Can Unsuccessful Tests Enhance Learning?

Lindsey E. Richland (l.e.richland@uci.edu)
Department of Education, University of California, Irvine
2001 Berkeley Place, Irvine, CA 92697-5500 USA

Liche Sean Kao (lskao@uci.edu)
Department of Education, University of California, Irvine
2001 Berkeley Place, Irvine, CA 92697-5500 USA

Nate Kornell (nkornell@ucla.edu)
Department of Psychology, University of California, Los Angeles
1285 Franz Hall, Los Angeles, CA 90095 US

Abstract

The testing effect is the phenomenon that testing enhances memory for previously studied content. Memory is particularly enhanced for items successfully retrieved during testing. Three experiments investigated the effects of testing before studying - a time when participants were unlikely to successfully retrieve content. Participants read excerpts from an essay on vision. They were either asked about embedded concepts before reading the passage (test condition) or they read the passage for a longer time (read condition). In both conditions the tested concepts were highlighted (presented in bold letters or italics) to distinguish the effects of testing from attention direction. Although participants failed on initial tests, memory performance on a final posttest was better in the tested condition in all experiments. Retrieving the correct answer from memory does not appear to be the only reason for the testing effect—simply being asked seems to enhance future learning.

Introduction

The purpose of this title is to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging State academic achievement standards and state academic assessments. This purpose can be accomplished by —

> *ensuring that high-quality academic assessments, accountability systems, teacher preparation and training, curriculum, and instructional materials are aligned with challenging State academic standards so that students, teachers, parents, and administrators can* ***measure progress against common expectations for student academic achievement*** *(No Child Left Behind, 2001, p.1, bolding added)*

The No Child Left Behind (NCLB) act was the largest political restructuring of the U.S. educational system in history. In NCLB, in political rhetoric, and in the modern assessment-heavy climate both have created, testing is often viewed purely as an instrument of performance assessment. Test results are very commonly presented as reasonably stable measurements of children's knowledge and proficiency. Policy-makers argue that children's learning will be indirectly improved by such measurements via teacher and administrators' adjustment of their curricula and pedagogical practices. This view, as well as NCLB and the political rhetoric, are valid but ignore an additional, important empirical finding: testing enhances learning. One would imagine that policy-makers would pounce on this well replicated psychological result if they were aware of it.

A survey of naive undergraduates reflects that this view of testing is a common one in the United States. Kornell & Bjork (2007) asked undergraduates whether they tested themselves when they were studying, and if so, why. While most students did report testing themselves (91%), most of those 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" (p. 222).

Testing as a Learning Event

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 (Roediger & Karpicke, 2006). Research examining the effect of testing following learning suggests that such tests not only provide a measure of learners' knowledge; but rather also become a learning event in their own right. Like Heisenberg uncertainty principle of physics, measurement of learners' proficiency actually alters their knowledge representations.

Caveat: Failed tests lack value

There has been a caveat to these hopeful analyses of testing as an instrument serving larger instructional goals, however. The benefits of testing after learning are most pronounced for test items that were answered correctly (Butler & Roediger, 2007; Karpicke & Roediger, 2007; Leeming, 2002; McDaniel et al, 2007; Roediger & Karpicke, 2006). Generally, items not retrieved correctly at time of test see minimal if any benefits for testing when compared to being allowed additional study time (for exceptions see Izawa, 1970; Kornell, Hays, & Bjork,

2007). Thus, when testing is viewed as the ultimate step in knowledge acquisition, any missed test items are unlikely to see knowledge gains. Providing detailed, personalized feedback after a test can ameliorate some of these challenges (Kang, McDermott, Roediger, 2007), but this is burdensome and often not feasible. Thus when testing the lowest performing students, as is NCLB's first and foremost priority, the benefits of testing for learning may be minimal if failed tests lack value.

Can failed tests improve future learning?

The current paper posits that this caveat is not universal, and should not be taken as a rationale for critiquing the value of testing as a learning event. Rather, three experiments were conducted to evaluate the impact of re-structuring 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 quite high, but we were able to assess whether trying but failing at such tests actually improved learners' retention of the instruction provided subsequently.

