THE QUARTERLY JOURNAL OF EXPERIMENTAL PSYCHOLOGY
2012, 65 (5), 962–975

Psychology Press
Taylor & Francis Group

Decomposing the interaction between retention interval and study/test practice: The role of retrievability

Yoonhee Jang[1], John T. Wixted[1], Diane Pecher[2], René Zeelenberg[2], and David E. Huber[1]

[1]Department of Psychology, University of California, San Diego, CA, USA
[2]Department of Psychology, Erasmus University Rotterdam, Rotterdam, The Netherlands

Even without feedback, test practice enhances delayed performance compared to study practice, but the size of the effect is variable across studies. We investigated the benefit of testing, separating initially retrievable items from initially nonretrievable items. In two experiments, an initial test determined item retrievability. Retrievable or nonretrievable items were subsequently presented for repeated study or test practice. Collapsing across items, in Experiment 1, we obtained the typical cross-over interaction between retention interval and practice type. For retrievable items, however, the cross-over interaction was quantitatively different, with a small study benefit for an immediate test and a larger testing benefit after a delay. For nonretrievable items, there was a large study benefit for an immediate test, but one week later there was no difference between the study and test practice conditions. In Experiment 2, initially nonretrievable items were given additional study followed by either an immediate test or even more additional study, and one week later performance did not differ between the two conditions. These results indicate that the effect size of study/test practice is due to the relative contribution of retrievable and nonretrievable items.

Keywords: Testing effect; Retrievability; Forgetting.

An effective technique for retaining information over the long term is to engage in test practice rather than additional study. A large number of experiments have reported that participants remember more material on a final test when they were given an intervening test, and this phenomenon is known as the *testing effect* (e.g., Gates, 1917; Glover, 1989; Spitzer, 1939; for a review, see Bjork, 1988; Dempster, 1996; Roediger & Karpicke, 2006a). The testing effect is quite robust. It has been found with different types of material (e.g., words or passages) in a variety of educational situations (for details, see Roediger & Karpicke, 2006a).

We begin by discussing two remarkable aspects of the testing effect. One aspect is that the testing effect is found even in the absence of feedback (e.g., Allen, Mahler, & Estes, 1969; Roediger & Marsh, 2005; Runquist, 1983; Wheeler & Roediger, 1992). It is not surprising that an intervening test with correct-

Correspondence should be addressed to Yoonhee Jang, Department of Psychology, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093–0109, USA. E-mail: yhjang@ucsd.edu

This research was supported by National Institute of Mental Health Grant RMH081084A to David Huber.

http://www.psypress.com/qjep
http://dx.doi.org/10.1080/17470218.2011.638079

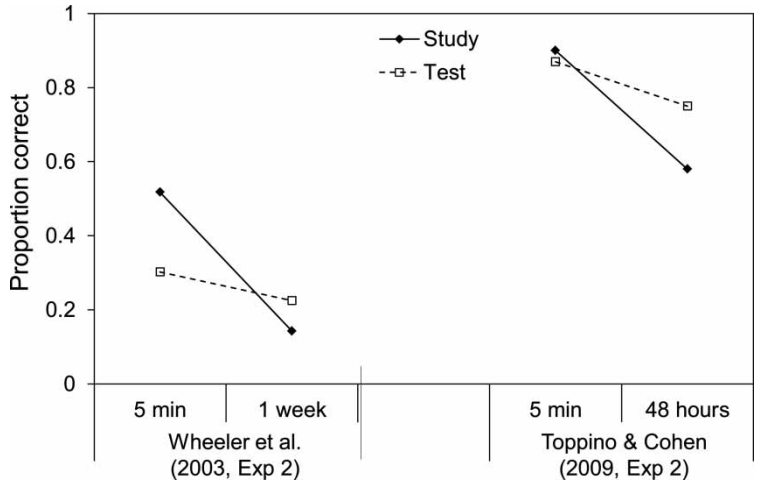

Figure 1. *The results of Wheeler et al. (2003, Experiment 2, Figure 2, p. 576), and Toppino and Cohen (2009, Experiment 2, Figure 2, p. 255): left and right sides, respectively. These graphs were created based on visual approximation from the previously reported figures. Note that the* x*-axis differs in scale across the two experiments—that is, for the delayed final test, 1 week (left) and 48 hours (right).*

answer feedback often enhances future test performance (e.g., Butler, Karpicke, & Roediger, 2007; Cull, 2000; McDaniel & Fisher, 1991; Pashler, Cepeda, Wixted, & Rohrer, 2005). Testing with feedback provides not only an opportunity for retrieval practice if the correct answer is recalled, but also an opportunity for additional study even if the correct answer is not recalled: Regardless of performance, feedback provides an opportunity for additional encoding. The other remarkable aspect of the testing effect is that in the absence of feedback, performance of the tested participants (or items) is not only better than that of the nontested participants (or items), but in certain circumstances, it is better than that of a condition that includes additional study rather than testing. A powerful demonstration of this compares intervening study versus intervening test, without feedback, as they affect an immediate final test versus a delayed final test (hereafter, we refer to an intervening study or intervening test between initial study and final test as simply *study* or *test*). In these situations, a cross-over interaction is often found: benefits of study for an immediate final test and benefits of test for a delayed final test (e.g., Roediger & Karpicke, 2006b; Thompson, Wenger, & Bartling, 1978; Wheeler, Ewers, & Buonanno, 2003; for a review, see Roediger & Karpicke, 2006a).

