Journal of Experimental Psychology:
Learning, Memory, and Cognition
2011, Vol. 37, No. 4, 801–812

0278-7393/11/$12.00 DOI: 10.1037/a0023219

When Does Testing Enhance Retention? A Distribution-Based Interpretation of Retrieval as a Memory Modifier

Vered Halamish
University of Haifa

Robert A. Bjork
University of California, Los Angeles

Tests, as learning events, can enhance subsequent recall more than do additional study opportunities, even without feedback. Such advantages of testing tend to appear, however, only at long retention intervals and/or when criterion tests stress recall, rather than recognition, processes. We propose that the interaction of the benefits of testing versus restudying with final-test delay and format reflects not only that successful retrievals are more powerful learning events than are re-presentations but also that the distribution of memory strengths across items is shifted differentially by testing and restudying. The benefits of initial testing over restudying, in this view, should increase as the delay or format of the final test makes that test more difficult. Final-test difficulty, not the similarity of initial-test and final-test conditions, should determine the benefits of testing. In Experiments 1 and 2 we indeed found that initial cued-recall testing enhanced subsequent recall more than did restudying when the final test was a difficult (free-recall) test but not when it was an easier (cued-recall) test that matched the initial test. The results of Experiment 3 supported a new prediction of the distribution framework: namely, that the final cued-recall test that did not show a benefit of testing in Experiment 1 should show such a benefit when that test was made more difficult by introducing retroactive interference. Overall, our results suggest that the differential consequences of initial testing versus restudying reflect, in part, differences in how items distributions are shifted by testing and studying.

Keywords: memory, testing, retrieval, distribution-based interpretation

Tests have traditionally been used to measure learning. A constantly growing body of research demonstrates, however, that tests are themselves learning events. The retrieval processes triggered by tests enhance subsequent recall, sometimes to a much greater degree than do comparable opportunities to restudy the information in question (for a recent review of testing effects, see Roediger & Karpicke, 2006a). There is often, though, an interaction such that restudying appears to be better than testing on the short term, whereas an advantage for testing emerges at longer retention intervals (e.g., Thompson, Wenger, & Bartling, 1978). Testing also appears to have greater benefits for subsequent free-recall or cued-recall testing than it does for forms of testing that are less dependent on recall, such as tests of recognition or priming (see, e.g., Hogan & Kintsch, 1971; Roediger & Blaxton, 1987).

In this paper we focus on the boundaries of the testing effect. When is testing better than restudying as a means of practice, in terms of subsequent performance? When is it not? We start by reviewing evidence for testing effects and briefly describing theories that have been used to explain such effects. Next, we focus on two factors that have been found to moderate the testing effect: final-test delay and final-test format. We then present a distribution-based interpretation of retrieval as a memory modifier to explain why the effects of testing interact with retention interval and final-test format. Finally, we report the results of three experiments that examined predictions derived from the framework.

Evidence for Testing Effects

The primary evidence for the benefits of tests consists of two findings: (a) Later recall profits from having an earlier test (or tests) of to-be-remembered information versus not having a prior test, and (b) later recall profits, under some conditions, more from having an earlier test (or tests) than it does from having an opportunity (or opportunities) to restudy the information in question. It is the second, more surprising finding that is the focus of the present paper.

In the typical experiment contrasting the benefits of testing with the benefits of additional study, participants initially study information for a final criterion test of some kind. They then, usually after a delay, go through either another study phase (study condition) or a test phase (test condition) in which they are tested on the initially studied information (hereafter referred to as the *practice test* or the *initial test*). Even more common are experiments in which the initial study phase is followed by several study cycles (repeated study condition) versus several test cycles (repeated test condition). In some experiments, feedback is provided after each test, but our focus in the present paper is on situations in which no

This article was published Online First April 11, 2011.

Vered Halamish, Department of Psychology, University of Haifa, Haifa, Israel; Robert A. Bjork, Department of Psychology, University of California, Los Angeles.

This research was support by Grant 29192G from the James S. McDonnell Foundation. We thank Nate Kornell for his insights into implications of the distribution framework; Alice Healy, Mark McDaniel, and Henry Roediger for their comments on an earlier draft of this paper; and Mia Nunez and Daniel Scheiffer for their help with data collection.

Correspondence concerning this article should be addressed to Vered Halamish, Department of Psychology, University of Haifa, Haifa, Israel. E-mail: halamish@research.haifa.ac.il

feedback is provided following a given test. Providing feedback in the testing condition typically enhances later recall (but see Hays, Kornell, & Bjork, 2010), but our interest is in comparisons of testing versus restudying in situations where time on task is held essentially constant and any benefits of testing can be attributed to the tests themselves rather than to the subsequent feedback. Any benefits of testing over restudying in situations when there is no feedback are both interesting and surprising. It is only when retrieval of the tested information succeeds that there is a reexposure of sorts to that information, whereas there is always a reexposure when information is restudied.

When memory is tested on a subsequent, criterion test (hereafter referred to as the *final test*) in such experiments, performance is often higher in the test (repeated test) condition than in the study (repeated study) condition. For example, Roediger and Karpicke (2006b, Experiment 2) used short prose passages to examine the contributions of testing to learning. Participants studied a given passage for 5 min; then—before a final test—they restudied the passage during three 5-min periods, were tested on their free recall of the passage during three periods, or restudied the passage during two periods and were then tested during one period. When the final test was administered after only a 5-min delay, free recall was higher in the restudy condition, but when the final test was delayed 1 week, the repeated test condition produced the best final recall (and by a large amount). We elaborate more on this type of crossover interaction after the next section.

Theories of Tests as Learning Events

Two explanations of the benefits of testing have been most influential (see Roediger & Karpicke, 2006a, for a more comprehensive review, including of less influential accounts). The first is the *retrieval hypothesis* (Bjork, 1975; Dempster, 1996; Gardiner, Craik, & Bleasdale, 1973; Jacoby, 1978), by which some aspects of retrieval itself, like the effort involved in it, improve subsequent memory. According to the retrieval hypothesis, the testing effect is a *desirable difficulty*, a term used by Bjork (1994) to describe various learning strategies that enhance long-term memory although they apparently slow initial learning. Restudying information involves much more fluent processing than taking a test, but it is the more difficult, less fluent, activity that results in better subsequent memory. According to the retrieval hypothesis, the more difficult the initial test is, provided retrieval succeeds, the larger the benefit of testing should be. Indeed, this prediction has been supported in a number of studies (e.g., Auble & Franks, 1978; Benjamin, Bjork, & Schwartz, 1998; Gardiner et al., 1973; McDaniel & Masson, 1985).