Many studies have demonstrated benefits of pre-training activities such as advanced organizers (see Huntley & Davies, 1976; Mayer, 1979), outlines (e.g. Snapp & Glover, 1990), and even pre-test questions (Huntley & Davies, 1976; Pressley et at, 1990; Shapiro, 2000). However, these studies have not sought to fully distinguish the effects of cognitively attempting to answer unknown test questions from these items' impact as guides for learners' attention during subsequent learning.

We report three experiments in which participants studied a science text about vision. Participants were either tested prior to learning, or they were given additional time to study.

Experiment 1

We predicted that testing before study would enhance future recall overall, in spite of learners' failure to successfully answer test questions. We also predicted that tested items would be recalled more than untested items.

Method

Participants

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

Materials

Study materials were selected from Sacks (1995). A two-page text was developed by combining passages from a narrative about cerebral achromatopsia (colorblindness caused by brain damage). This text was selected due to its rich scientific content in combination with an engaging narrative. 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.

Within the reading packet, ten sentences were identified as testable items. Test materials were constructed from the ten testable sentences. Two counterbalanced Time 1 tests were constructed such that each contained questions about five of the italicized sentences. Questions were written as fill-in-the-blank or short free response items (e.g., "What is total colorblindness caused by brain damage called?" and "How does Mr. I distinguish red and green traffic lights?).

A final test included all ten of the test items in randomized order. Thus for all participants in the test conditions, five of these questions had been tested previously (tested during Time 1) and five had not been previously tested (untested). Items from the two Time 1 test versions were always interspersed.

Procedure

The experiment was conduced in a group setting. Participants were randomly assigned to an *Extended Study* condition (N = 27) or *Test and Study* condition (N = 36).

Test and Study Condition. Participants in the Test condition were first administered one of two counterbalanced tests, during Time 1, and allowed two minutes to answer the questions. They were instructed to provide an answer to all five questions, regardless of whether they knew the answer. After completion, the tests were collected. Participants were then given the text passage and told to study it for eight minutes.

Extended Study Condition. Participants in the extended study condition were first given the study passage. They were given 10 minutes to study the materials—the same total time that participants in the test condition spent in testing and study of the material.

Time 2 Test. After the timed study periods were complete in all conditions, text passages were collected. Participants were then immediately administered the Time 2 Test which consisted of 10 questions. This was untimed to ensure that time pressure did not impact performance.

Data

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 at Time 1 were removed from the following analyses of Time 2 test scores by individual.

An independent-samples t-test first examined the effects of testing by comparing mean scores for tested items in the test and study condition with the overall mean score in the extended study condition. As shown in Figure 1, the difference was significant, t (61) = 4.25, $p < .001$, revealing a benefit for testing over extra time spent studying the same material.

In order to better understand the impact of testing before studying, the test and study condition means for tested items was compared with the mean for untested items. A paired-samples t-test t (35) = 5.03, $p < .001$,

again revealed a benefit of testing over reading only, in spite of the fact that the participants generally failed the initial test opportunity. The benefit of testing, however, did not spread to untested items. Importantly, however, neither did it hurt, since a comparison between the extended study mean and the untested items in the test and study condition revealed no differences, t (61) = 1.3, $p = .20$.

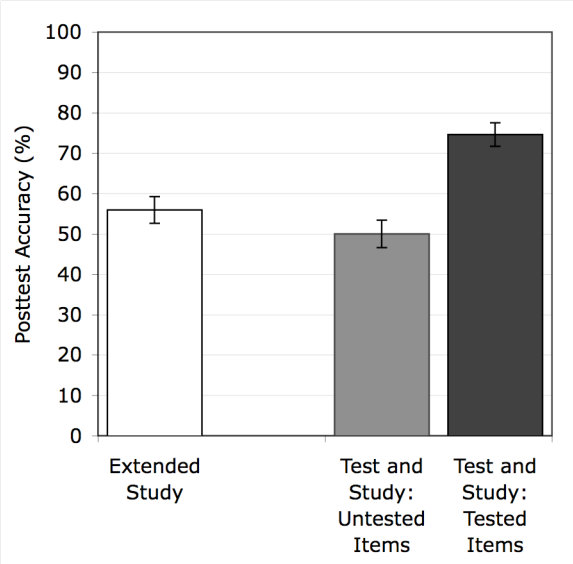

Figure 1. Experiment 1: performance on a final test across conditions when studying an unmarked text.