To further investigate this issue, we took a closer look at the cross-over interaction reported by Wheeler et al. (2003, Experiment 2), which is shown in the left side of Figure 1. For an immediate final test occurring 5 minutes after practice, the study group had better recall than the test group, but there was a significant reversal for the 1-week delayed final test. Participants in the study group were provided with all 40 words for each of the four study phases between initial learning and the final test (in total, 40×5 presentations). In contrast, participants in the test group studied all 40 words only once (initial learning) and then were tested on all 40 for each of the four test phases, but on average they were only able to recall 11 of the words on each test. Thus, there was less opportunity for additional encoding of the items in the test condition—for example, $40 + (11 \times 4)$ presentations. For the test group, performance was consistent across the four tests and also across the final tests. Little forgetting occurred during the 1-week retention interval: Participants in the test group recalled approximately 12 and 9 words on the immediate and delayed final tests, respectively. This consistency suggests that the initially retrievable items were more or less permanently stamped in by test practice: The advantage of test over study is

driven mainly by the items that are *initially retrieved*, or *retrievable*. In other words, if retrieval practice is effective when the item is actually correctly retrieved, then only an item that is retrieved on the first test will benefit from subsequent tests.

A widely accepted explanation of the testing effect is the retrieval practice hypothesis (e.g., Bjork, 1988; Dempster, 1996; Glover, 1989; Roediger, 2000; Spitzer, 1939). Providing support for this hypothesis, Wheeler and Roediger (1992) found that recall at a 1-week delay steadily increased as a function of the number of additional tests without feedback. Validating a key prediction of the retrieval practice hypothesis, conditional analyses revealed that the benefit of test practice without feedback increases with the number of times that the item is correctly recalled during test practice (e.g., Karpicke & Roediger, 2007). However, these analyses did not assess retrievability followed only by study or test practice; instead, participants experienced a mixture of study and test practice throughout learning. It seems fairly obvious that the long-term benefit of test over study practice must be due to the retrievable items, although, to date, no work has compared study versus test practice in the situation that separately considers the role of retrievable versus nonretrievable items. Of greater interest is the question of how retrievability relates to the short-term benefit of study over test practice.

A clue comes from Toppino and Cohen (2009, Experiment 2). Unlike many other studies, their experiment used a high proportion of retrievable items. They provided participants with multiple initial learning opportunities (eight times) so that they could achieve substantially high recall rates (at least 85% correct) on the initial test immediately after initial learning. Therefore, the difference in the number of opportunities for additional encoding between the study and test conditions was smaller than the difference used in many other studies. As shown in the right side of Figure 1, they found no significant benefit of study for an immediate test occurring 5 minutes after practice but a substantial benefit of testing for a delayed test occurring 2 days after practice. Thus, as study does not seem to benefit highly retrievable items, the typical finding that study produces an advantage as compared to testing for an immediate test might be caused almost exclusively by the items that are *not initially retrieved*, or *nonretrievable*.

When retrieval practice via testing is followed by correct-answer feedback, however, a different pattern emerges (Pashler et al., 2005): Nonretrievable items benefit more from testing than retrievable items. Pashler et al. compared a group that received correct-answer feedback to groups that received no feedback or only acknowledgment of their accuracy following each response. When all items were analysed, the correct-answer feedback group showed the best performance both for an immediate test (Test 2, Day 1) and for a 1-week delayed test. Pashler et al. further examined performance conditionalized on performance for the initial test (Test 1, Day 1). When performance for the initial test was correct (i.e., items that were initially retrieved), correct-answer feedback produced no significant benefit, regardless of retention interval. In contrast, when performance for the initial test was incorrect (i.e., items that were not initially retrieved), correct-answer feedback produced substantial benefits for both the immediate and delayed tests.

At first glance, these findings appear contradictory because the first two studies (Toppino & Cohen, 2009; Wheeler et al., 2003) imply that retrievable items are important for testing effects whereas the results of Pashler et al. (2005) suggest that nonretrievable items are important for testing effects. However, this puzzle can be solved when we carefully consider the procedure of Pashler et al. Of critical importance, their experimental design did not include a pure study condition (i.e., study practice without testing) but instead examined the role of correct-answer feedback. For initially retrievable items, it is likely that these items were successfully retrieved during subsequent test practice, and so it is not surprising that the items were unaffected by the type of feedback. In contrast, for initially nonretrievable items, correct-answer feedback provides an opportunity for additional encoding. Thus, the apparent testing effect for nonretrievable items may have had more to do with the additional opportunity

to study these items rather than practice recalling the items. Therefore, the results of Pashler et al. can be consistent with the idea that initially retrievable items primarily benefit from test practice whereas initially nonretrievable items primarily benefit from additional study.

To examine the role of retrievability, our study used a full breakdown of item retrievability and appropriate control conditions for each type of item. We investigated whether the cross-over interaction between retention interval and practice type exists for both retrievable and nonretrievable items. Rather than manipulating retrievability through different amounts of study (e.g., Toppino & Cohen, 2009, across two different experiments), we measured naturally occurring differences in item retrievability with a pretest (or initial test) after initial learning.

