Journal of Experimental Psychology:
Learning, Memory, and Cognition
2014, Vol. 40, No. 1, 115–124

0278-7393/14/$12.00 DOI: 10.1037/a0034252

On the Transfer of Prior Tests or Study Events to Subsequent Study

Benjamin C. Storm
University of California, Santa Cruz

Michael C. Friedman, Kou Murayama, and Robert A. Bjork
University of California, Los Angeles

Tests, as learning events, are often more effective than are additional study opportunities, especially when recall is tested after a long retention interval. To what degree, though, do prior test or study events support subsequent study activities? We set out to test an implication of Bjork and Bjork's (1992) new theory of disuse—that, under some circumstances, prior study may facilitate subsequent study more than does prior testing. Participants learned English–Swahili translations and then underwent a practice phase during which some items were tested (without feedback) and other items were restudied. Although tested items were better recalled after a 1-week delay than were restudied items, this benefit did not persist after participants had the opportunity to study the items again via feedback. In fact, after this additional study opportunity, items that had been restudied earlier were better recalled than were items that had been tested earlier. These results suggest that measuring the memorial consequences of testing requires more than a single test of retention and, theoretically, a consideration of the differing status of initially recallable and nonrecallable items.

Keywords: memory, testing effects, relearning, retrieval practice

Testing does more than measure memory—it modifies memory in a way that makes information retrieved on a test more likely to be remembered in the future than it would have been otherwise (Bjork, 1975). An appreciation for the beneficial consequences of testing is not new. William James (1890) wrote about the benefits of active retrieval over passive repetition, and some of the earliest research in cognitive psychology buttressed his case (e.g., Abbott, 1909; Gates, 1917; Jones, 1923–1924; Sones & Stroud, 1940; Spitzer, 1939). As of today, there are hundreds of published studies documenting the potential of testing as a powerful tool for learning (for reviews, see Roediger & Karpicke, 2006a; Roediger, Putnam, & Smith, 2011).

The most explored benefit of testing, referred to as the *testing effect*, is observed when the act of taking an initial test improves performance on a later test. Evidence for the testing effect has come from research showing that participants who are given an initial test perform better on a final retention test than participants who are not given such a test (e.g., Bjork, 1975; Carpenter & DeLosh, 2005; Carrier & Pashler, 1992; Cull, 2000; Glover, 1989; Hogan & Kintsch, 1971; Karpicke & Roediger, 2008; Landauer & Ainslie; 1975; Landauer & Bjork, 1978; Landauer & Eldridge, 1967; Pyc & Rawson, 2007; Roediger & Karpicke, 2006b; Wheeler & Roediger, 1992; Whitten & Bjork, 1977).

Importantly, the testing effect is not simply the consequence of re-exposure to the to-be-learned information. In fact, the benefit of testing is often greater than the benefit that would have accrued from additional study (e.g., Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Karpicke & Roediger, 2008). That is, learners who are given an initial test often outperform learners who are given a restudy opportunity, especially when the later test is given after a long retention interval. A study by Roediger and Karpicke (2006b), for example, revealed that an initial study opportunity followed by three tests (without feedback) was inferior to four study opportunities when the criterion test was administered after a 5-min retention interval, but was superior when the criterion test was administered after a 1-week retention interval.

Another important feature of the testing effect—one that also demonstrates that testing effects are not simply a matter of re-exposure—is that a more difficult initial test can sometimes lead to better later recall than does a less difficult initial test, even when no feedback is provided, and especially after long retention intervals (e.g., Landauer & Bjork, 1978; Pyc & Rawson, 2009; Whitten & Bjork, 1977). Given that more items are successfully retrieved on an easier test, it may seem that more items should profit from an easier test, but the learning benefits of successful test-induced retrievals do not appear to be equal for easier and harder tests. More specifically, it appears that the more difficult or involved a retrieval effort, provided it succeeds, the greater the benefit of that retrieval as a learning event. Under some circumstances, then, such greater benefits can more than offset the disadvantage of a harder initial test—namely that fewer items are strengthened by the retrieval process.

This article was published Online First August 26, 2013.

Benjamin C. Storm, Department of Psychology, University of California, Santa Cruz; Michael C. Friedman, Kou Murayama, and Robert A. Bjork, Department of Psychology, University of California, Los Angeles.

The present research was supported in part by James S. McDonnell Foundation Grant 29192G. Some of these results were presented at the 2009 annual meeting of the Psychonomic Society in Boston, Massachusetts. We thank Elizabeth Ligon Bjork and members of CogFog for helpful comments. We also thank Genna Angello, Cynthia Flores, Monika Holser, Laura Hurley, Andrew Marian, Maxwell Mansolf, and Victor Sungkhasettee for their assistance with data collection.

Correspondence concerning this article should be addressed to Benjamin C. Storm, Department of Psychology, University of California, Santa Cruz, Social Science 2, 1156 High Street, Santa Cruz, CA 95064. E-mail: bcstorm@ucsc.edu

A Distribution-Based Interpretation of Retrieval as a Memory Modifier

Recently, Halamish and Bjork (2011) and Kornell, Bjork, and Garcia (2011) have tested and found support for a distribution-based interpretation of testing effects. This interpretation incorporates several assumptions. The first is that—owing to differences across participants, items, and level of item-by-item encoding efficiency during an initial study phase—the to-be-learned items are distributed continuously on some memory-strength dimension. The second assumption is that an opportunity to restudy the items moves this distribution to the right—that is, all items get their memory strength boosted by some amount. A third and critical assumption is that an initial test bifurcates the distribution: Items that are successfully recalled on the initial test are strengthened, and strengthened more than when those same items are restudied, whereas items that are not recalled successfully on the test are left with the same memory strength they had before the test. A final assumption is that the more difficult the initial test, the more the memory strength of successfully retrieved items is incremented.