Discussion

Overall, these results revealed that failed tests *can* impact learning for educational content. Although participants largely failed on the initial test, testing led to increased retention of studied content.

The explanation for the improvement 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 a unique benefit above and beyond the impact of attention. Experiment 2 used the same procedure but all testable sentences were italicized in the text. Thus participants' attention to key sentences should be equivalent, and differences between conditions attributable to impact of the test itself.

Experiment 2

We predicted that testing before study would enhance future recall more than studying italicized key sentences in the instructional text.

Method

Participants

Participants in this study were 61 undergraduates who were given extra credit course points. Data from two participants were excluded from analyses due to failure to respond to final test questions.

Materials

Study materials were the same text and testable sentences as used in Experiment 1.

The key difference was that within the reading packets, the ten testable sentences were italicized. Like many science textbooks that highlight the key elements of a chapter, this was considered an educationally relevant way to ensure that all participants were equally alerted to what was deemed to be important information. 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, p18).

Test materials were the same as in Experiment 1.

Procedure

The procedure was exactly the same as for Experiment 1. Participants were randomly assigned to the extended study condition (N = 26) or the test and study condition (N = 33).

Data

In the Test and Study condition, on the initial test that preceded the presentation of the passage, participants answered 22**%** of the questions correctly. Any items answered correctly at Time 1 were removed from the following analyses of Time 2 test scores by individual.

An independent-samples t-test first examined the effects of testing by comparing mean scores for tested items in the test and study condition with the overall mean score in the extended study condition. As shown in Figure 1, the difference was significant, t (57) = 2.3, $p < .05$, revealing a benefit for testing over extra time spent studying the same material.

In order to better understand the impact of testing before studying, the test and study condition means for tested items was compared with the mean for untested items. A paired-samples t-test t (33) = 3.27, $p < .01$, again revealed a benefit of testing over reading only, in spite of the fact that the participants generally failed the initial test opportunity. The benefit of testing, however, did not spread to untested items. Importantly, however, neither did it hurt, since a comparison between the extended study mean and the untested items in the test and study condition revealed no differences, t (57) = .15, $p = .88$.

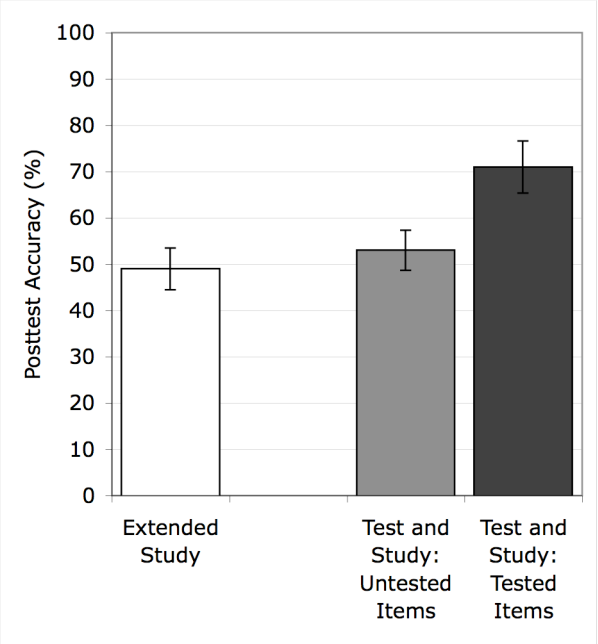

Figure 2. Experiment 2: performance on a final test across conditions when studying text with italicized key sentences.

Experiment 3 replicated the same methodology with two changes. The final test was delayed to a week after the initial learning phrase to investigate the robustness of memory benefits of testing.

