progress and as a particularly egregious waste of needy students' time (e.g., Garrison, Jeung, & Inclán-Rodríquez, 2006; Hursh, 2007; Mathison & Freeman, 2003).

The current research lays the foundation for arguing that these testing days might be profitably integrated into the curriculum, and could actually facilitate subsequent learning for the unsuccessfully retrieved content. It is crucial, however, that after a test students be provided with an opportunity to restudy the tested material; a test that is not followed by instruction or feedback is likely to be of little use for items that were not answered correctly on the test. A recent study of standardized testing without aligned instruction suggests the same. When undergraduates and high school students were tested on retired SAT II questions without feedback, the tests led to an increase in posttest recall for participants who scored fairly well on the initial test, but led to no change in lower performing undergraduates and costs for high school students (Marsh, Agarwal, & Roediger, in press). Thus, in the absence of feedback or posttest learning opportunities, standardized testing may well be more problematic for lower performing students than higher performing students.

Reconsideration of these tests as learning events as well as assessments could have far-reaching implications for those reforms and interrelations with curriculum decisions. The alignment between tests and instruction has been the subject of substantial reform on the instructional side to ensure that instruction aligns with tested standards (e.g., see Amrein & Berliner, 2002; Herman & Golan, 1993; Stecher et al., 2008). Much less discussion has turned to tests' potential to directly affect students' learning and retention. Addressing this issue would require greater integration between testing and instruction, and such a shift could potentially address a common concern with high-stakes testing, namely, that the tests do not currently assess important aspects of the curriculum, thus reducing instructional depth.

Finally, another potential advantage of integrating standardized testing with instruction is self-regulatory and motivational, even in instances of failures. Although this was not the focus of the current analyses, "constructive failures" on test items may motivate learners and assist self-regulatory processes, such that learners become aware of what they do not currently know (e.g., Boekaerts, Pintrich, & Zeidner, 2000; McCaslin, 2006; Paris & Winograd, 1990; Paris, Byrnes, & Paris, 2001). Emphasizing the role of tests as learning events rather than as performance assessments may alleviate some of the pressure, and accompanying anxiety, that tests create for students when viewed solely as ultimate performance measures. In addition, reorienting to tests as learning tools could facilitate learners' ability to use them as prompts for deliberate practice as described by Ericsson, Krampe, and Tesch-Römer (1993).

Limitations and Future Directions

The present studies are merely a first step in demonstrating the benefits of unsuccessful tests for future learning and, we hope, in encouraging educators and policymakers to consider the benefits of tests as learning events. In spite of the relatively clear results, several aspects of this study limit the breadth of interpretation for educational practice, and more must be done to establish the practical utility of this work. First, the current analyses exclusively focused on posttest scores for items that were answered incorrectly on a pretest. This allowed us to directly examine the impact of attempting yet failing to answer pretest questions, but may have underestimated the overall effects of testing. The extended study condition never had items removed, so these participants may have answered a small number of the posttest questions correctly regardless of their learning from the study materials. Future experiments should be conducted to develop a more precise measure of the effect size for providing pretest questions.

Second, future studies are necessary to ensure that the results generalize across diverse texts and are not tied to some particulars of the current experimental text. Third, we tested only fact-based questions; extending this research to additional types of questions would be useful in future studies. Fourth, the experimental materials were not embedded into participants' educational curricula more broadly, so there may be various differences in the way

participants engaged with the content, although we would anticipate that the cognitive underpinnings of testing and retrieval would remain constant.

Finally, before long-term gains can be made through changes to school practices, bridging studies must be conducted to better understand the relevant teacher-, school-, and district-level engagement with standardized and unit-level tests. The challenge is not only dissemination, but also of determining (a) how to effectively reframe testing as learning during everyday educational practice, and (b) whether modifying the well-established orientation to testing as a performance measure would lead to student gains. In short, the current results suggest that testing may have broad potential to directly enhance learning, whether or not the tests are successful, but, as with any finding in cognitive psychology, further research is necessary to demonstrate that such testing will effectively translate into classroom settings.