These assumptions are sufficient to account for how the benefits of an initial test, versus the benefits of a restudy opportunity, interact with initial-test difficulty, retention interval, and final-test difficulty. Thus, for example, as shown in Figure 1, the advantage of a restudy opportunity over an initial test when a criterion test is administered at a very short retention reflects, in this view, that the memory strength of *every* item is boosted by a restudy opportunity, whereas only those items successfully retrieved on the initial test get a boost in their memory strength. At a long retention interval, however, where criterion-test performance will tend to reflect the number of items that have high memory strength, the larger boost

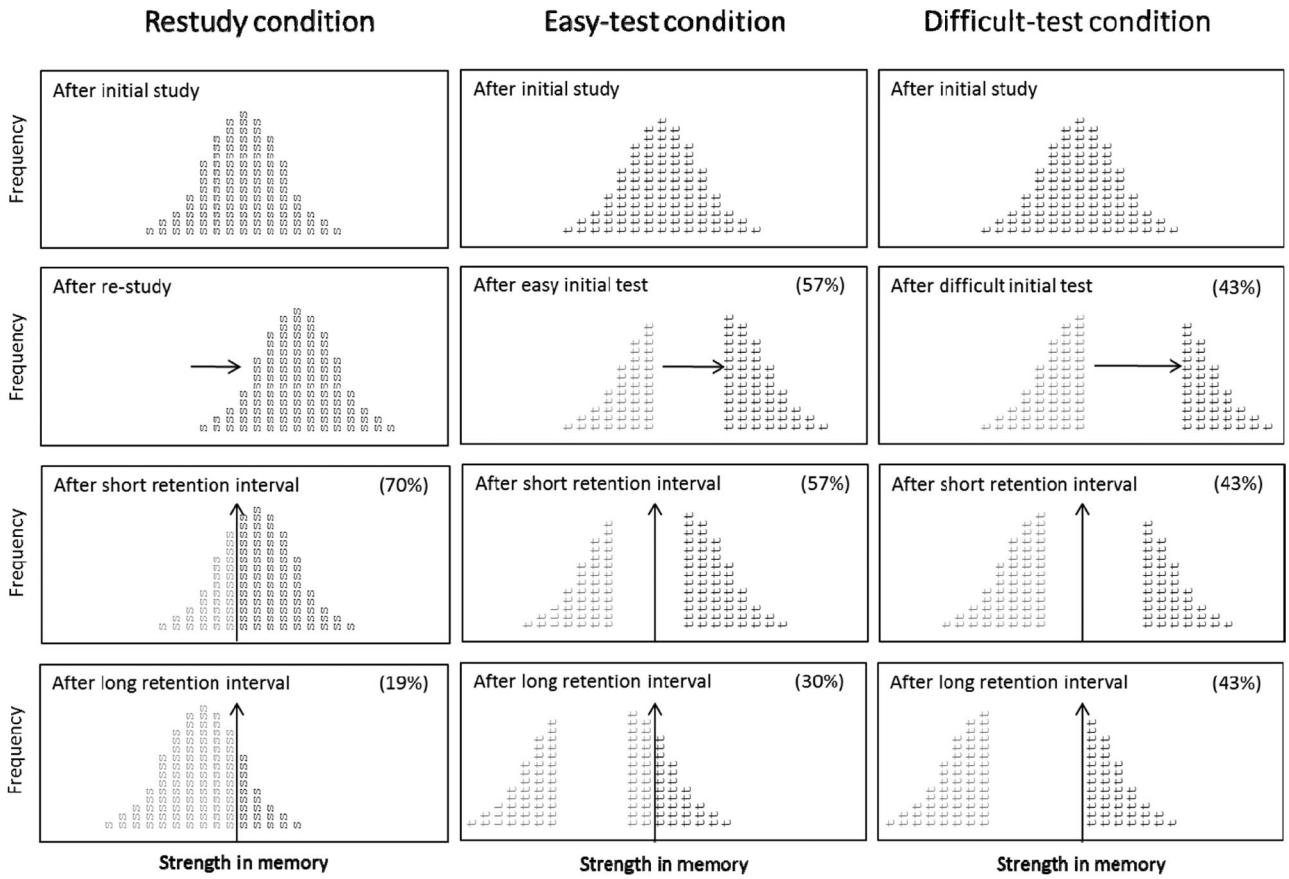

Figure 1. Simulated memory strength for three hypothetical sets of 100 items. The left column represents items that were restudied. The middle column represents items that were tested using easy tests without feedback. The right column represents items that were tested using difficult tests without feedback. The top row of panels shows memory strength after initial study. In the second row of panels, all of the restudy items gain memory strength equally, whereas the tested items become bifurcated. Items that are successfully retrieved gain more strength than items that are restudied, with difficult retrievals leading to larger gains than easy retrievals. Items not successfully retrieved do not gain any memory strength. The vertical arrows represent the recall threshold (i.e., items to the right are recallable). The third and fourth rows of panels represent memory strength after a short and long retention interval, respectively. All items are forgotten at the same rate, but the bifurcated distributions in the test conditions appear to prevent forgetting when measured as the percentage of items that remain above threshold. Note that more items remain above threshold in the easy-test condition than in the difficult-test condition after a short retention interval, whereas the opposite is true after a long retention interval.

to successfully retrieved items in an initial-test condition, versus those same items in a restudy condition, is the critical factor. A similar argument applies to the effects of initial-test difficulty: As shown in the right two columns of Figure 1, an easy initial test can yield better performance at a short delay because more items are helped by such a test than by a hard initial test, but a hard initial test can yield better performance at a long retention interval because the boost to the items that are recalled on the initial test is larger.

Both Halamish and Bjork (2011) and Kornell et al. (2011) point out that the distribution framework just described, though sufficient for their experimental purposes, is clearly an oversimplification—in part because there is abundant evidence that it is unrealistic to think that items in memory vary on a single dimension of strength. Using alternative measures of memory, for example, researchers have shown that one type of encoding condition may appear to have produced greater "strength" than some other encoding condition when measured by an explicit test, such as free recall, whereas the opposite may appear to be true when measured by an implicit test, such as priming (see, e.g., Richardson-Klavehn & Bjork, 1988; Roediger, 1990).