A related explanation of the benefits of testing is that the retrieval processes engaged by tests constitute practice for a later criterion test. This explanation, referred to as a *retrieval practice* interpretation of tests as learning events (Bjork, 1988), has been embellished by Roediger and Karpicke (2006a) in the form of their *transfer-appropriate-processing hypothesis*. The basic idea is that tests enhance memory on subsequent tests because they give learners the opportunity to engage in retrieval processes of the type that will be required on a later test.

A simple prediction based on this hypothesis is that the benefits of practice tests should increase as a function of their similarity to the final test. Indeed, Thomas and McDaniel (2007) observed that a detailed-oriented encoding task (letter reinsertion) resulted in higher memory performance on a final detailed test than on a final conceptual test, whereas conceptual-oriented encoding task (sentence sorting) resulted in higher memory performance on a final conceptual test than on a final detailed test (see also Duchastel & Nungester, 1982).

This brief review of the currently dominant interpretations of why testing sometimes yields better later performance than does restudying is presented, in part, to provide a contrast with the distribution-based framework we propose. We do not attempt in our framework to explain why being (successfully) tested on a specific item strengthens its memory more than does restudying that item. We simply assume, based on the literature, that it does. Rather, we attempt to account for when advantages of testing should and should not be expected. Our model, therefore, is not intended to be a competing process model of why tests are learning events, and it may not capture other dynamics in test-effect experiments, such as the possibility of increased rehearsal of tested items. It does, though, make somewhat different predications than do the explanations reviewed above, especially as to why the relative benefits of testing and restudying interact with final-test format.

Moderators of the Benefits of Testing

This basic finding—that testing improves long-term memory relative to an equal amount of time spent restudying, even in the absence of feedback—has been replicated in numerous experiments with various types of materials, such as single words (e.g., Thompson et al., 1978; Zaromb & Roediger, 2010), paired associates (e.g., Allen, Mahler, & Estes, 1969; Toppino & Cohen, 2009), and text materials (e.g., Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b). For present purposes, however, testing is not always better than restudying. Sometimes restudying is better, and sometimes the two conditions do not differ significantly. We focus below on two of the factors that have been found to moderate the benefits of testing and that have received the most attention in the literature: final-test delay and final-test format.

Final-Test Delay

Studies that used *delayed* final tests have yielded consistent results: When the final test is delayed for a matter of days or weeks after the practice, a benefit of testing is usually observed (e.g., Allen et al., 1969; Hogan & Kintsch, 1971; Nungester & Duchastel, 1982; Roediger & Karpicke, 2006b; Wheeler, Ewers, & Buonanno, 2003; Wheeler & Roediger, 1992). In a study by Nungester and Duchastel, for example, participants studied a history passage and then restudied the passage, took a test on the passage, or engaged in an unrelated activity. In a final test administered 2 weeks later, participants in the test group outperformed participants in the other two groups.

Studies that have used final tests administered immediately or shortly after the practice phase have, however, yielded less consistent results. Tulving (1967), for example, used immediate testing and observed that study and test practice periods resulted in an equivalent amount of learning. Although no testing advantage was observed in this study, it served as an important demonstration of the power of testing in maintaining memory, given that there was

much less exposure to the studied information in the test condition than in the study condition. Similar results have been reported by other investigators (e.g., Thompson et al., 1978, Experiment 1). Other studies, however, reported a benefit of testing on immediate final tests (e.g., Gates, 1917; Kuo & Hirshman, 1996). Yet other studies that used immediate testing reported a disadvantage of testing (Thompson et al., 1978, Experiment 2).

The various relevant studies used not only different final-test delays but also different materials, different participants, and somewhat different designs. It is therefore difficult to pinpoint the effect of test delay on the benefits of testing across these different studies. More direct evidence for the role of test delay comes from studies that used different final-test delays within a single experimental design. Consistently, these studies have obtained an interaction of the testing effect with final-test delay (e.g., Chan, 2010; Roediger & Karpicke, 2006b; Runquist, 1983; Thompson et al., 1978; Toppino & Cohen, 2009; Wheeler et al., 2003). Results from these studies converge to suggest that the benefit of testing over restudying increases as final-test delay increases.

A good example of finding such an interaction with test delay is a study by Wheeler et al. (2003) in which participants studied single words and then practiced them under repeated study or repeated (free-recall) test conditions. A final free-recall test was administered after 5 min, 2 days, or 7 days. On the 5-min-delay test, more items were recalled in the repeated study condition than in the test condition. After 2 days, however, this effect was eliminated and the groups did not differ significantly; after 1 week, the repeated testing group outperformed the repeated study group. Similar results were obtained by Roediger and Karpicke (2006b, Experiment 1), as described earlier.

There is strong evidence, then, that the benefit of testing emerges and increases as the final test is delayed. This interaction with final-test delay has been explained by suggesting that whereas restudy strengthens memory storage, testing strengthens item retrievability (Birnbaum & Eichner, 1971; Bjork, 1975; Wheeler et al., 2003; see also Roediger & Karpicke, 2006a). The model that we present in this paper suggests a simpler and more comprehensive solution to the puzzle.

Final-Test Format

Final-test format is another factor that has been shown to moderate the testing effect, although the evidence is sparser than it is for the moderating role of final-test delay. The effects of initial-test format have also been investigated (see, e.g., Duchastel, 1981; Kang, McDermott, & Roediger, 2007), but our focus is on the effects of final-test format. In their classic study, Hogan and Kintsch (1971) found that participants, after studying a list of single words, free-recalled more of those words after having had an intervening test versus a second study opportunity, independent of whether the initial test was a recognition or a free-recall test. When the final test was a recognition test, however, no benefit of testing was obtained: Restudying and testing did not differ significantly when the initial test was a recognition test, and restudying was better when the initial test was a free-recall test.

Along similar lines, several other studies examined final-test format by comparing short-answer tests, which involve active production of information, to multiple-choice tests, which involved less active production. The results of those studies suggest that short-answer tests are more sensitive to the benefits of testing than are multiple-choice tests (R. C. Anderson & Myrow, 1971; Kang et al., 2007, Experiment 1; but see Glover's 1989 study, in which similar benefits of testing were obtained on free-recall, cued-recall, and recognition tests). In the Kang et al. study (Experiment 1), for example, there was a numerical (thought not a significant) benefit of testing (using a multiple-choice test) over restudying when the final test was a short-answer test but not when it was a multiple-choice test.

Hogan and Kintsch (1971) interpreted their results in term of a two-stage theory of recall. They attributed the moderating role of final-test format to the qualitatively different processes that are involved in free recall and recognition. In particular, they suggested that testing improves subsequent retrievability, which underlies free recall, but not subsequent recognizability, which underlies both free recall and recognition. As a consequence, free-recall performance benefits from testing, but recognition performance does not. Our interpretation, as sketched below, suggests an alternative interpretation.