EXPERIMENT 1

Experiment 1 used a standard manipulation of practice type and retention interval with three important additions. First, we included an initial test after initial learning to determine item retrievability. Second, we included a control condition in which items were neither studied nor tested during intervening practice. Because we did not use multiple study and test phases during initial learning (such as used by Izawa, 1966; Karpicke & Roediger, 2007; 2008; Tulving, 1967), our control provided a baseline measure of performance without any opportunity for additional study of test practice beyond the initial test that established retrievability. Third, participants in the immediate final test condition also returned 1 week later to take another final test. Inclusion of this condition allowed us to ascertain the long-term benefit of an immediate test following additional study. This is analogous to a student who crams for an exam (e.g., massed study of material just before an exam), who is then tested at a later date (e.g., an encounter with the material in a surprise quiz after the exam)—does the act of taking the exam shortly after massed study serve as test practice that promotes long-term retention?

Method

Participants

One hundred and forty-eight undergraduate students at the University of California, San Diego were recruited and received credit for psychology courses in return for their participation. Both the immediate final test group and the 1-week delayed final test group consisted of 74 randomly assigned participants. Both groups were further divided into participants that received only retrievable items (36 for immediate and 35 for delayed) and participants that received only nonretrievable items (38 for immediate and 39 for delayed) during the intervening phase.

Materials

The stimuli were 90 moderately high-frequency (an average frequency of 60 per million: Kucera & Francis, 1967), singular noun word pairs, from 4 to 7 letters in length. The two words of a pair were semantically and phonologically unrelated.

Design

A $2 \times 2 \times 3$ mixed-factorial design was used. Retention interval (immediate versus 1 week) and item type (retrievable versus nonretrievable) were between-subjects factors whereas practice type (study, test, and control) was a within-subjects factor.

Procedure

Figure 2 shows the experimental procedure. This experiment consisted of two sessions separated by 1 week. At the beginning of Session 1, participants were told that they would be asked to remember a list of word pairs. They were then presented with 90 pairs, of which one third were randomly assigned to each of the study, test, and control conditions, and this assignment was unknown to the participant. The initial learning consisted of three presentation blocks (items were presented three times), and each item was presented once per block for 4 s. A pilot study determined that use of three presentations would produce approximately 50% recall, which was needed for equal numbers of retrievable and nonretrievable items. Presentation order was randomized anew per block and per participant. After initial learning,

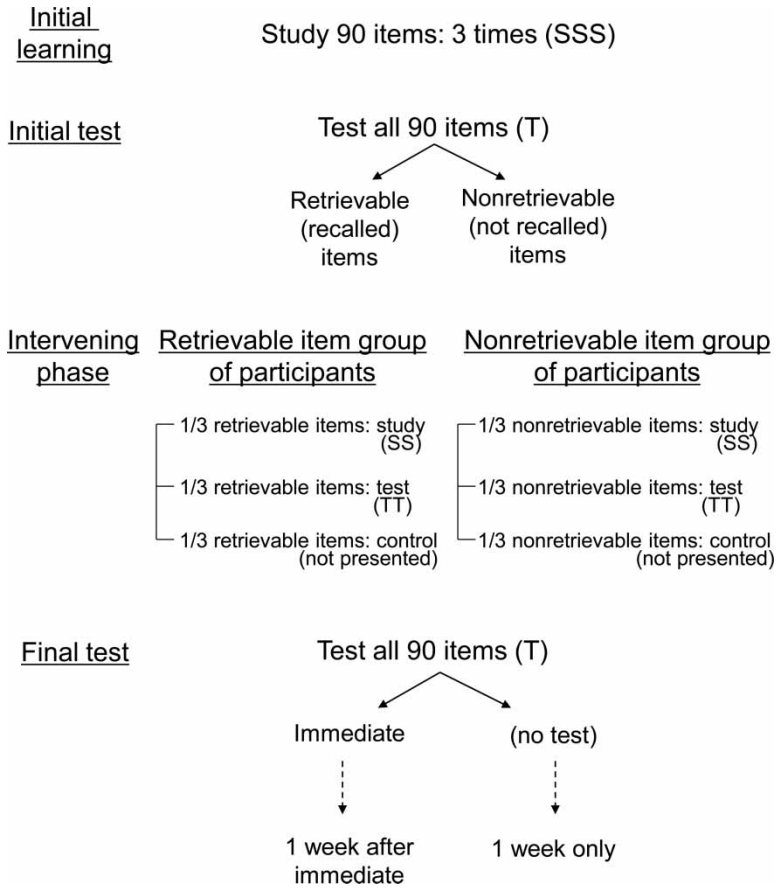

Figure 2. *The procedure of Experiment 1. S = study; T = test.*

participants took an initial cued recall test on all 90 items. Each typed response was self-paced, and no feedback was given as to the correctness of answers. Based on this initial test, the items were divided into retrievable and nonretrievable, and participants only received one type of item during the subsequent intervening phase, depending on whether the participant was assigned to the retrievable item group or the nonretrievable item group.

The intervening phase consisted of two study (S) blocks and two test (T) blocks (items were either presented twice or tested twice). To counterbalance, block order was either STTS or TSST. During each study block, each item was presented for 5 s, and during each test block, participants were given 5 s to type in a response to a cue word. The item order within a study or test block was randomized anew for each participant. The items assigned to the control condition were not presented or tested during the intervening phase.