Conclusion

When a learner makes an unsuccessful attempt to answer a question, both learners and educators often view the test as a failure, and assume that poor test performance is a signal that learning is not progressing. Thus, compared with presenting information to students, which is not associated with poor performance, tests can seem counterproductive. Tests are rarely thought of as learning events (Kornell & Bjork, 2007), and most educators would probably assume that giving students a test on material before they had learned it would have little impact on student learning beyond providing teachers with insight into their students' knowledge base. In terms of long-term learning, however, unsuccessful tests fall into the same category as a number of other effective learning phenomena (e.g., the spacing effect; see Dempster, 1988): Providing challenges for learners leads to low initial test performance, thereby alienating learners and educators, while simultaneously enhancing long-term learning (Bjork, 1994; Schmidt & Bjork, 1992). The current research suggests that tests can be valuable learning events, even if learners cannot answer test questions correctly, as long as the tested material has educational value and is followed by instruction that provides answers to the tested questions.

References

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives, 10*(18). Retrieved from http://epaa.asu.edu/epaa/v10n18/

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 9, pp. 89–132). New York: Academic Press.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory II* (pp. 396–401). London: Wiley.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80,* 139–144.

Boekaerts, M., Pintrich, P. R., & Zeidner, M. (2000). *Handbook of self-regulation.* San Diego, CA: Academic Press.

Bransford, J. S., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11,* 717–726.

Bransford, J. S., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education, 24,* 61–100.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13,* 273–281.

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

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619–636.

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

Craig, S., Gholson, B., Ventura, M., Graesser, A. C., & the Tutoring Research Group. (2000). Overhearing dialogues and monologues in a virtual tutoring session: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education, 11,* 242–253.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level-reasoning-question effect: The role of dialogue and deep-level-reasoning questions during vicarious learning. *Cognition and Instruction, 24,* 565–591.

Craik, F., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 67–84.

Davis, D., & Loftus, E. F. (2007). Internal and external sources of misinformation in adult witness memory. In M. P. Toglia, J. D. Read, D. F. Ross, & R. C. L. Lindsay (Eds.), *Handbook of eyewitness psychology (Vol. l). Memory for events* (pp. 195–237). Mahwah, NJ: Erlbaum.

Dempster, F. N. (1988). Informing classroom practice: What we know about several task characteristics and their effects on learning. *Contemporary Educational Psychology, 13,* 254–264.

Ericsson, K. A., Krampe, R. Th., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406.

Frase, L. T. (1968). Effect of question location, pacing, and mode upon retention of prose material. *Journal of Educational Psychology, 59,* 244–249.

Frase, L. T. (1970). Boundary conditions for mathemagenic behaviors. *Review of Educational Research, 40,* 337–347.

Garrison, C., Jeung, B., & Inclán-Rodríquez, R. (2006, March). Meeting English language learners' needs under No Child Left Behind. *Trends: Issues in Urban Education, 21,* 1–7.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 40,* 104.

Ghatala, E. S. (1981). The effect of internal generation of information on memory performance. *American Journal of Psychology, 94,* 443–450.

Ghatala, E. S. (1983). When does internal generation facilitate memory for sentences? *American Journal of Psychology, 96,* 75–83.

Gholson, B., & Craig, S. D. (2006). Promoting constructive activities that support vicarious learning during computer-based instruction. *Educational Psychology Review, 18,* 119–139.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399.

Glynn, S. M., Britton, B. K., & Tillman, M. H. (1985). Typographic cues in text: Management of the reader's attention. In D. Jonassen (Ed.), *The technology of text* (Vol. 2., pp. 192–209). Englewood Cliffs, NJ: Educational Technology Publications.

Glynn, S. M., & DiVesta, F. J. (1979). Control of prose processing via instructional and typographical cues. *Journal of Educational Psychology, 71,* 595–603.