Aside from differences in explicit and implicit tests of memory, it is also unrealistic, in our view, to think that performance on one test, even a delayed test, provides a pure measure of learning. One of the goals of the present research was to demonstrate that a tacit assumption in the testing-effect literature—namely, that if participants recall more items in a testing condition than in a restudy condition, we can then conclude that testing was the more effective condition for promoting learning—is not always appropriate. Although testing may lead to more items being recalled, such an advantage in the proportion of items recalled may belie the fact that less learning occurred overall across the entire set of to-be-learned items. Thus, for example, as shown in Figure 1, the "strength" of the restudied items not recallable on the final test—though not sufficient to meet some recall criterion—may be greater than the strength of the tested items that are not recallable. Although the long-term benefits of testing for the items that are successfully retrieved may often outweigh the costs of these failures—thus leading to a testing effect in terms of the proportion of items recalled on the final test, as shown in Figure 1—the total amount of learning accrued across the entire distribution of items may actually be less, particularly when the initial testing is very difficult.

Storage Strength Versus Retrieval Strength

Again, although the unidimensional representation in Figure 1 may be sufficient to make the point that final-test performance is not a good measure of the total effect of restudying versus testing, it is not realistic. There is ample evidence, dating back to the 1930s–1950s heyday of learning theory, that performance is not a reliable measure of learning and that, conceptually, what Estes (1955a, 1955b) labeled *habit strength* must be distinguished from what he labeled *response strength*. Such a distinction has been resurrected and embellished in Bjork and Bjork's (1992) new theory of disuse (NTD), a theoretical framework in which an item's *storage strength* in memory is distinguished from its *retrieval strength*. According to the NTD, storage strength, which is assumed to reflect learning, is a measure of how interassociated a given memory representation is with related representations in memory, whereas retrieval strength is assumed to be a measure of current ease of access—that is, how primed or activated the representation is in the presence of current cues. Recall at any given point in time is assumed to be solely a function of an item's current retrieval strength, whereas storage strength acts as a latent variable in the theory, one that retards the loss of retrieval strength across a retention interval (forgetting) and enhances the gain of retrieval strength during relearning. Importantly, an item high in storage strength can have either low or high retrieval strength (e.g., a childhood phone number and one's current phone number, respectively).

An additional assumption of the NTD framework is that storage strength grows as a pure accumulation process with additional study or retrieval, which implies two additional assumptions: that there is no limit to the amount of information that can be stored in long-term memory and that once gained, storage strength is never lost. Increments in storage strength are assumed, however, to be a decreasing function of current storage strength. That is, the higher the current storage strength, the less there is to be gained toward some maximum, so the increment that results from additional study or successful retrieval becomes smaller and smaller.

Retrieval strength, however, is assumed to be limited and, given sufficient time and interpolated events, will fall to zero with disuse, even though the knowledge or procedure in question remains in memory from a storage strength standpoint. It is in this respect that the NTD is a "new theory of disuse"—by comparison to Thorndike's (1914) original "law of disuse," which postulated that with disuse, items decayed from memory. A decrease in an item's retrieval strength can be attributed to the cue-dependent nature of retrieval and the constant adjustments in retrieval strength of other related items. With new learning, changes in context, or the additional practice of competing items in memory, the retrieval strength of a target item will be reduced. A hotel room number, for example, may be readily recallable (i.e., have high retrieval strength) while one is away on vacation, but then become nonrecallable rapidly when one returns home, owing to the change of cues and the quite rapid loss of retrieval strength of that number, given that it, typically, never gets beyond a low level of storage strength.

An important additional assumption of the NTD is that successfully retrieving an item from memory produces a larger increment in storage strength than does a more passive opportunity to restudy. It is for this reason that retrieving an item on an initial test is believed to make that item more recallable on a later test. Provided that retrieval is successful, initial testing provides a far more potent opportunity to accumulate storage strength, which in turn promotes a more lasting effect on retrieval strength. Although additional study (as opposed to initial testing) may promote accumulation of storage strength, such accumulation is likely to be relatively weak compared with initial testing and may not be sufficient to create the combination of storage strength and retrieval strength that will support retrieval success on a long-delayed final test.

Finally, and critical to the NTD being able to provide an account for a variety of basic learning and memory phenomena, are the theory's assumptions as to how storage strength and retrieval strength interact. Gains in retrieval strength are assumed to be a decreasing function of current retrieval strength (the greater the

current retrieval strength, the less there is to gain) and an increasing function of current storage strength (i.e., storage strength potentiates the gain of retrieval strength). Gains in storage strength, however, are assumed to be a decreasing function of both current storage strength *and* current retrieval strength. Thus, and somewhat unintuitively, the more accessible an item is at a given point in time, the less the increment in learning (storage strength) that results from a study or retrieval opportunity. Forgetting (loss of retrieval strength) is, in that sense, the friend of learning (gaining storage strength). Losses in retrieval strength are an increasing function of current retrieval strength (the more there is to lose, the greater the decrement) and a decreasing function of current storage strength, as mentioned earlier.

These assumptions lead to very interesting predictions in relation to the distributions of items created in typical testing-effect experiments. In the testing condition, some items are successfully retrieved during initial testing, resulting in large accumulations of storage strength and retrieval strength for those specific items, whereas items that are not successfully retrieved will not accumulate much storage strength at all. Consequently, if a participant in the testing condition was given the opportunity to study all of the items again after a delay, the items successfully retrieved on the initial test would benefit little because of their already high retrieval strength, and the items not successfully retrieved on the initial test would benefit little because of their low storage strength. In the restudy condition, however, although most of the items would have relatively low retrieval strength, all of the items would have accumulated some degree of storage strength, thus potentiating their subsequent study.