Second, in order to rule out the possibility that learners were unfamiliar with the meaning of italics within text and thus this was not a sufficiently equivalent control, Experiment 2 used the same procedure with bolded keywords rather than italicized sentences. Textbooks more often contain bolded key words than italicized sentences, so we anticipated that this might act as a stronger attention cue. Experiment 2 thus examined the impact of testing when compared with studying text in which the key test items were bolded.

Bolding was manipulated in a within-subjects design to better distinguish between the effects of bolding and testing. Testing versus extended study remained a between-subjects manipulation.

Experiment 3

We predicted that final test performance at a delay would be higher for participants in the test condition, even if they were unsuccessful at answering any questions, than for the extended study condition. We also predicted that bolding would aid retention, but that testing would still be overall more advantageous than bolding.

Method

Participants

Participants in this study were 158 undergraduates who were given extra credit course points.

Materials

Study materials and procedure were exactly the same as those used in Experiment 2 with one variation in the bolding. Rather than bolding all ten key words that were tested on the final test, as in Experiment 2, only five items were bolded. In the Test and Study condition, the bolded items were the same as those on the Time 1 test. Thus for any participant, five items on the posttest had been given additional emphasis during initial study (testing and bolding or bolding only), and five items had not. Tested/ bolded items were counterbalanced across participants. This allowed us to interpret effects of bolding compared with no bolding and testing versus no testing or bolding in a within-subjects design.

Procedure

The procedure was the same as for Experiments 1 and 2, with the exception of the timing for the final test. Also in order to control the delay timing, participants were tested individually. The first session followed the identical procedure as for Experiments 1 and 2. After completion of the first session, participants were asked to return one week later at the same time of day. At that time, participants were administered the Time 2 Test.

Participants were randomly assigned to an *Extended Study* condition (N = 79) or *Test and Study* condition (N = 79).

Data

In the Test and Study condition, on the initial test that preceded the presentation of the passage, participants answered 10% of the questions correctly. Any items answered correctly at Time 1 were removed from the following analyses of Time 2 test scores by individual.

Analyses were conducted slightly differently from Experiments 1 and 2 since the posttest performance for the extended study condition now could be separated into bolded and unbolded items. A repeated measures ANOVA was conducted with testing/ bolding as a within-subjects variable to compare performance on tested/ bolded items and untested/ unbolded items. Condition was included as a two-level between-subjects variable.

The analysis revealed a main effect of condition, such that testing led to overall higher performance than extended study, F (1, 156) = 4.9, $p < .05$, $h_p^2 = .03$. Additionally, as shown in Figure 3, there was a main effect of testing/bolding, $F(1, 156) = 14.0$, $p < .001$, $h_p^2 = .08$. There was not an interaction between condition and testing/ bolding $F(1, 156) = 2.3$, $p = .13$, $h_p^2 = .02$. When examined separately, there was a significant difference between conditions on the tested/bolded items $F(1, 156) = 7.9$, $p < .01$, $h_p^2 = .05$, though not on the untested/ unbolded items $F(1, 156) = .34$, $p = .56$, $h_p^2 = .00$.

A paired-samples t-test examining only the testing condition revealed that retention was higher for tested than untested items t (79) = 3.27, $p < .01$.