A Distribution-Based Interpretation of Retrieval as a Memory Modifier

In this section, we present a distribution-based interpretation of retrieval as a memory modifier, an interpretation that has evolved across the second author's many years of trying to understand the factors that moderate the benefits of testing (e.g., Bjork, Hofacker, & Burns, 1981; Gelfand, Bjork, & Kovacs, 1983). As we discuss at the end of this section, the distribution framework we present in this paper is oversimplified in certain respects, but it makes the same predictions in the present context as do more complex versions of the distribution-based interpretation of retrieval effects.

The basic assumptions of the distribution model are depicted in Figure 1. Panel A depicts a situation in which some material is initially studied and then practiced by restudying, and panel B depicts the corresponding testing (without feedback) situation. The first assumption that we make is that before an initial study phase—and across all participants and all studied items—the studied items end up normally distributed, to a first approximation, on a memory-strength distribution. This distribution is depicted by the dotted-line distributions shown in Figure 1, representing memory strength before the initial study. Second, after an initial study phase, all items are strengthened, as depicted by the dashed-line distributions—representing memory strength after the initial study—being to the right of the dotted-line distributions.

A third and critical assumption is that, following the practice phase, items that are successfully recalled on the initial, practice test are strengthened more than are items that are restudied, as indicated by the right-hand solid curve on panel B (memory strength following successful testing) being to the right of the solid curve in panel A (memory strength following restudy). We assume, in addition, that a failure to recall in the testing conditions leaves those failed items with the same strength in memory they had before, as shown by the left-hand solid curve on panel B (memory strength following failed retrieval attempt), which overlaps the dashed curve (prepractice memory strength). Under some circumstances, the assumption that nonrecalled items are left with the same strength may be unrealistic, given the evidence for *retrieval-induced forgetting* (M. C. Anderson, Bjork, & Bjork,

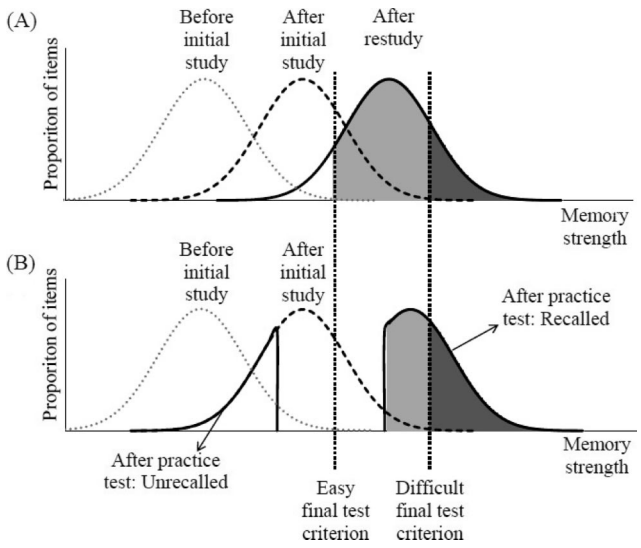

Figure 1. A distribution-based theory of retrieval as a memory modifier. Dotted curves represent memory strength prior to the initial study; dashed curves represent prepractice memory strength after initial study and before additional practice; solid curves represent postpractice memory strength following restudy (panel A) and test (panel B). Dotted vertical lines represent final-test recall criterion for easy tests and difficult tests. Shaded areas represent items recalled on the final test.

1994)—that is, that recalling some items associated with a given cue can impair the subsequent recall of other items associated with the cue. But in the research that we report using cued-recall procedures on the initial test, each target item is associated with a unique cue.

Figure 1 depicts a situation in which about 70% of the initially studied information is successfully retrieved on the initial test. A final assumption is that the more difficult the initial test, the larger the benefit for the items successfully recalled on that test. This assumption is not necessary for the purpose of the experiments presented in the current paper, but it is germane to issues addressed in the General Discussion.

Based on these assumptions, the main prediction of the model is that final-test difficulty moderates the testing effect. Test difficulty is represented in Figure 1 by the dotted vertical lines. Items that are to the right of these lines (marked by the gray areas in the figure) are items that are stronger than the criterion for recall and thus will be retrieved; items that are to the left of these lines will not be retrieved on the final test. In the specific situation depicted in Figure 1, there is a benefit of restudying over testing for the easy test and a benefit of testing over restudying for the difficult test (i.e., a crossover interaction). In general, though, what the model predicts is that there will be an interaction of final-test difficulty and whether information is initially tested or restudied. Whether a crossover interaction is observed or not should depend on where the easier and harder final-test criteria fall on the memory-strength dimension. With two hard final tests, for example, performance following initial testing may exceed performance following initial restudying, but the benefit should be larger for the more difficult final test. Conversely, with two easy final tests, there should be a benefit of restudying over testing, but the advantage of restudying should be larger for the easier test.

We define final-test difficulty in terms of its retention criterion, that is, the minimum memory strength at the end of the practice phase that is required for an item to be retained on the final test. Final-test difficulty is assumed to be a function of properties of the final test itself, such as final-test format (e.g., recall vs. recognition) or the amount of cue support in cued-recall tests, but it also varies with whatever happens prior to the test itself. The same memory task can be easy when given just after the study test but difficult when given after a long interval. Similarly, if the final test follows an interfering task, it will be more difficult—in the sense that fewer items will be retrieved—than when the test follows a less interfering activity.[1]

Predicting the Effects of Final-Test Delay and Final-Test Format

The model predicts the effects of final-test delay in a straightforward way. At a short delay, the benefit of restudying, which is that all items are strengthened, will often outweigh the benefit of testing, which is that the retrieved items get a larger boost in memory strength than do corresponding studied items. At a long delay, on the other hand, the key factor will tend to be which condition produces a subset of the strongest items (i.e., which condition produces the most items that are strong enough to exceed a higher final-test criterion). As the final test is delayed, the larger boost given the items recalled on the initial test (vs. that given restudied items) will play a greater role, and advantages of testing over restudying will become increasingly apparent.

The arguments with respect to the effects of final-test format are similar. Free-recall tests are more difficult than recognition tests, for example, and thus should be more revealing of the benefits of testing. Note, however, that this explanation of why tests such as free recall and cued recall should show larger benefits of testing differs substantially from prior, existing explanations of test effects. Others have explained the interaction of test effects with format of the final test by saying, for example, that tests strengthen item retrievability whereas restudy strengthens item storage, and retrievability is a larger factor during tests of free recall, say, than it is during tests of recognition. It could well be that there are qualitative differences in how items are strengthened by testing versus by restudying, but our approach in the present paper is to examine the degree to which the effects of testing can be accounted for without appealing to such qualitative differences.