The bivariate distributions shown in Figure 2 illustrate these arguments. In the top two panels are shown the presumed long-term consequences of initial restudying versus initial testing, similar to the bottom panels of Figure 1, but with the addition of marginal distributions showing storage strength as well as retrieval strength. In this example, the scatterplots shown reflect an assumption that initial retrieval strength and initial storage strength are quite highly, but not perfectly, correlated, which—combined with the assumptions mentioned earlier as to how storage strength and retrieval strength interact—is why the marginal distributions become distorted normal distributions. The vertical line indicates the level of retrieval strength necessary for an item to be recallable.

As can be seen in Figure 2, initial testing creates more items that are high in both storage strength and retrieval strength than does initial restudy opportunities, but initial testing also results in more items that are low in both retrieval strength and storage strength. At the time of the delayed test, because prior restudying will have incremented the retrieval and storage strengths of all the to-be-learned items, not just those that are recallable during initial testing, it can then potentiate the effectiveness of subsequent study, as shown in the two bottom panels (a more detailed simulation of such a bivariate model is available in Bjork & Murayama, 2013).

Experiment 1

In summary, we believe that a single test of retention in testing-effect experiments may, under some conditions, misrepresent the relative accumulation of storage strength across the entire distribution of to-be-learned items in the testing and restudy conditions and that even when testing leads to a significant advantage in terms of recall, that advantage may be significantly reversed if learners are given an opportunity to study all of the items again at the time of the final test. We tested these predictions by administering a testing-effect experiment in which participants received additional study opportunities through feedback at the time of the final test. Participants first studied 36 Swahili–English pairs. A subset of the pairs was then repeatedly tested without feedback, whereas another subset was repeatedly restudied. A third subset served as baseline and received no testing or restudy opportunities. Participants were later given a cued-recall test after a 1-week delay in which they were provided the Swahili words and asked to recall the English associates. Importantly, this final test involved feedback. That is, after the conclusion of each delayed test trial, participants were briefly shown the intact Swahili–English pair. Following the first session of testing and feedback, participants were given a second test, once again with feedback. In total, participants were given six delayed tests with feedback. We predicted that a typical testing effect would be observed on the first test of retention but that this advantage would be significantly reversed on subsequent tests of retention. That is, despite recalling more items in the testing condition on the first test, participants would recall more items in the restudy condition on each of the subsequent tests.

Method

Participants. Twenty-four English-speaking undergraduate students from the University of California, Los Angeles (17 women, mean age = 20.6), received course credit for their participation.

Materials. Thirty-six Swahili–English pairs (e.g., Wingu–Cloud, Pombe–Beer, etc.) were selected. The words were simple nouns ranging in length from three to eight letters in both Swahili and English. All of the Swahili words were pronounceable. For counterbalancing purposes, the 36 pairs were divided into three subsets of 12. These subsets were created such that they had approximately the same association strength based on the norms established by Nelson and Dunlosky (1994). For a given participant, one subset of items served in the testing condition, another subset served in the restudy condition, and a final subset served in the baseline condition. The particular subset assigned to a given experimental condition was counterbalanced across participants. All pairs were shown in the center of a white background in Arial black font, size 44.

Procedure.

Initial study phase. In the first phase of the experiment, participants studied a list of 36 Swahili–English pairs presented one at a time on a computer screen for 12 s each. Participants were told that they would be tested on their ability to remember the English words when given the Swahili words as cues.

Initial testing/restudy phase. Immediately following the initial study phase, participants were given repeated testing and study practice for a portion of the Swahili–English pairs. Specifically, 12 of the pairs were shown intact for the participants to study again (restudy condition), 12 of the pairs were shown with the English word missing (testing condition), and 12 of the pairs were not shown at all (baseline condition). During this task, items in the restudy and testing conditions were shown on the screen for 4 s, and participants were instructed to say the English word out loud for the experimenter to record. Partici-