Finally, participants assigned to the immediate final test group took a final test on all 90 items at the end of Session 1 (i.e., the immediate final test) and returned 1 week later to take another final test during Session 2 (i.e., the 1-week-after-immediate final test). Participants assigned to the 1-week delayed final test group did not receive an immediate final test during Session 1 and received only a final test 1 week later during Session 2 (i.e., the 1-week-only final test). All final tests were self-paced cued recall, and all items were presented in random order for each participant.

Results

Statistical significance was determined with an alpha of .05. First, we consider performance on the initial and intervening tests to check for equality between conditions: The complete data are reported in Appendix A. Performance on the initial test was .49 on average ($SE = .046 \sim .050$), and there were no significant effects for any of the independent variables, suggesting that retrievability was equally distributed across the conditions, as expected considering that all conditions were treated identically until the initial test. Also, performance on the intervening tests was not significantly different between immediate and 1-week-only final test conditions, regardless of item type. Next, we present the results of the final tests.

Interactions between practice type and retention interval

Figure 3 shows the mean proportion correct recall when the data were combined over retrievable and nonretrievable items. A 2 × 3 analysis of variance (ANOVA) was conducted with retention interval (immediate versus 1-week-only final test) and practice type (study, test, and control). Performance was better when retention was measured immediately than after a 1-week delay, $F(1, 146) = 113.34$, $\eta_p^2 = .44$. There was a main effect of practice type, $F(2, 292) = 28.77$, $\eta_p^2 = .16$, and an interaction between practice type and retention interval, $F(2, 292) = 26.32$, $\eta_p^2 = .15$. Specifically, study produced better performance than test for the immediate final test, $t(73) = 8.22$, $d = 1.73$, but performance did not differ between test and control, $t(73) < 1$, whereas test produced better performance than study for the 1-week-only final test, $t(73) = 2.13$, $d = 0.16$, but performance did not differ between study and control, $t(73) = 1.63$, $p = .11$. These results replicate the testing effect, revealing a cross-over interaction between study/test and retention interval.

Before considering the results broken down by retrievability, we consider whether there were basic performance differences between the group of participants who practised retrievable items versus the group who practised nonretrievable items. One way to check for this is with the control conditions, which were common to all groups (for all groups, there were some retrievable and some nonretrievable control items that were not practised). We found no significant differences between the groups: The complete data are reported in Appendix B. The failure to find any differences between participant groups in terms of these control conditions suggests that practice of retrievable or nonretrievable items did not differentially affect motivation or otherwise change performance. In the analyses that follow, control items were only analysed as indicated by the participant group (e.g., for participants who

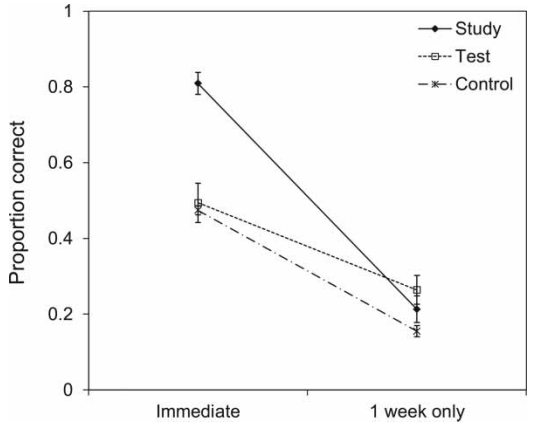

Figure 3. *Proportion correct recall on the final test as a function of retention interval and practice type of Experiment 1: All items. Error bars depict ±1 standard error of the mean.*