In the context of qualitative differences, it is important to emphasize that the version of the distribution model, as depicted in Figure 1, is indeed oversimplified in a major way, namely, in the assumption that items can be characterized as differing on a single strength dimension. Research on implicit and explicit measures of memory and on differences between recognition and recall pro-

[1] The more complex version of the model assumes (a) that the final-test criterion is not absolute but is subject to oscillations around some central value as a function of moment-to-moment changes, such as in the mental set of a participant, and (b) that increments to a recalled item's memory strength are a decreasing function of the item's current retrieval strength and an increasing function of its current storage strength (see Bjork & Bjork, 1992). Because neither complicating assumption affects the model's predictions in the context of the present experiments, we have presented the simpler version.

cesses has demonstrated that memory representations are multidimensional (for reviews, see Richardson-Klavehn & Bjork, 1988; Roediger, 1990). To account for other phenomena, such as performance on implicit tests versus explicit tests, one must assume a bivariate normal distribution, with items differing on two dimensions of strength, but a one-dimensional version suffices for the present purposes.

In addition, we need to emphasize that what is depicted in Figure 1 is *retrieval strength*, to use Bjork and Bjork's (1992) term, or *response strength*, to use Estes's (1955) term—not *storage strength* or *habit strength*, respectively. For many purposes it is necessary to consider not only the level of retrieval/response strength, which is assumed to completely determine the probability of recall, but also storage/habit strength, a latent variable assumed to influence the rate of forgetting and learning. Again, though, that is a consideration that can be ignored for the purposes of this paper.

Novel Predictions of the Framework

Our distribution-based framework makes a somewhat different prediction than does the retrieval-practice (or transfer-appropriate-processing) interpretation of tests as learning events. Those ideas predict that the overlap of initial test and final test format should be crucial for the benefit of testing. Thus, an initial cued-recall test should exercise processes demanded by a final cued-recall test more than it does the processes demanded by final free-recall test, and the converse should be true as well. By contrast, according to our framework, final-test difficulty is the crucial factor for the benefit of testing. Final free recall, therefore, should profit more from initial cued-recall testing than should final cued recall, regardless of the initial test format. Experiments 1 and 2 were designed to test that prediction.

The framework also predicts that another factor, degree of retroactive interference between the initial test or restudy opportunity and the final test, which has not been explored in prior research, should moderate the benefits of testing. Basically, the more retroactive interference, the more difficult is retention on the final test and, therefore, the more the benefits of initial testing versus restudying. Experiment 3 was designed to test that prediction.

The Current Experiments: General Procedure

The basic procedure was similar in all three experiments. Participants studied a list of related word pairs (e.g., *RENT*: *HOUSE*) under repeated study (SSS) or repeated test (STT) conditions (within participants). The participants then read a second list of paired associates for about 8.5 minutes, before taking the final test. In Experiments 1 and 2 the second list was unrelated to the first list, and the critical manipulation was the format of the final test, which was a test either of cued recall or of free recall, between participants. In Experiment 3 the second list was designed to contain pairs that either interfered with or did not interfere with individual pairs in the first list.

Experiment 1

In Experiment 1, final-test difficulty was manipulated by using three different final-test formats (between participants): (a) recall cued with the cue word and a fragment of target word; (b) recall cued with cue word only; and (c) free-recall tests. These three final tests were designed to serve as easy, intermediate, and difficult final tests, respectively.

Method

Participants and design. The participants were 72 students from the University of California, Los Angeles, 24 in each between-participants condition, who participated for course credit. The design was a 2 × 3 mixed design: Type of practice (restudy/test) and final-recall test format (recall cued with cue word and fragment/recalled cued with cue word only/free recall) were manipulated within and between participants, respectively.

Materials. The materials included 54 related word pairs (e.g., *RENT*: *HOUSE*), taken from Jacoby (1996). The pairs were chosen from a range of association frequencies according to various norms, and the highest frequency associate of a cue word was never selected. Association frequency ranged from .03 to .59 (M = .27). Target words were four or five letters in length. Target fragments, with two or three missing letters in each, were also adopted from Jacoby (1996), and the fragments were unique to a given target. The fragments were constructed such that there was at least one other associate that would complete a given target's fragment (e.g., *RENT*: —*SE* could be completed with *house* and with *lease*). The additional set of associates (e.g., *LEASE*) was not used in the current experiment but was used in Experiment 3.

Two lists of 24 pairs each were randomly selected from this pool of pairs. One list served as the critical list that was used in the study phase, practice phase, and final test phase; the other list was used in the distraction phase. Which list was assigned to which role was counterbalanced across participants. Each list was randomly split into two sublists of 12 word pairs, with one sublist assigned to the restudy condition and the other assigned to the test condition (counterbalanced across participants) when that list served as the critical list. Six additional word pairs served as three primacy buffers and three recency buffers in the study phase.

Procedure. There were four phases for the experiment: study, practice, distractor, and final test. During the study phase, the critical list was studied: Word pairs were presented one by one, in random order, for 3 s each, with a 1-s interpair interval. Participants were instructed to study the word pairs for a later memory test and to think about the association between the two words in each pair, because doing so might help them remember the pair. Participants were not given any information about the format of the final test.

During the practice phase that followed the study phase, participants went through two consecutive cycles of practicing the critical list. Items from one sublist of 12 pairs were restudied on both cycles, whereas the items from the other sublist of 12 items were tested on both cycles. Items from the two sublists were intermixed and presented in a different random order on each practice cycle, with a 1-s interpair interval. On study trials, a word pair was re-presented for 6 s. On test trials, the cue word and fragment of the target were presented for 6 s. Participants were asked to retain the appropriate target word and to type in the entire word while the cue word and fragment were present on the screen.

During the distractor phase, participants were asked to read a second list of related word pairs. In Experiments 1 (and 2, in

contrast to Experiment 3), word pairs from the second list were always unrelated to any of the word pairs from the critical list. Thus, this phase was designed as an activity that would provide nonspecific interference with—and prevent rehearsal of—items from the critical list, with the goal being to avoid recency effects and ceiling effects on the final test. Participants went through three consecutive cycles of reading the second list, and the pairs were presented in a different random order on each cycle. Each word pair was presented for 6 s, with a 1-s interpair interval. The cover story used was that we were interested in how prior study of the first list would affect the processing of the second list. Participants were instructed to read each word pair and say it to themselves. They were also asked to try to read the pairs at a constant rate, without thinking about the previous study phase when reading.

During the final test that immediately followed the distraction phase, participants were asked to recall the target words from the critical list. They were instructed in particular to recall the target (right-hand) words from the first (i.e., critical) list. In the *easy* test condition, participants were presented with the cue word and a fragment of the target word (e.g., *RENT*: —*SE*) and asked to type in the target. The test was self-paced, but participants were limited to 20 s per item, after which the computer advanced to the next item. The *intermediate* test condition was similar to the easy test condition, but no target fragments were provided (i.e., participants were cued with the cue words alone). In the *difficult* test condition, a free-recall test was administered: Participants were given a blank sheet of paper and were asked to write down the target words. The retrieval interval was self-terminated.