Guthrie, E. R. (1952). *The psychology of learning* (rev. ed.). Oxford, England: Harper Bros.

Herman, J. L., & Golan, S. (1993). The effects of standardized testing on teaching and schools. *Educational Measurement: Issues and Practice, 12,* 21–25, 41–42.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567.

Huntley, J., & Davies, I. K. (1976). Preinstructional strategies: The role of pretests, behavioral objectives, overviews and advance organizers. *Review of Educational Research, 46,* 239–265.

Hursh, D. (2007). Exacerbating inequality: The failed promise of the No Child Left Behind Act. *Race Ethnicity and Education, 10,* 295–308.

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

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70,* 626–635.

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

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

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal, 29,* 303–323.

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal, 31,* 338–368.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review, 14,* 219–224.

Kornell, N., Hays, M. J., & Bjork, R. A. (in press). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29,* 210–212.

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (in press). Memorial consequences of answering SAT II questions. *Journal of Educational Psychology: Applied.*

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). Memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14,* 194–199.

Mathison, S., & Freeman, M. (2003). Constraining elementary teachers' work: Dilemmas and paradoxes created by state mandated testing. *Education Policy Analysis Archives, 11*(34). Retrieved from http://epaa.asu.edu/epaa/images/logo.gif

Mayer, R. (1979). Twenty years of research on advance organizers: Assimilation theory is still the best predictor of results. *Instructional Science, 8,* 133–167.

Mayer, R. E. (2008). *Learning and instruction* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice Hall.

McCaslin, M. (2006). Student motivational dynamics in the era of school reform. *The Elementary School Journal, 5,* 479–490.

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385.

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

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

No Child Left Behind Act of 2001, 20 U.S.C. §6319 (2008).

Paris, S. G., Byrnes, J. P., & Paris, A. H. (2001). Constructing theories, identities, and actions of self-regulated learners. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed., pp. 253–287). Mahwah, NJ: Erlbaum.

Paris, S. G., & Winograd, P. (1990). How metacognition can promote learning and instruction. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 15–51). Hillsdale, NJ: Erlbaum.

Pashler, H., Bain, P. M., Bottge, B. A., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (Report No. NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

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

Pressley, M., Tanenbaum, R., McDaniel, M. A., & Wood, E. (1990). What happens when university students try to answer prequestions that accompany textbook material? *Contemporary Educational Psychology, 15,* 27–35.

Reynolds, R. E., & Anderson, R. C. (1982). Influence of questions on the allocation of attention during reading. *Journal of Educational Psychology, 74,* 623–632.

Reynolds, R. E., Standiford, S. N., & Anderson, R. C. (1979). Distribution of reading time when questions are asked about a restricted category of text information. *Journal of Educational Psychology, 71,* 183–190.

Richland, L. E., Bjork, R. A., & Linn, M. C. (2007). Cognition and instruction: Bridging laboratory and classroom settings. In F. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, & T. Perfect (Eds.), *Handbook of applied cognition* (2nd ed., pp. 555–584). Chichester, England: Wiley.

Richland, L. E., Kao, L. S., & Kornell, N. (2008). Can unsuccessful tests enhance learning? In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the Twenty-Eighth Annual Conference of the Cognitive Science Society* (pp. 2338–2343). Mahwah, NJ: Erlbaum.

Rickards, J. P. (1976). Interaction of position and conceptual level of adjunct questions on immediate and delayed retention of text. *Journal of Educational Psychology, 68,* 210–217.

Rickards, J. P., & Hatcher, C. W. (1977–1978). Interspersed meaningful question learning questions as semantic cues for poor comprehenders. *Reading Research Quarterly, 13,* 538–553.

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

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

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155–1159.

Rothkopf, E. Z. (1965). Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumboltz (Ed.), *Learning and the education process* (pp. 193–221). Chicago: Rand McNally.

Rothkopf, E. Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3,* 241–249.

Rothkopf, E. Z. (1982). Adjunct aids and the control of mathemagenic activities during purposeful reading. In W. Otto & S. White (Eds.), *Reading expository material.* New York: Academic Press.