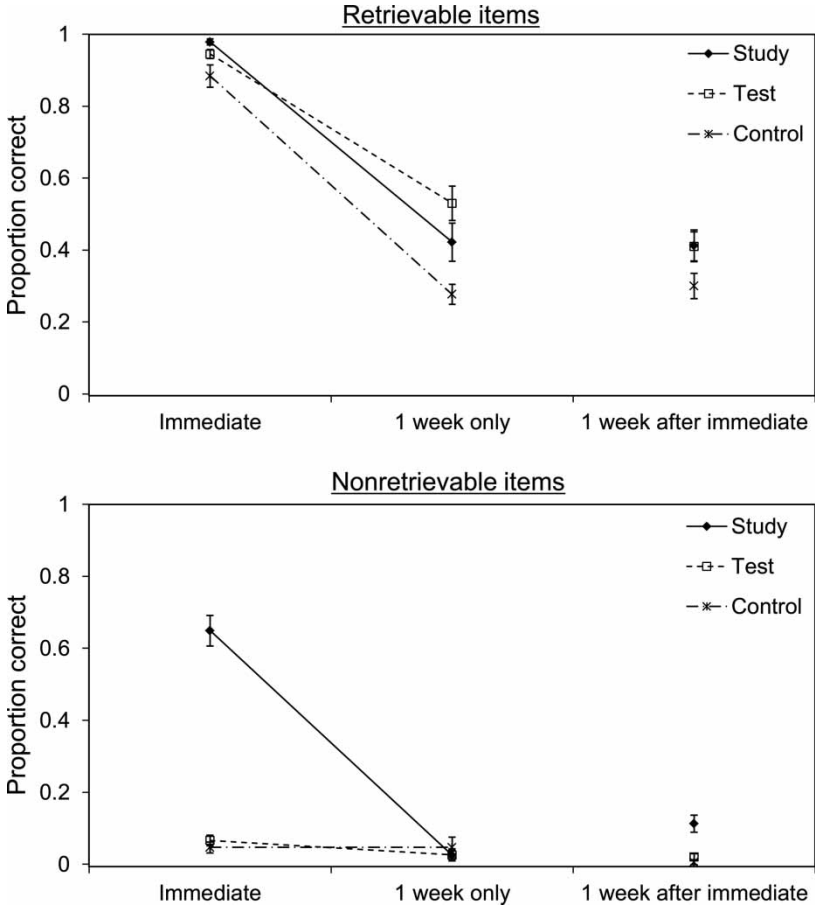

Figure 4. *Proportion correct recall on the final test as a function of final test type and practice type of Experiment 1: Top shows retrievable items, and bottom shows nonretrievable items. Error bars depict ±1 standard error of the mean.*

practised retrievable items, only the retrievable control items were analysed). In this manner, we were able to use item type (retrievable versus nonretrievable) as a fully between-subjects factor and practice type as a fully within-subjects factor. Figure 4 shows the results broken down by item type, and we next consider separately the results for retrievable items and then the results for nonretrievable items.

The results for initially retrievable items are presented in the top panel of Figure 4. A 2 × 3 ANOVA with retention interval (immediate versus 1-week-only final test) and practice type (study, test, and control) revealed main effects of retention interval and practice type, $F(1, 69) = 189.07$, $\eta_p^2 = .73$; and $F(2, 138) = 22.57$, $\eta_p^2 = .25$, respectively, as well as an interaction, $F(2, 138) = 8.37$, $\eta_p^2 = .11$. Specifically, study produced better performance than test for the immediate final test, $t(35) = 2.28$, $d = 0.51$, but performance was not different between test and control, $t(35) = 1.75$, $p = .09$. Furthermore, test produced better performance than study for the 1-week-only final test, $t(34) = 2.42$, $d = 0.36$, and study also was better than control, $t(34) = 3.71$, $d = 0.61$. Thus, for the final test of retrievable items, there was a small benefit of study over test at the immediate test whereas there was a substantial advantage of test over study at the 1-week-only final test. This benefit of test over study accounted almost completely for the test advantage in the

collapsed data at the 1-week test. Note that the small (but significant) benefit of study at the immediate test is inconsistent with the result of Toppino and Cohen (2009, Experiment 2) who obtained no difference between the study and test conditions for an immediate test of highly retrievable items (as shown in the right side of Figure 1). However, their failure to find a difference may have been a Type II error considering that the numerical advantage of study over test was about .03 for both experiments. Besides the robust testing effect at the 1-week-only final test, our results demonstrate that there are long-term benefits of study for retrievable items, as shown by significantly better performance in the study condition than in the control condition.

The results for initially nonretrievable items are presented in the bottom panel of Figure 4. A 2 × 3 ANOVA revealed main effects of retention interval and practice type, $F(1, 75) = 81.38$, $\eta_p^2 = .52$; and $F(2, 150) = 129.18$, $\eta_p^2 = .63$, respectively, as well as an interaction, $F(2, 150) = 139.78$, $\eta_p^2 = .65$. Specifically, study produced better performance than test for the immediate final test, $t(37) = 15.02$, $d = 2.95$, but performance did not differ between test and control, $t(37) = 1.43$, $p = .16$. Furthermore, there was no difference between the three practice types for the 1-week-only final test, $F(2, 76) < 1$. Thus, for the immediate final test of nonretrievable items, there was a large advantage of study over test. This advantage accounted almost completely for the study advantage in the collapsed data at the immediate test.

Does an immediate test help?

To ascertain whether an immediate test enhances long-term retention, we compared the two delayed final tests. Separately for each type of item, we conducted a 2 × 3 ANOVA with type of delayed final test (1-week-after-immediate versus 1-week-only test) and practice type (study, test, and control).

For retrievable items (Figure 4, top), there was no significant difference between the 1-week-after-immediate test and 1-week-only test, $F(1, 69) < 1$. Both the main effect of practice type and the interaction were significant, $F(2, 138) = 31.45$, $\eta_p^2 = .31$; and $F(2, 138) = 6.51$, $\eta_p^2 = .09$, respectively. As seen in the figure (top), this interaction was due to the three practice types becoming more similar to each other for the 1-week-after-immediate final test than for the 1-week-only final test. This result is sensible because in the 1-week-after-immediate test condition, all of the words received additional testing during Session 1 (the immediate final test): It is likely that all of the words benefited from this immediate final test considering that these were initially retrievable items.

For nonretrievable items (Figure 4, bottom), there was no significant difference between the two delayed final tests, $F(1, 75) = 1.30$, $p = .26$. Both the main effect of practice type and the interaction were significant, $F(2, 150) = 5.01$, $\eta_p^2 = .06$; and $F(2, 150) = 8.40$, $\eta_p^2 = .10$, respectively. As seen in the figure (bottom), this interaction was entirely due to the study condition: The only practice type that differed between the two delayed final tests was the study condition, which was significantly better in the 1-week-after-immediate test condition, $t(75) = 3.42$, $d = 0.79$. That is, for material that was not initially retrieved, study followed by an immediate test enhanced long-term retention.