Results and Discussion

The distribution framework predicts that the relative or absolute advantage of initial testing over restudying, if any, should increase with test difficulty. Thus, the test of free recall should have been most reflective of any benefits of test, the cue-plus-fragment test the least reflective, with the cue-alone test falling in between. The retrieval practice idea, on the other hand, without additional assumptions, predicts the opposite pattern. The processes engaged by the cue-plus-fragment initial testing should overlap most that same condition on the final tests, overlap to a lesser degree the cue-alone final test condition, and overlap least with the free-recall test.

Practice-phase performance. Overall, averaged across the two practice cycles and three final-test conditions, participants recalled .82 of the studied targets during the practice phase. Performance did not differ significantly as a function of final-test condition ($F < 1$).

Final-test performance. The proportions of targets recalled correctly on the final test are shown in Figure 2 as a function of practice condition, restudy or test, and type of final test. Final test performance was subjected to a 2 (practice condition: repeated study vs. repeated test) × 3 (final-test difficulty: easy, intermediate, difficult) mixed-design analysis of variance.

The three different final tests were designed to represent different levels of test difficulty, and indeed that was the case overall, $F(2, 69) = 105.15$, $MSE = .05$, $p < .001$, $\eta_p^2 = .75$. However, as is apparent in Figure 2, a Tukey post hoc test revealed that the level of recall on the easy (cue-alone) test (.84) and the intermediate (cue-plus-fragment) test (.81) did not differ significantly and were both much higher ($p < .05$) than recall on the difficult (free-recall) test (.26). Thus, the three different final tests used in this study appear to have represented only two levels of final-test difficulty: a relatively easy level (cued recall with or without target fragments) and a relatively difficult level (free recall).

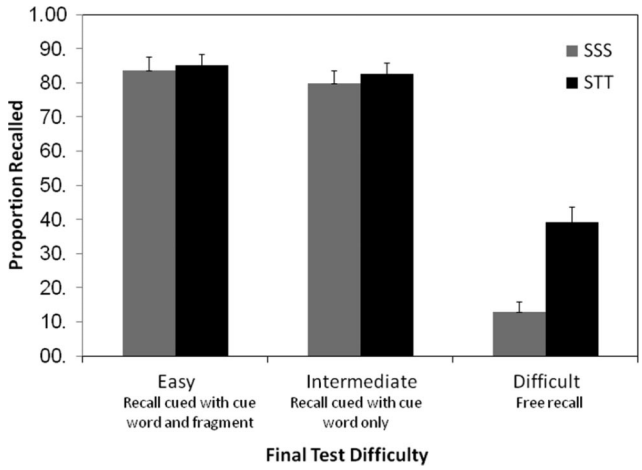

Figure 2. Mean proportion of targets recalled on the final test as a function of practice condition (repeated study, SSS/repeated test, STT) and final-test difficulty (easy/intermediate/difficult) in Experiment 1. Error bars represent standard error of the means.

A general testing effect was obtained: Across the different test-difficulty conditions, items were recalled better in the repeated test condition (.69) than in the repeated study condition (.58), $F(1, 69) = 27.07$, $MSE = .01$, $p < .001$, $\eta_p^2 = .28$. That advantage, though, as predicted by our distribution model, depicted in Figure 1, was a consequence of there being a large advantage of testing over restudying when the final test was the difficult (free-recall) test (.39 vs. .13), $t(23) = 7.05$, $p < .001$, $d = 1.52$. When, on the other hand, the final test was a relatively easy test (cued recall with/without target fragments), there was no advantage of testing over restudying (.85 vs. .83, respectively), $t(23) = 0.69$, *ns*; (.83 vs. .80, respectively), $t(23) = 0.85$, *ns*. This pattern resulted in a significant two-way interaction between practice condition and final-test difficulty, $F(2, 69) = 27.07$, $MSE = .02$, $p < .001$, $\eta_p^2 = .31$, as predicted by the distribution framework.

In Experiment 1, there was not a benefit of restudying over testing when the final test was a relatively easy test. In the context of the distribution model, the easy test criterion in Experiment 1 was a little higher than (i.e., to the right of) the one depicted in Figure 1.

From the perspective of the distribution framework, an aspect of Experiment 1 that may have worked against finding an advantage of restudying on the easy final test is that performance on the practice tests was over 80%. An advantage restudying has over testing is that all items are helped, whereas only the successfully retrieved items are helped via testing (though they are helped more). Experiment 2 was designed to examine whether, as implied by the distribution framework, reducing the proportion recalled on the initial tests in the testing condition would yield a crossover interaction, as depicted in Figure 1.

Experiment 2

Experiment 2 was similar to Experiment 1 except that a more difficult practice test—recall cued with cue word only—was used. In addition, final-test difficulty was manipulated by using only two different final-test formats (between participants): (a) recall cued with cue word only and (b) free recall.

Method

Participants and design. The participants were 40 students from the University of California, Los Angeles, 20 in each between-participants condition, who participated for course credit. The design was a 2 × 2 mixed design: Type of practice (restudy/test) and final recall test format (recalled cued with cue word only/free recall) were manipulated within and between participants, respectively.

Materials and procedure. The materials were identical to those used in Experiment 1. The procedure was the same as in Experiment 1, except that during the practice phase participants were cued with the cue words only and that only two levels of final test difficulty were used.

Results and Discussion

Again, the distribution framework predicts that any benefits of testing should emerge as the final test becomes more difficult, whereas the retrieval-practice idea predicts that any such benefits should be largest when the processes engaged by the initial and final tests overlap the most. In Experiment 2, this period corresponds to the cued-recall final-test condition, given that the initial test was also cued recall.

Practice-phase performance. Overall, averaged across the two practice cycles and two final-test conditions, participants recalled .73 of the studied targets during the practice phase. Performance did not differ significantly as a function of final-test condition, $t(30) = 1.10$, *ns*.[2]

Final-test performance. The proportions of targets recalled correctly on the final test are shown in Figure 3 as a function of practice condition, restudy or test, and type of final test. Final test performance was subjected to a 2 (practice condition: repeated study vs. repeated test) × 2 (final-test difficulty: easy, difficult) mixed-design analysis of variance.

The two different final tests were designed to represent different levels of test difficulty. Indeed, that was the case overall, as recall on the easy test (.84) was far higher than recall on the difficult test (.23), $F(1, 38) = 165.28$, $MSE = .05$, $p < .001$, $\eta_p^2 = .81$.