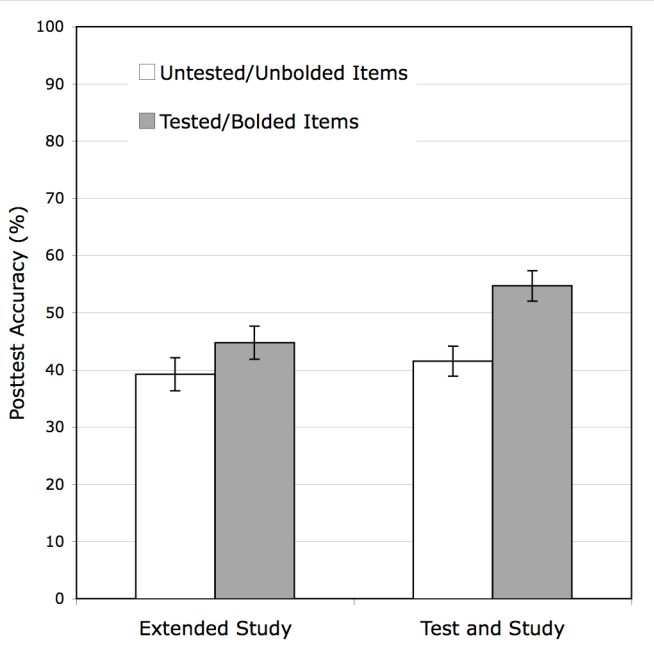

Figure 3. Experiment 3: One-week delayed performance on a final test across conditions when keyword bolding was manipulated within-subjects.

Discussion

These results extend findings from Experiments 1 and 2, to show that failed tests can impact learning for educational content even after a delay. Testing items before learning was a more potent learning opportunity than bolding keywords in the text. 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.

General Discussion

In sum, the three experiments reported herein support the argument that testing should be considered a potent learning opportunity, rather than simply an assessment measure. The current results extend previous findings by showing that even if tests are not answered successfully, they have the potential to improve future learning. This is true on both immediate and delayed performance measures, suggesting that rethinking testing as learning could have lasting implications on learners' content acquisition. Moreover, because most science texts bold key words/topics, these findings require little translation for use in educational learning contexts.

At present, teachers in many states spend two weeks or more of instruction time on standardized testing throughout the school year, which might increase as more districts align with NCLB and assessment pressures. Teachers and administrators alike often describe these days as outside of instruction and as reducing an already impacted curriculum schedule. Failed tests get used as markers of lack of student progress and as a particularly egregious use of these needy students' time.

The current paper lays the foundation for arguing instead 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 to provide future learning opportunities after a failed test, however, since a test that is not followed by instruction or feedback is likely to impact learning less. While feedback on tests aids learning (Kang, McDermott & Roediger, 2007), our data suggest that instruction following testing need not be individualized. Rather, instruction that draws attention to key content may build on the cognitive acts performed when attempting to answer related test questions.

As argued by Bransford & Schwartz (1999), measuring the impact of instruction on future learning is an important aspect of judging learning opportunities. The current experiments demonstrate that the act of attempting to answer test questions, even if one is ultimately unsuccessful, can serve to prepare learners for future knowledge acquisition. In sum, this paper seeks to provide evidence that will encourage educators and policy-makers to expand theories of testing to include the benefits of tests, even if failed, as learning events.

Acknowledgments

The authors thank the Office of Naval Research, #N000140810186 for support. We also thank Seungyeon Lee and Keara Osborne for invaluable assistance.

References

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4-5), 514-527.

Bransford, J.S., Schwartz, D. L. (1999). Rethinking Transfer: A Simple Proposal with Multiple Implications. *Review of Research in Education*, *24*, pp. 61-100.

Huntley, J., Davies, I. K. (1976). Preinstructional Strategies: The Role of Pretests, Behavioral Objectives, Overviews and Advance Organizers. *Review of Educational Research, 46(2)*, pp. 239-265.

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

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*(4-5), 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*(2), 151-162.

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. (2007, November). Failed tests can enhance learning. Poster presented at the 48th annual meeting of the Psychonomic Society, Long Beach, CA.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology, 29*(3), 210-212.

Mayer, R. (1979). Twenty years of research on advance organizers: Assimilation theory is still the best predictor of results *Instructional Science, 8 (2)*

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*, 494-513.

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*(2), 200-206.

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. Vol 15(1),* pp. 27-35.

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

Sacks, O. (1995). *An Anthropologist On Mars: Seven Paradoxical Tales by Oliver Sacks*.

Shapiro, A. M. (2000). The effect of interactive overviews on the development of conceptual structure in novices learning from hypermedia. *Journal of Educational Multimedia and Hypermedia*, *9*, pp. 57-78.

Snapp & Glover (1990). Advanced Organizers and Study Questions. *Journal of Educational Research*, *83*, pp 266-71.