Discussion

Experiment 1 found the testing effect in the combined data, showing a typical cross-over interaction between retention interval and practice type. Furthermore, although test was no better than control for an immediate final test, test produced the best performance after 1 week. By dividing the data into retrievable versus nonretrievable items, we examined different contributions to this interaction. As expected for an immediate test of retrievable items, performance was near ceiling, but there was still a slight benefit of study over test. For the 1-week-only final test of retrievable items, the benefit of test over study was .10, which fully accounted for the .05 advantage in the collapsed data. However, there were long-term benefits of study as demonstrated by the advantage for retrievable items in the study condition as compared to the control condition (.42 versus .28). Thus, even for items that had been successfully encoded, additional study (i.e., overlearning) was

beneficial. For nonretrievable items, there was a large advantage of study over test for an immediate final test (.58), which almost completely explained the advantage in the collapsed data (.31), given a very small advantage for retrievable items (.03). These results explain why some experiments find a robust testing effect with delay while others find an advantage of study for an immediate test—the proportion of initially retrievable items serves to produce one data pattern or the other.

Considering that there was little opportunity for retrieval practice for nonretrievable items, it is not surprising that there was no testing benefit following a 1-week delay. Surprisingly, however, although participants were able to successfully learn the initially nonretrievable items immediately after additional study (.65 or around two thirds of these items became retrievable), they forgot much of what they learned 1 week later (.02 for the 1-week-only test condition, and .12 for the 1-week-after-immediate test condition). However, it is difficult to compare forgetting rates for retrievable versus nonretrievable items considering that performance on the immediate test was different for these items (e.g., performance was near ceiling for the immediate test of retrievable items). To equate performance, we conducted an additional analysis using the study condition from the 1-week-after-immediate test, selecting only items that were recalled on the immediate test. One week later, performance for the retrievable items ($M = .45$, $SE = .06$) was still greater than performance for the nonretrievable items ($M = .19$, $SE = .03$), $t(72) = 4.19$, $d = 0.99$. This result suggests that difficult material (initially nonretrievable items) is highly vulnerable to forgetting over the long term, even if it has been successfully retrieved after additional study practice. Of course, it is possible that memory strength was not perfectly equated for the retrievable and nonretrievable items despite the fact that both were recalled on a test following study practice. Still, the results are suggestive of a forgetting rate difference.

Experiment 1 revealed the role of retrievability in the testing effect. For initially nonretrievable items, although there was a large advantage of additional study on the immediate final test, there were no differences between study, test, and control with a 1-week delay. Thus, no single form of practice helped nonretrievable items in terms of long-term retention. However, taking an immediate test after additional study (i.e., a mixed practice strategy) slightly but significantly enhanced long-term retention, which suggests that there may be hope for initially nonretrievable items. This result is seen by comparing the two 1-week delayed final tests that differed in whether there was an immediate final test at the end of Session 1. However, Experiment 1 did not include the necessary comparison to ascertain whether this apparent testing effect for initially nonretrievable items was any greater than would have occurred if additional study was followed by more study rather than an immediate test. Next, we investigated this effect in greater detail.

EXPERIMENT 2

To investigate whether a testing effect can be obtained for initially nonretrievable items after additional study, Experiment 2 included the study condition followed by an immediate test, as in Experiment 1 while also examining one that included even more additional study instead of an immediate test. Experiment 2 was similar to Experiment 1, but in Experiment 2, participants only practised nonretrievable items during the intervening phase, and there were only two practice types (study and test).

Method

Participants

Fifty-two undergraduate students at the University of California, San Diego were recruited and received credit for psychology courses in return for their participation.

Materials

The same 90 word pairs as those in Experiment 1 were used.

Design

The design was a within-subject design composed of study and test conditions.

Procedure

The procedure was identical, except as noted, to the study condition of Experiment 1 in which participants only practised nonretrievable items and were given both an immediate test and then a 1-week delayed final test (1-week-after-immediate test condition). While this condition was termed a study practice condition in Experiment 1, we term it a test practice condition in Experiment 2 because we examined the effect of the immediate test on performance after a 1-week delay. In addition to this test condition, we included a true study condition that consisted of studying nonretrievable items one more time during the intervening phase instead of an immediate test.

The intervening phase consisted of four blocks. During each of the first two blocks, all nonretrievable items were presented one at a time for 5 s. Then, a simple maths distractor task was given for 30 s, which was followed by a final study block for half of the items and a test block for the other half of the items. Block order was counterbalanced for these last two blocks (either ST or TS).

Results and discussion

Initial test performance was not significantly different between the study ($M = .58$, $SE = .04$) and test ($M = .57$, $SE = .04$) conditions, $t(51) < 1$, suggesting that retrievability was equal for these conditions.

Figure 5 shows the mean proportion correct recall for the immediate and 1-week final tests. We replicated the substantial benefit of additional study for an immediate test of initially nonretrievable items (.64, comparable to the .65 found in Experiment 1).