More important, and as is apparent in Figure 3, the benefits of testing versus restudying interacted with final-test difficulty, $F(1, 38) = 14.00$, $MSE = .02$, $p < .001$, $\eta_p^2 = .27$, and did so in a way that is consistent with the distribution model. When the final test was difficult, there was an advantage of testing over restudying (.29 vs. .18), $t(19) = 2.97$, $p < .01$, $d = 0.68$, whereas when the final test was easy, there was an advantage of restudying over testing (.89 vs. .79), $t(19) = -2.36$, $p < .05$, $d = 0.60$. Across the two test-difficulty conditions, there was no main effect of testing versus restudying (.54 vs. .53; $F < 1$).

Thus, final-test difficulty, not the overlap of initial- and final-test formats, moderated whether initial testing or restudying led to better recall on the final test. An initial cued-recall test had maximum benefits, relative to restudying, not when the final test was the exact same test but when the final test was a free-recall test without the cues being re-presented. The obtained results also pose difficulties for Hogan and Kintsch's (1971) dual-process interpretation of why testing, in their research, improved later free recall more than did restudying but did not have the same advantage on later recognition testing. According to that interpretation, testing enhances subsequent retrievability more than does restudying but does not enhance subsequent recognizability relative to restudying. Given, however, that both the final cued-recall and free-recall tests employed in Experiment 2 are tests of item retrievability, such a dual-process model cannot account—at least without added assumptions—for why there was an advantage of testing over restudying when the final test was a test of free-recall whereas the opposite was true when the final test was cued recall.

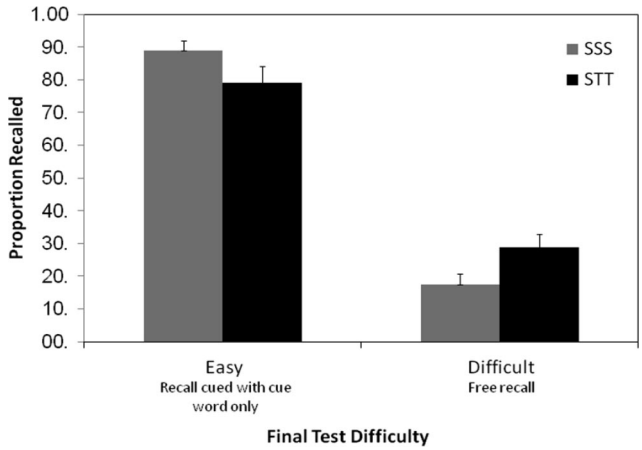

Figure 3. Mean proportion of targets recalled on the final test as a function of practice condition (repeated study, SSS/repeated test, STT) and final-test difficulty (easy/difficult) in Experiment 2. Error bars represent standard error of the means.

Experiment 3

Experiment 3 was designed to examine whether another way of increasing final-test difficulty, providing a higher level of retroactive interference prior to the final test, would also, as predicted by the distribution framework, increase the benefits of testing versus restudying. To our knowledge, the effect of rate of forgetting, as manipulated by degree of retroactive interference, on the relative effectiveness of initial testing versus restudying has not been examined in the literature.

As in the previous experiments, participants in Experiment 3 studied a list of related word pairs under repeated study (SSS) or repeated test (STT) conditions (within participants). The critical

[2] As intended, mean practice test performance was lower in Experiment 2 (.73) than in Experiment 1 (.82), $t(87) = 2.52$, $p < .05$. Therefore, although performance on the cued-recall test with cue words only and performance on the cued-recall test with cue words and target fragments did not differ on the final test in Experiment 1, they did differ between experiments when used as practice tests.

manipulation, crossed with practice condition, was introduced in the subsequent distraction phase, during which participants read a second list of related word pairs that was designed to contain pairs that either interfered with or did not interfere with particular pairs in the first list. The final test was a cued-recall test with recall cued by the cue word and a fragment of the target word. The motivation for using a final-test format that did not, in Experiment 1, show a benefit of testing over restudying was to examine in a different way the key prediction of the distribution framework (i.e., that it is final-test difficulty, not format, which is the key factor in whether initial testing produces better later performance than does restudying). The framework predicts that any way of making the final test difficult—in this case by introducing retroactive interference—should increase the likelihood that initial testing will lead to better final performance than does restudying.

Method

Participants and design. The participants were 36 students from the University of California, Los Angeles, who participated for course credit. The design was a 2 × 3 design: Type of practice (restudy/test) and type of distraction (repetition, AB-AB/interference, AB-AC/filler, AB-DE) were manipulated within participants.

Materials. We used the same set of cue–target pairs that was used in Experiments 1 and 2 along with the corresponding competitive fragment completions from Jacoby (1996). The study materials, therefore, included 54 items, each containing a cue and two competitive targets that share the same fragment (e.g., *RENT*: —*SE*; *HOUSE, LEASE*). The association frequency of the competitive fragment completions ranged from .03 to .69 ($M = .35$).

We randomly created nine sublists of six quadruplets each from this pool of 54 quadruplets. The study list was composed of six sublists, one for each of the six within-participants conditions. Quadruplets on two sublists were used as fillers in the distraction phase, and those on the ninth sublist was used as primacy and recency buffers during the study phase. The assignment of specific sublist to each of these options was rotated across participants. We also counterbalanced which set of targets was used in the study list and which was used in the distraction phase for the interference condition.

Procedure. The experiment was similar to the previous experiments, with study, practice, distractor, and final-test phases. The study and practice phase were identical to the ones used in Experiment 1, except that a longer, 36-item list was used.

In the distraction phase, participants were asked to read a second list of 36 related word pairs that included 12 pairs of each of the following: (a) *interfering pairs* (e.g., *KNEE–BEND*), which shared the same cue word and target fragment (e.g., B-N-) with a corresponding first-list pairs (e.g., *KNEE–BONE*) but had a different target word; (b) *filler pairs*, which were unrelated to any of the pairs from the first list; and (c) *repeated pairs,* which were identical to pairs from the first list. The repetition condition was added to support the interference manipulation: If the second list had included only interfering and filler items, participants could have used the rule "if a target has been presented in the previous [distraction] phase, it is not the first-list target that I am searching for" to limit interference. The cover story and instructions were the same as in the previous experiments. To equate the overall duration of the distraction phase—that is, the retention interval—across all three experiments despite the change in list length from 24 items in Experiment 1 and 2 to 36 items in Experiment 3, we used only two cycles of reading the second list in this experiment, resulting in 8.4 min of distraction. A different random order was used in each cycle.

The final test phase was identical to the one used in the easy test condition in Experiment 1. Participants were cued with a cue word and fragments of the associated target word from List 1 and asked to recall the first-list target.