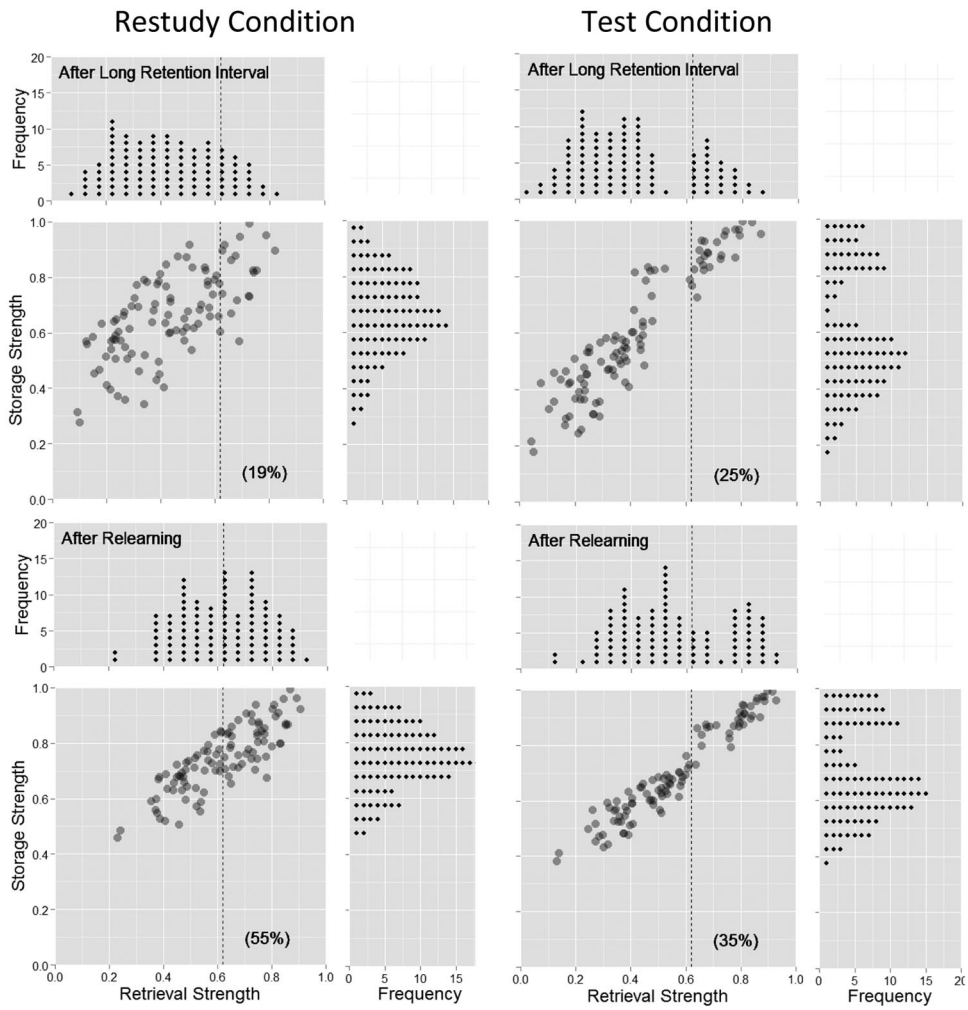

Figure 2. Simulated bivariate distributions showing retrieval strength and storage strength for two hypothetical sets of 100 items. The left column represents items that were restudied during initial learning. The right column represents items that were tested without feedback during initial learning. The top panels show distributions after a long retention interval. The bottom panels show distributions after a long retention interval and subsequent relearning. The dotted lines represent the recall threshold (i.e., items to the right are recallable). Prior to relearning, all items in the restudy condition are shown to have accumulated some amount of retrieval strength and storage strength. Owing to the benefits of testing, items in the test condition that are successfully retrieved gain substantially more retrieval strength and storage strength than do items in the restudy condition. Items in the test condition that are not successfully retrieved, however, do not accumulate any retrieval strength or storage strength. Because storage strength is assumed to potentiate new increments in retrieval strength, and because items in the restudy condition remain closer to the recall threshold following the delay, a greater proportion of the items in the restudy condition than in the test condition are shown to surpass the recall threshold after relearning.

pants were told to say the English word regardless of whether the pair was shown intact (e.g., saying "cloud" when shown "Wingu–Cloud") or with the English word missing (saying "beer" when shown "Pombe– _____"). No feedback was given. The restudy and test trials were randomly intermixed, with the only constraint that no more than three consecutive trials were of the same type. In total, participants received six sessions of practice with each session, separated by a 1-min distractor task. A different order was used for each session of restudy/test trials, but each item in the testing condition was always tested, and each item in the restudy condition was always studied. After the final session of restudy/test practice, participants were excused and instructed to return exactly 1 week later.

Delayed final tests. Upon returning, participants were told that they would be tested on each of the 36 Swahili–English word pairs they had studied in the initial study phase of the experiment. They were also informed that each test trial would be followed by feedback in which the English word would be presented along with the Swahili word and that they should pay attention to this feedback because each pair would be tested again.

Each test trial consisted of the Swahili word appearing on the screen for 4 s. During this time, participants were instructed to say out loud the associated English word for the experimenter to record. Responses were scored as correct if they were provided within the allotted 4 s. Immediate feedback was provided after each test trial such that the English response word was shown in green font next to the Swahili word. The 36 pairs were tested in a random intermixed order, with the only constraint that pairs in the same condition (testing, restudy, baseline) were not tested more than three times in a row.

Immediately following the completion of the first test, participants were given a second test for the same 36 pairs (in a new randomized order). This process repeated for a total of six test/feedback sessions such that every item was tested a total of six times.

Results and Discussion

Performance during initial testing practice. Participants successfully recalled the English words on 28% (SD = 16%) of the initial test trials. Although performance did improve numerically from the first trial (M = 27%, SE = 3%) to the last trial (M = 30%, SE = 4%), this difference was not statistically significant, $t(23) = 1.81$, $p = .08$, $d = .37$.

Performance on the first delayed final test. Recall performance on the first delayed test was analyzed as a function of condition (test vs. restudy vs. baseline) using a one-way analysis of variance (ANOVA). This analysis revealed a significant main effect, $F(2, 46) = 31.82$, $MSE = .01$, $p < .001$. As shown in the top panel of Figure 3, subsequent t tests confirmed that performance on the first 1-week delayed test was significantly better in the testing (M = 25%, SE = 3%) and restudy (M = 18%, SE = 3%) conditions than it was in the baseline condition (M = 5%, SE = 1%), ts(23) > 5, ps < .001. More importantly, a significant testing effect was observed such that participants recalled more items in the testing condition than in the restudy condition, $t(23) = 2.68$, $p = .01$, $d = .55$. This pattern nicely replicates previous testing-effect findings. Repeated testing led to better performance on a delayed test even when the initial tests were difficult and even though feedback was not provided.

Performance on the subsequent delayed final tests. A very different pattern of results was observed on the subsequent delayed tests. To analyze these data, we conducted a 3 (condition: test vs. restudy vs. baseline) × 5 (test trial: Test 2 vs. Test 3 vs. Test 4 vs. Test 5 vs. Test 6) repeated measures ANOVA. Not surprisingly, a main effect of test trial was observed such that participants performed increasingly better on each test trial, $F(4, 92) = 120.20$, $MSE = .01$, $p < .001$. More importantly, a main effect of condition was observed, $F(2, 46) = 32.83$, $MSE = .06$, $p < .001$. As shown in the top panel of Figure 3, participants remembered significantly more items in the test and restudy conditions than in the baseline condition. This time, however, a significant testing effect was not observed. In fact, performance on the subsequent five delayed tests was significantly better in the restudy condition than it was in the test condition, $t(23) = 4.77$, $p < .001$, $d = .97$. A 2 (first test vs. subsequent tests) × 2 (testing vs. restudy) ANOVA confirmed that the interaction between condition and test trial was significant, $F(1, 23) = 76.87$, $MSE = .01$, $p < .001$.