For nonretrievable items, there was no significant difference between the study and test conditions (i.e., SS followed by S, versus SS followed by T), $t(51) = 1.09$, $p = .28$. This indicates that there was no testing effect for initially nonretrievable items. Nevertheless, it is notable that performance did not differ between nonretrievable items of the study condition and retrievable items, $t(51) = 1.56$, $p = .12$; nor between nonretrievable items of the test condition and retrievable items, $t(51) < 1$. In other words, performance for initially nonretrievable items after a 1-week delay was brought up to the level of the initially retrievable items that did not receive any additional practice.

Note that a quantitative comparison between Experiments 1 and 2 is puzzling in several regards. For instance, in Experiment 1, nothing was retained from the first two study blocks of nonretrievable items, but in Experiment 2, the third study block enabled long-term retention that was equivalent to test practice. However, there is one

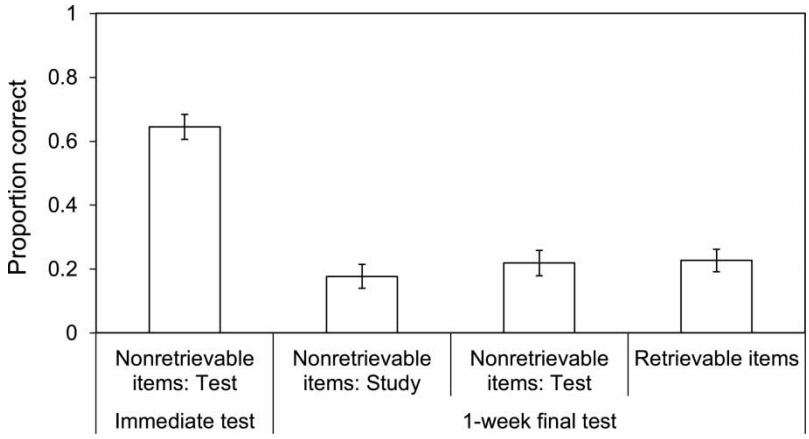

Figure 5. *Proportion correct recall on the intervening and 1-week final tests of Experiment 2. Error bars depict ± 1 standard error of the mean.*

key procedural difference that makes it difficult to compare the two experiments. The immediate test list in Experiment 1 was roughly four times longer than the immediate test list in Experiment 2 because the immediate test list in Experiment 2 only included nonretrievable items of the test condition (i.e., test practice only on nonretrievable items at the end of Session 1). By contrast, all of the items were tested in Experiment 1 (i.e., reexposing retrievable items as well). In any event, the procedure used in Experiment 2 failed to reveal a long-term advantage of test over study for nonretrievable items.

GENERAL DISCUSSION

We performed two experiments examining the testing effect in the absence of feedback during test practice. Experiment 1 investigated the role of retrievability by using an initial test to label items as retrievable (initially recalled) or nonretrievable (initially not recalled). Aggregating across all items, we found the typical cross-over interaction between practice type and retention interval: Performance was better with study practice than with test practice for an immediate final test, but the opposite was true after a 1-week delay. Analysing the results separately for retrievable and nonretrievable items, we identified different contributing sources to this interaction. For retrievable items, the benefit of study over test for the immediate final test accounted for a tiny fraction of the benefit seen in the aggregate data (around 5%) whereas the retrievable items fully accounted for the advantage of test over study for the delayed final test (around 99%). Nevertheless, for retrievable items, there was a long-term benefit of study as compared to a control condition without any additional practice. For nonretrievable items, additional study provided a huge increase for the immediate final test, which was a large fraction of the benefit seen in the aggregate data (around 95%), whereas there was virtually no difference between study and test for the delayed final test: The nonretrievable items did not account for the testing effect in the aggregate data (if any, around 1%). To briefly summarize, the long-term benefits of test practice as compared to study practice were entirely due to the retrievable items. In addition, our study is the first to demonstrate that the short-term benefits of study over test practice were almost entirely due to nonretrievable items. Thus, the cross-over interaction between practice type and retention interval is largely explained by using a mixture of items that differ in retrievability: The effect size of study/test practice is due to the relative contribution of retrievable and nonretrievable items.

In retrospect, these results may seem obvious. However, most studies of the testing effect do not consider item retrievability, and yet Experiment 1 clearly demonstrates that a higher proportion of nonretrievable items will tip the interaction pattern towards study benefits with an immediate test but little difference with delay (similar to the previous results shown in the left side of Figure 1) whereas a higher proportion of retrievable items will tip the interaction pattern towards little difference on an immediate test, but robust test benefits following a delay (similar to the previous results shown in the right side of Figure 1).

Experiment 1 also provided some evidence that a mixed strategy of additional study followed by testing may provide a long-term benefit for initially nonretrievable items. However, Experiment 1 did not include the necessary condition to ascertain whether there was a benefit of test over study in this case. Therefore, Experiment 2 focused on this particular comparison examining practice for initially nonretrievable items. Somewhat to our surprise, there was no difference between study versus test as measured after a 1-week delay. None of these items were recalled on the initial test, but after two additional study episodes, around two thirds of the items became retrievable for an immediate test (similar to the result in Experiment 1). Nevertheless, the additional study failed to convert the nonretrievable items into retrievable items in the sense of exhibiting the benefit of test over study as measured on a delayed test.