Results and Discussion

On the basis of the results of the previous two experiments, no benefit of testing over restudying was predicted in the filler condition. According to our distribution-based framework, however, the benefit of testing was expected to emerge in the interference condition.

Practice-phase performance. Overall, participants recalled .82 of the studied targets during the practice phase. Performance did not differ significantly as a function of final-test condition ($F < 1$).

Final-phase performance. The proportions of targets recalled correctly on the final test are shown in Figure 4 as a function of practice condition (restudy or test) and distraction condition (repetition, filler, or interference). Final test performance was subjected to a 2 (practice condition: repeated study vs. repeated test) × 3 (distraction condition: repetition, filler, or interference) within-participants analysis of variance.

The different distraction conditions were designed to represent different levels of test difficulty, and indeed that was the case overall, $F(2, 70) = 33.63$, $MSE = .06$, $p < .001$, $\eta_p^2 = .49$. Tukey post hoc tests revealed that repetition of pairs during the distraction phase resulted in the highest level of recall (.92) and interference resulted in the lowest level of recall (.62), with the filler condition falling in between (.88; $ps < .05$).

The comparisons of key interest involve the filler (AB-DE) and interference (AB-AC) conditions. As predicted by the distribution model, the relative benefits of testing and restudying

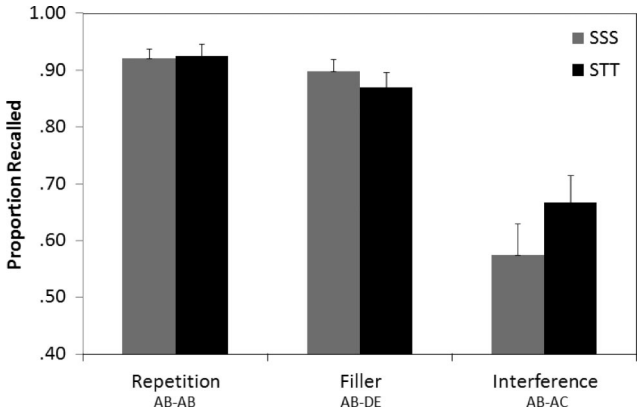

Figure 4. Mean proportion of targets recalled on the final test as a function of practice condition (repeated study, SSS/repeated test, STT) and distraction condition (repetition/interference/filler) in Experiment 3. Error bars represent standard error of the means.

interacted with degree of retroactive interference prior to the final test, $F(2, 70) = 4.33$, $MSE = .02$, $p < .05$, $\eta_p^2 = .11$. In the interference condition there was an advantage of initial testing over restudying (.67 vs. .57), $t(35) = 2.41$, $p < .05$, $d = 0.41$, whereas in the filler condition there was a nonsignificant difference in favor of the restudy condition (.87 vs. .90), $t(35) = -0.92$, $p = .36$.

As mentioned before, the repetition (AB-AB) condition, which was added only to support the interference manipulation, is not of primary interest. One might have expected, though, given prior evidence that tests potentiate subsequent learning (e.g., Izawa, 1970; Kornell, Hays, & Bjork, 2009), to see an advantage of prior testing in this condition. No such significant effect was obtained (.92 vs. .93 for repeated study and repeated test conditions, respectively), $t(35) = .18$, $p = .86$. This is perhaps not surprising, given that the levels of recall are so close to the ceiling.

The effect of retroactive interference on memory for tested information. To examine further the benefits of testing in insulating memories from retroactive interference, we compared performance on the second of the initial tests during the practice phase with performance on the final test, administered after the distraction (second list) phase. In the repetition condition (AB-AB), performance on the final test was expected to benefit from the additional study of those items during the distractor phase. Participants indeed recalled significantly more items in this condition on the final test (.93) than on the initial test (.82), $t(35) = 4.38$, $p < .001$, $d = 0.76$.

Of more interest are the filler (A-B, D-E) and interference (A-B, A-C) conditions. A 2 (test phase: practice vs. final) $\times$ 2 (distraction condition: interference vs. filler) within-participants analysis of variance revealed that the difference between initial test performance and final-test performance interacted with level of retroactive interference induced by the second list, $F(1, 35) = 16.65$, $MSE = .02$, $p < .001$, $\eta_p^2 = .32$. There was a significant decrease in recall from the second practice trial to the final test in the interference condition (.83 vs. .67), $t(35) = 4.02$, $p < .001$, $d = 0.75$, whereas there was an increase in the filler condition (.84 vs. .87), $t(35) = -2.24$, $p < .05$, $d = 0.40$, which might be simply attributed to practice with the task demands (i.e., typing in the responses within the 6-s time frame).

Thus, tested items did suffer from retroactive interference: 83% of the tested pairs remained available on the practice test, but only 67% of the tested items were recallable on the final test, resulting in a forgetting rate of 19% following retroactive interference ($100 - [67/83] \times 100$). However, restudied items suffered much more from retroactive interference: From the participants' point of view, 100% of the pairs were present during the practice trials in the restudy condition, but then only 57% of those pairs could be recalled on the final test, resulting in a forgetting rate of 43% following retroactive interference. Therefore, although there is no absolute benefit of testing in insulating memories from the negative consequences of retroactive interference (as repeated testing did not make memory for the List 1 pairs immune from retroactive interference altogether), there is a substantial relative benefit of testing over restudying in doing so (as testing protect those pairs from interference much better than did restudying).

General Discussion

We have referred to the model in Figure 1 as a framework, interpretation, or model because it is not a theory in the process-model sense. The model simply assumes that a successful test results in larger increase in the subsequent accessibility of the retrieved item from memory than does restudying that item.[3] That assumption, together with the heterogeneity of items within and across participants, has produced the interactions of testing/restudying with final-test delay and final-test format that have been reported in the literature. The model says that final-test difficulty by itself moderates whether initial testing exhibits a benefit over initial restudying. Thus, the model predicts that any way of making the final test more difficult, such as by increasing retroactive interference (tested in Experiment 3), should increase the relative benefits of initial testing and that final-test difficulty, not the overlap of initial-test and final-test formats, should increase the benefits of testing versus restudying (tested in Experiments 1 and 2). The results of Experiments 1, 2, and 3 confirm those predictions.

Overlap of Initial-Test and Final-Test Formats

In Experiments 1 and 2, an advantage of testing over restudying was observed when the final test was a test of free recall but was not observed (Experiment 1) or was reversed (Experiment 2) when the final test was a test of cued recall, even though the initial-test format was cued recall in both experiments. This finding is consistent with previous work that suggested that final-test format moderates the testing effect (e.g., Hogan & Kintsch, 1971), but the interpretations given those prior findings differ from the present interpretation. Thus, for example, Hogan and Kintsch (1971) explained their findings that recall tests (or short-answer tests) exhibited benefits of testing, whereas recognition tests (or multiple-choice tests) did not, by assuming that recognizability is not enhanced by testing, whereas retrievability does benefit from testing. Our distribution-based framework suggests, instead, that such findings were obtained because recall (or short-answer) tests are more difficult than are recognition (or multiple-choice) tests.