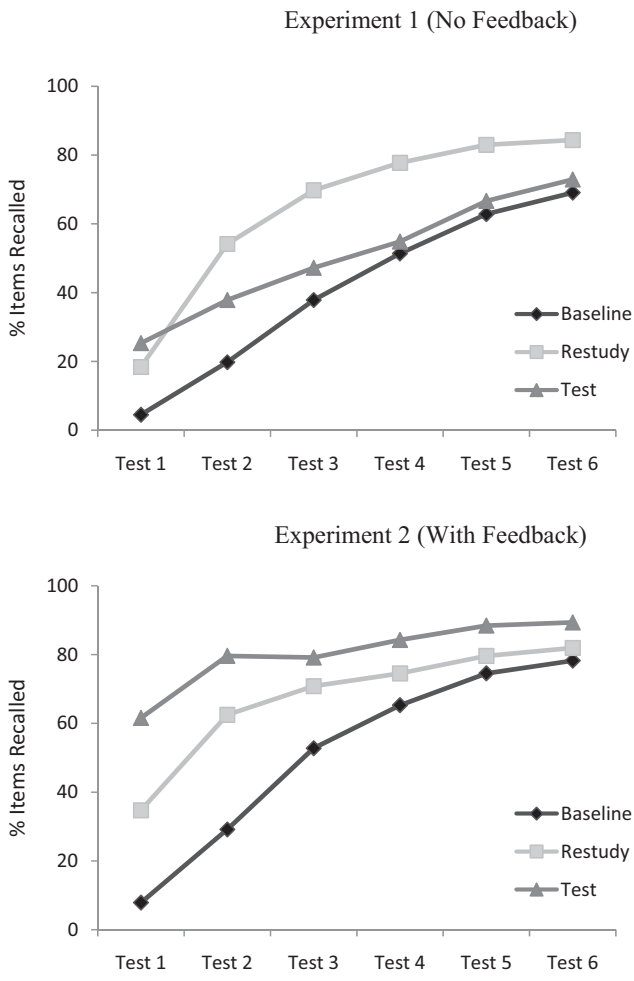

Figure 3. Recall performance on the six delayed tests in Experiment 1 (top panel) and Experiment 2 (bottom panel).

Amazingly, the testing effect was reversed after only a single test/feedback trial. On the first test following feedback (Delayed Test 2), participants recalled significantly more items in the restudy condition (M = 54%, SE = 5%) than in the test condition (M = 38%, SE = 4%), $t(23) = 3.38$, $p < .01$, $d = .69$. This reversal was striking. Whereas the 2-s study opportunity provided by feedback increased performance in the restudy condition from 18% to 54%, it only increased performance in the test condition from 25% to 38%. A paired samples t test confirmed that the difference in improvement across the two conditions was robustly significant, $t(23) = 6.09$, $p < .001$, $d = 1.24$. Furthermore, performance continued to be better in the restudy condition than in the test condition on each of the final four delayed test trials (all ts > 2.90, ps < .01). The observation that the difference between the restudy and test conditions became somewhat smaller across tests is difficult to interpret because performance approached ceiling.

Another way to convey the magnitude of the reversal in the testing advantage is to examine the proportion of participants who exhibited a testing effect on each delayed test session. During the first delayed test, 67% of the participants exhibited

a testing effect (others exhibited no difference between the testing and restudy conditions, or an advantage for the restudy condition). During the subsequent five tests, however, only 25%, 8%, 4%, 13%, and 8% of the participants exhibited testing effects, respectively. Finally, it is worth noting that all of the analyses reported above remained significant even when we limited our sample to the participants who actually exhibited testing effects on the first delayed test.

The results of Experiment 1 are in some ways even more compelling than we had anticipated. In typical testing-effect experiments, participants are given a single test of retention. Had we also used only a single test of retention, we would have observed a significant testing effect, which would have provided an incomplete, if not misleading, representation of the consequences of testing. Although performance on the initial delayed test was significantly better in the test condition than in the restudy condition, simply providing 2 s of feedback per item led to a substantial reversal in performance. Thus, though repeated opportunities to restudy failed to promote successful retrieval on the first delayed test, they did appear to potentiate the learning of nonretrieved items during feedback. Consequently, during each of the subsequent delayed test trials, participants performed significantly better for items in the restudy condition than for items in the testing condition. This pattern of results suggests that under certain conditions, testing can lead to better recall performance without necessarily potentiating future study opportunities and without necessarily leading to more overall learning across the entire distribution of to-be-learned information.

Experiment 2

An important feature of the first experiment was that initial testing was administered without feedback. Presumably, it was this lack of feedback—combined with the difficulty of the initial tests—that created the conditions for the observed results of Experiment 1, as illustrated in Figure 2. From the standpoint of the NTD and the item distribution framework illustrated in Figure 2, providing feedback during initial testing should produce an even larger testing effect on the first delayed test, and, more importantly, the testing effect should persist across the subsequent delayed testing/feedback trials. In other words, providing feedback should not only increase the storage strength and retrieval strength of nonretrieved as well as retrieved items during the initial testing trials, but also—by promoting success during the initial-test trials—increase the proportion of items benefitting from testing, thus making the testing condition a more powerful condition for learning than is the restudy condition, even when an additional opportunity to study the items is provided at final test. We tested this prediction in Experiment 2.

Method

Participants. Eighteen English-speaking undergraduate students from the University of California, Los Angeles (13 women, mean age = 20.2), received course credit for their participation.