In two experiments, we made the following three observations regarding long-term retention. First, the advantage of testing without feedback rather than studying is entirely due to items that

are initially retrievable, supporting the retrieval hypothesis. Our experiments provide strong evidence that successful retrieval is an important factor underlying the testing effect. Second, both testing and studying help items that are initially retrievable. In contrast to this result, Karpicke and Roediger (2007, Experiment 2; also see Karpicke & Roediger, 2008) found that additional study of retrievable items produced no advantage. However, both their additional study condition (STST) and control condition (STSnT, where Sn is study of only the items that were not recalled for the first T) were followed by additional testing (the final T during the intervening phase) prior to the delayed final test. Therefore, any benefit of additional study for initially retrievable items may have been obscured by this final test before delay, which may have brought the control condition up to the level of the study condition. In contrast, we used a control condition without any additional testing (a true baseline) beyond the first test to determine initial retrievability, and unlike those studies, we found that retrievable items benefited from additional study. Third, it appears that initially nonretrievable items are fundamentally different in terms of the long-term effectiveness of study/test practice. We failed to find any advantage of test over study for initially nonretrievable items even when test practice performance was quite high following study practice. However, there are a variety of factors that enhance the strength of testing effects, such as the number of successful retrievals, the type of test practice (e.g., free recall versus recognition), and the spacing between retrieval attempts (e.g., Glover, 1989): A parametric investigation of these factors may demonstrate that it is possible to rehabilitate initially nonretrievable items to the point where they show a testing effect.

At the conclusion of this study, we quote a passage from William James (1890) who contemplated the steep forgetting rate following massed study of material right before an exam:

> The reason why *cramming* is such a bad mode of study is now made clear. I mean by cramming that way of preparing for examinations by committing "points" to memory during a few hours or days of intense application immediately preceding the final ordeal, little or no work having been performed during the previous course of the term. Things learned thus in a few hours, on one occasion, for one purpose, cannot possibly have formed many associations with other things in the mind. Their brain-processes are led into by few paths, and are relatively little liable to be awakened again. *Speedy oblivion* [italics added] is the almost inevitable fate of all that is committed to memory in this simple way. (p. 663)

Our results are in agreement with these musings, showing that initially nonretrievable items are destined for speedy oblivion shortly after additional study (e.g., around 96% of forgetting rate over 1 week found in Experiment 1) even if they are subjected to a mixture of study and test practice.

Original manuscript received 17 June 2011
Accepted revision received 29 October 2011
First published online 8 February 2012

REFERENCES

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning & Verbal Behavior*, *8,* 463–470.

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: Current research and issues* (Vol. 1, pp. 396–401). New York, NY: Wiley.

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.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14,* 215–235.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Academic Press.

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

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

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, *18,* 879–919.

James, W. (1890). *The principles of psychology*. New York, NY: Holt. Retrieved July 31, 2006, from http://www.archive.org/details/theprinciplesofp01jameuoft

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

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

Kucera, H., & Francis, W. H. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16,* 192–201.

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

Roediger, H. L., III. (2000). Why retrieval is the key process to understanding human memory. In E. Tulving (Ed.), *Memory, consciousness, and the brain: The Tallinn conference* (pp. 52–75). Philadelphia, PA: Psychology Press.

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 & Cognition*, *31,* 1155–1159.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, *11,* 641–650.

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

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning & Memory*, *4,* 210–221.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*, *56,* 252–257.

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

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11,* 571–580.

Wheeler, M. A., & Roediger, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3,* 240–245.

APPENDIX A

Results for the initial and intervening tests in Experiment 1

The mean proportion correct recall during the initial test for all conditions and the mean proportion correct recall during the intervening tests for the test condition

Table A1. *Mean proportion correct recall for the initial and intervening tests in Experiment 1*

		Retrievable item group			*Nonretrievable item group*		
		Initial (all items)	*Intervening*		*Initial (all items)*	*Intervening*	
		T	*T1*	*T2*	*T*	*T1*	*T2*
Immediate	Study	.49 (.048)			.46 (.047)		
	Test	.52 (.049)	.89 (.018)	.92 (.019)	.46 (.048)	.06 (.018)	.08 (.019)
	Control	.50 (.049)			.48 (.047)		
1 week only	Study	.46 (.049)			.50 (.046)		
	Test	.48 (.050)	.90 (.019)	.92 (.020)	.50 (.047)	.04 (.018)	.06 (.019)
	Control	.49 (.049)			.50 (.047)		

Note: Standard errors of the mean are in parentheses. T = test.

APPENDIX B

Results of the control condition for the final test in Experiment 1

The complete breakdown of the control condition

Table B1. *Mean proportion correct recall of the control condition for the final test for each group of the participants in Experiment 1*

		Participants				
Final test	*Control items*	*Retrievable item group*	*Nonretrievable item group*	*df*	*t*	*p*
Immediate	Retrievable	.88 (.03)	.89 (.02)	72	<1	
	Nonretrievable	.07 (.02)	.05 (.02)	72	1.21	.23
1 week only	Retrievable	.28 (.03)	.32 (.03)	72	1.13	.26
	Nonretrievable	.03 (.02)	.05 (.03)	72	<1	

Note: Standard errors of the mean are in parentheses.