Rothkopf, E. Z., & Bisbicos, E. E. (1967). Selective facilitative effects of interspersed questions on learning from written materials. *Journal of Educational Psychology, 58,* 56–61.

Sacks, O. (1995). *An anthropologist on Mars: Seven paradoxical tales by Oliver Sacks.* New York: Knopf.

Sagerman, N., & Mayer, R. E. (1987). Forward transfer of different reading strategies evoked by adjunct questions in science text. *Journal of Educational Psychology, 79,* 189–191.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217.

Skinner, B. F. (1958, October 24). Teaching machines. *Science, 128,* 969–977.

Snapp, J. C., & Glover, J. A. (1990). Advanced organizers and study questions. *Journal of Educational Research, 83,* 266–271.

Stecher, B. M., Epstein, S., Hamilton, L. S., Marsh, J. A., Robyn, A., McCombs, J. S., et al. (2008). *Pain and gain: Implementing No Child Left Behind in three states, 2004–2006.* Santa Monica, CA: Rand Corporation.

Terrace, H. S. (1963). Discrimination learning with and without "errors." *Journal of the Experimental Analysis of Behavior, 6,* 1–27.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175–184.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 1–60.

Waller, R. (1991). Typography and discourse. In R. Barr, D. R. Pearson, M. L. Kamil, & P. B. Mosenthal (Eds.), *Handbook of reading research* (Vol. 2, pp. 341–380). Mahwah, NJ: Erlbaum.

Watts, G. H., & Anderson, R. C. (1971). Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology, 62,* 387–394.

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465–478.

Yost, M., Avila, L., & Vexler, E. B. (1977). Effects of learning of post-instructional responses to questions of differing degrees of complexity. *Journal of Educational Psychology, 69,* 398–401.

Appendix A

Full List of Test Questions

1. What color is tomato juice to Mr. I?°^
2. How does Mr. I distinguish red and green traffic lights?*
3. It had been shown in the 1960s that there were cells in the primary visual cortex of monkeys (in the area termed V_1) that responded specifically to_____, but not to color.°
4. V4 specializes for responding to _____.
5. _____, in his famous prism experiment in 1666, showed that while light was composite—could be decomposed into, and recomposed by, all the colors of the spectrum.°
6. _____ _____, in 1802, feeling that there was no need to have an infinity of different receptors in the eye, each turned to a different wavelength postulated that 3 types of receptors would be enough.
7. What is total color blindness caused by brain damage called?°
8. Total color blindness caused by brain damage can reveal to us the mechanisms of _____ construction, specifically, here, how the brain "sees" (or makes) color.*
9. Color blindness, as ordinarily understood is something one is born with—a difficulty distinguishing red and green, or other colors, or (extremely rarely) an inability to see any colors at all, due to defects in color responding cells, the _____ of the retina.°
10. When given a large mass of yarns, containing 33 separate colors, how did he sort them?°

Appendix B

Replacement Items for Experiment 4

11. How does Mr. I distinguish flowers?°
12. Why was Mr. I stopped by the police when he decided to go to work again after the accident?°

Appendix C

Rewritten Items for Experiment 5

13. Tomato juice appears _____ to Mr. I.
14. There are cells in the primary visual cortex of monkeys that respond specifically to_____, but not to color.
15. Total color blindness caused by brain damage is called _____.
16. _____ showed that white light was composite.
17. Color blindness, as ordinarily understood, is something one is _____, rather than acquired later.
18. When given a large mass of yarns, containing 33 separate colors, Mr. I separated them by _____.
19. Mr. I distinguishes flowers by _____.
20. Mr. I was _____ when he decided to go to work again after the accident.

*Replaced for Experiments 3–5.

°Rewritten into standardized form for Experiment 5.

^Question numbering is for clarity and was not fixed in this order.

Received June 30, 2008
Revision received April 24, 2009
Accepted April 27, 2009 ■