Materials and procedure. The initial study phase and the delayed final test phase were identical to those of Experiment 1. In the initial testing/restudy phase, however, two important changes were made. First, all initial test trials were followed by 2 s of feedback. Specifically, for trials in the test condition, participants were given 4 s to recall each English word followed by an additional 2 s of feedback in which the English–Swahili pair was shown intact. Feedback was provided regardless of whether participants succeeded or failed to recall the English word. Second, to control for the total time of exposure, participants were provided 6 s for each trial in the restudy condition (as opposed to the 4 s provided in Experiment 1).

Results and Discussion

Performance during initial testing practice. Averaging across the entire initial testing/restudy phase, participants successfully recalled 70% (SD = 25%) of the English words given their Swahili translation. Not surprisingly, owing to the feedback, performance improved substantially across the six initial-test trials (Test 1: M = 38%, SD = 26%; Test 2: M = 59%, SD = 33%; Test 3: M = 69%, SD = 31%; Test 4: M = 78%, SD = 27%; Test 5: M = 85%, SD = 23%; Test 6: M = 91%, SD = 20%).

Performance on the first delayed final test. As shown in the bottom panel of Figure 3, performance on the first 1-week delayed test was better in the testing (M = 62%, SE = 7%) and restudy (M = 35%, SE = 6%) conditions than it was in the baseline condition (M = 8%, SE = 3%), ts(17) > 5, ps < .001. Most importantly, a significant testing effect was observed such that participants recalled more items in the testing condition than in the restudy condition, $t(17) = 2.68$, $p = .01$, $d = 1.44$. Ninety-two percent of the participants recalled more items in the testing condition than in the restudy condition.

Performance on the subsequent delayed final tests. Participants continued to exhibit a testing effect during each of the subsequent delayed tests. To analyze these data, we conducted a 3 (condition: test vs. restudy vs. baseline) × 5 (test trial: Test 2 vs. Test 3 vs. Test 4 vs. Test 5 vs. Test 6) repeated measures ANOVA. A main effect of test trial was observed such that participants performed increasingly better on each test trial, $F(4, 68) = 65.11$, $MSE = .01$, $p < .001$. A main effect of condition was also observed, $F(2, 34) = 20.90$, $MSE = .06$, $p < .001$. Not surprisingly, performance in the restudy and test conditions was significantly better than in the baseline condition, ts(17) > 3, ps < .01. The more important finding was that there was a significant difference in performance between the test and restudy conditions, $t(17) = 2.91$, $p = .01$, $d = .69$. Unlike in Experiment 1, the testing effect remained significant across the subsequent test trials.

General Discussion

The results of Experiments 1 and 2 illustrate, quite dramatically, that the consequences of prior studying or testing are only partly revealed by performance on a later criterion test. In the testing-effect literature—as well as in many other literatures—it is not uncommon for researchers to assume, implicitly or explicitly, that performance on a single final test provides a good, if not entirely accurate, reflection of learning or memory strength, especially when such a test is administered after a long delay. If participants recall more items in the testing condition than in the restudy condition, then it is typically concluded—or

at the very least implied—that testing was the superior condition for learning. As evidenced by the present results, however, this conclusion is not always warranted. When initial testing is difficult and when feedback is not provided (such as was the case in Experiment 1), testing may be superior to restudying in terms of facilitating performance on a delayed criterion test while being substantially inferior to restudying in terms of facilitating the effectiveness of subsequent study.

More specifically, participants in Experiment 1 studied Swahili–English pairs before repeatedly restudying a subset of the pairs (restudy condition) and repeatedly being tested on another subset of the pairs (testing condition). Although a significant testing effect was observed on a criterion test administered 1 week later, a single additional study opportunity—provided in the form of feedback during the final test—was sufficient to reverse that effect such that items in the restudy condition became significantly more recallable than items in the testing condition. As shown in Figure 3, the magnitude of the reversal was striking, shifting from a 7% advantage for the testing condition on the first delayed test to a 16% disadvantage for the testing condition on the second delayed test. This reversal remained significant across each of the next four test/feedback sessions.

Importantly, a very different pattern of results was observed when initial testing was administered with feedback. In Experiment 2, participants were given 2 s of feedback after each initial test trial, thus ensuring that all items received some boost in memory strength even if retrieval was unsuccessful. Under these conditions, items in the test condition were recalled significantly better than items in the restudy condition on the first delayed test and continued to be better recalled on each of the subsequent delayed tests. Clearly, the shortcoming of difficult tests in terms of boosting storage strength across the entire distribution of items is largely limited to instances in which difficult tests are provided without feedback. Of course, the specific dynamics and predictions will vary depending on a number of important factors (e.g., the number and difficulty of initial-test/restudy trials, the final retention interval, the type of feedback provided, etc.).

Theoretical Implications

Although the distribution-based bifurcation model illustrated in Figure 1 can account for why difficult testing without feedback might lead to better performance than additional study on a delayed criterion test—a finding replicated in Experiment 1—the present results demonstrate that performance on such a test does not necessarily imply better learning or superior memory strength. A single criterion test cannot directly measure the strength of an item in memory, nor can it measure the average strength of a distribution of items in memory. Instead, a single criterion test can only measure the proportion of items that are sufficiently strong to surpass some threshold for recall. Although testing may ensure that a larger proportion of items surpass this threshold by providing a substantial boost in the strength of items that are successfully retrieved, it appears to do so at the cost of failing to increase the strength of items that are not successfully retrieved.

Interpreted in this context, many findings in the testing-effect literature may need to be reconsidered—or at least carefully qualified. It is possible that many of the studies demonstrating long-term benefits of testing without feedback compared with restudying in terms of promoting long-term retention would have also demonstrated significant impairments in terms of its ability to promote accumulations in storage strength across the entire set of to-be-learned information (e.g., Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b; Whitten & Bjork, 1977). Note, however, that the implications extend beyond comparisons of testing versus studying. For example, there has been some debate regarding the relative potential of expanding retrieval practice versus uniform retrieval practice for promoting long-term retention (see, e.g., Karpicke & Roediger, 2010; Landauer & Bjork, 1978; Logan & Balota, 2008; Storm, Bjork, & Storm, 2010). One of the key benefits of expanding retrieval practice is that it is designed to keep as many items as possible above the recall threshold, thus ensuring that each item benefits from every test opportunity. Ensuring success, however, may come at a cost; namely, each of the tests must be relatively easy, thus preventing the potentially large benefits that might have been accrued from more difficult tests. Consequently, when expanding retrieval practice is followed by a relatively lengthy retention interval, it may be less than optimally effective in terms of promoting recall performance while being much more effective in terms of promoting subsequent study.

Practical Implications

Finally, from an applied perspective, it is important to consider the goal of implementing tests as an educational tool. On one hand, if one hopes to facilitate the largest overall boost in memory strength across the entire distribution of to-be-learned information, then repeated testing without feedback may be inferior to repeated studying even if it leads to better performance on a final delayed test. One might argue as well that in an ideal educational world, there should always be relearning after a delay (see, e.g., Bahrick, 1979; Rawson & Dunlosky, 2011)—that such relearning is essential for the long-term retention of key facts, procedures, and concepts—so conditions that manipulate the effectiveness of such relearning should be the conditions of choice.

There are, though, offsetting considerations that can favor testing. If one hopes to ensure that the largest proportion of to-be-learned information is recallable after a delay, then repeated testing without feedback may be superior to repeated studying even if it fails to provide the largest overall boost in memory strength across the entire distribution of to-be-learned information. If a learner only cares about final recall performance, for example, then a manipulation that facilitates subsequent study opportunities is not going to be very useful.

Concluding Comment

An important point that emerges from this research, as well as from the research by Halamish and Bjork (2011) and Kornell et al. (2011), is that certain standard ways that we plot the results from experiments on learning and memory can hide dynamics that are complex, important, and interesting (cf. Estes, 1956). Each of the learning curves in Figure 3, for example, combines 288 learning curves (each of 12 items for each of the

24 participants). There are good reasons for testing multiple participants and having multiple to-be-learned items, but we can fall prey to thinking that learning or retention curves represent the acquisition or forgetting of each item by each learner, whereas such curves may not capture the learning or forgetting of any one item by any one participant.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs: General and Applied, 11,* 159–177. doi:10.1037/h0093018

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308. doi:10.1037/0096-3445.108.3.296

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., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Bjork, R. A., & Murayama, K. (2013). *Studying, testing, retaining, relearning: A distribution-based model of how retrieval and restudying impact learning and performance*. Unpublished manuscript.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology, 19,* 619–636. doi:10.1002/acp.1101

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. doi:10.3758/BF03202713

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology, 14,* 215–235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.0.CO;2-1

Estes, W. K. (1955a). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377. doi:10.1037/h0046888

Estes, W. K. (1955b). Statistical theory of spontaneous recovery and regression. *Psychological Review, 62,* 145–154. doi:10.1037/h0048509

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53,* 134–140. doi:10.1037/h0045156

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. doi:10.1037/0022-0663.81.3.392

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 801–812. doi:10.1037/a0023219

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. doi:10.1016/S0022-5371(71)80029-4

James, W. (1890). *The principles of psychology*. New York, NY: Holt. doi:10.1037/11059-000

Jones, H. E. (1923–1924). The effects of examination on the performance of learning. *Archives of Psychology, 10,* 1–70.

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

Karpicke, J. D., & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition, 38,* 116–124. doi:10.3758/MC.38.1.116

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language, 65,* 85–97. doi:10.1016/j.jml.2011.04.002

Landauer, T. K., & Ainslie, K. I. (1975). Exams and use as preservatives of course-acquired knowledge. *Journal of Educational Research, 69,* 99–104.

Landauer, T. K., & Bjork, R. A. (1978). Optimal rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London, England: Academic Press.

Landauer, T. K., & Eldridge, L. (1967). Effects of tests without feedback and presentation-test interval in paired-associate learning. *Journal of Experimental Psychology, 75,* 290–298. doi:10.1037/h0025047

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition, 15,* 257–280. doi:10.1080/13825580701322171

Nelson, T. O., & Dunlosky, J. (1994). Norms of paired-associate recall during multitrial learning of Swahili-English translation equivalents. *Memory, 2,* 325–335. doi:10.1080/09658219408258951

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74,* 18–22. doi:10.1037/0022-0663.74.1.18

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition, 35,* 1917–1927. doi:10.3758/BF03192925

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60,* 437–447. doi:10.1016/j.jml.2009.01.004

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. doi:10.1037/a0023956

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39,* 475–543. doi:10.1146/annurev.ps.39.020188.002355

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043–1056. doi:10.1037/0003-066X.45.9.1043

Roediger, H. L., & Karpicke, J. D. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. doi:10.1111/j.1745-6916.2006.00012.x

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

Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. In J. Mestre & B. Ross (Eds.), *Psychology of learning and motivation: Cognition in education* (pp. 1–36). Oxford, England: Elsevier. doi:10.1016/B978-0-12-387691-1.00001-6

Sones, A. M., & Stroud, J. B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology, 31,* 665–676. doi:10.1037/h0054178

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology, 30,* 641–656. doi:10.1037/h0063404

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances

long-term retention. *Memory & Cognition, 38,* 244–253. doi:10.3758/MC.38.2.244

Thorndike, E. L. (1914). *The psychology of learning*. New York, NY: Teachers College Press.

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. doi:10.1111/j.1467-9280.1992.tb00036.x

Whitten, W. B., & Bjork, R. A. (1977). Learning from tests: The effects of spacing. *Journal of Verbal Learning and Verbal Behavior, 16,* 465–478. doi:10.1016/S0022-5371(77)80040-6

Received May 13, 2013
Revision received July 12, 2013
Accepted July 25, 2013 ■