We were able to demonstrate in Experiments 1 and 2, from that standpoint of testing the distribution model, that the relative benefit of testing over restudying can be larger when initial-test and final-test formats differ than when they are the same. That is, an initial cued-recall test, versus restudying, enhanced later recall more on a final free-recall test than it did on a final cued-recall test. As Roediger and Karpicke (2006a) pointed out based on other findings, such as Kang et al.'s (2007) finding that performance on

[3] To directly examine the assumption that successful tests result in larger increase in memory of an item versus restudying that items, we reanalyzed the data from the interference condition in Experiment 3. To avoid any item-selection biases, we computed final-test recall rates following restudy, successful retrieval on practice, and unsuccessful retrieval on practice, for each of the 36 items, across participants. Results indicated that, as predicted, an item was more likely to be recalled on the final test when it was successfully retrieved on the practice test (.77) than when it was restudied on the practice phase (.57), $t(35) = 4.76$, $p < .001$. (This analysis also yielded that for items that were not recalled on practice test, the probability of recall on the final test was .15.)

a final multiple-choice test was fostered more by an initial free-recall test than by an initial multiple-choice test, this finding is problematic for interpretations such as the retrieval-practice interpretation ("Initial retrieval aids a later retrieval to the extent that it constitutes practice for that later retrieval—that is, to the extent that the processes involved in the initial retrieval overlap the processes required to retrieve that item later"; Bjork, 1988, p. 397) or the transfer-appropriate-processing hypothesis ("Performance on a final test should be best when that test has the same format as a previous test"; Roediger & Karpicke, 2006a, p. 200).

It may well be, as argued by Schmidt and Bjork (1992) and by Roediger and Karpicke (2006a), that, in Schmidt and Bjork's words, "the overlap of relevant processes does not necessarily mean that there is overlap of the objective conditions of performance" (1992, p. 215), but predicting the results of the present Experiments 1 and 2 then requires additional assumptions that are not required by our distribution framework. It would seem that, ultimately, the overlap of initial-test and final-test conditions has to matter, and certain findings in the generation-effects literature (e.g., deWinstanley, Bjork, & Bjork, 1996) suggest that such overlap may matter substantially under some circumstances. However, the distribution model can explain a range of findings without appealing to such an assumption.

Retroactive Interference and Testing Effects

In Experiment 3 we found, as predicted by the distribution framework, that degree of retroactive interference can determine whether there is or is not a benefit of initial testing over restudying. When an interfering task was given after the practice and before the final test, a benefit of testing over restudying as a practice activity was observed, whereas no such benefit was observed with a less interfering intervening task.

These results, which suggest that when competing materials are studied, tests can help to insulate the tested material against retroactive interference from subsequent competing materials, are consistent with those of other studies demonstrating that testing can protect information from proactive interference (Allen & Arbak, 1976; Arkes & Lyons, 1979; Darley & Murdock, 1971; Robbins & Irvin, 1976; Szpunar, McDermott, & Roediger, 2008; Tulving & Watkins, 1974). In the Szpunar et al. (2008, Experiment 3) study, for example, participants studied five lists of words, one at a time, expecting a final cumulative recall test. In one condition, a 1-min free-recall test was given after each list. In a second condition, a test was given only after List 5, and additional 1-min study time was given for each of Lists 1–4. In a third condition, a test was given only after List 5. Memory for List 5 words, both immediately and after a 30-min delay, was better when participants were tested on Lists 1–4, and there were fewer intrusions. Taken together, the present findings and the prior proactive-interference findings suggest that prior testing helps to distinguish tested information from competing information.

Other Implications of the Distribution Framework

As mentioned in the introduction, a range of existing findings (e.g., Whitten & Bjork, 1977) suggests that as an initial test is made more involved or difficult, the learning benefit for the items successfully retrieved becomes larger. Within the distribution framework, as depicted in Figure 1, that translates to saying that fewer items are shifted by means of successful retrieval on a more difficult initial test, but they are shifted more than they would have been by being retrieved on an easier initial test. The framework predicts, therefore, that there should be an interaction of initial-test and final-test difficulty: Relative to the benefits resulting from an easier initial test, the benefits of a more difficult initial test should be larger on a more difficult final test than they are on an easier final test. Results from an experiment by Hofacker (1982)—in which he covaried the delay between an initial study trial and a first test and the delay between the first test and a second test—support that prediction.

Comparing the results of Experiment 1 and 2 sheds light on related implication of the interaction between initial-test and final-test difficulty: As the practice test was made more difficult from Experiment 1 to Experiment 2, the benefit of restudying over testing on an easy final test increased. According to the distribution framework, this pattern reflects the fact that fewer items were successfully recalled—and, therefore, strengthened—on the more difficult practice test. Therefore, to the extent that a final test is easy enough such that more restudied items are recalled on it than items that were successfully retrieved on the practice test, the benefit of restudying over testing should be larger for the more difficult practice test. More systematic investigation is needed, though, to test this implication.

Another implication of the distribution framework, one that has been tested by Kornell, Bjork, and Garcia (2010), is that the frequent assertion that testing retards the subsequent forgetting rate for items retrieved on that test may be wrong or at least require modification. From the standpoint of the distribution model depicted in Figure 1, performance on a subsequent recall test reflects the proportion of items that remain above the threshold for that test, not memory strength per se. Previously restudied items will tend to be distributed normally. Previously retrieved items that are above threshold will tend to be farther above threshold than are previously restudied items. With the passage of time, the previously retrieved items will tend to remain above threshold as more restudied items cross the threshold toward retrieval failure. Thus, even if the two types of items lose strength at the same rate, the previously retrieved items will appear to be forgotten more slowly. The present findings and those of Kornell et al. (2010) converge on the conclusion that it is important to consider the role of item-distributions when understanding testing effects.

Concluding Comment

The results of the present experiments are consistent with the growing body of research demonstrating that tests not only measure learning but are also potent learning events. At a more detailed level, the results also suggest that effects of initial test and restudy events can interact with the distribution of memory strengths across items in a given experiment and with the difficulty of the final, criterion test. At a practical level, however, the message for teachers and students is more straightforward, if also not simple. In general, when there is a fixed amount of time that can be spent restudying or testing, the more difficult the anticipated criterion test, the more initial testing should be chosen. The present results suggest, though, that an assessment of test difficulty must take into account not only how the test will be formatted but

also when it will be administered and how much the intervening activities are likely to reduce the accessibility (i.e., cause forgetting) of the studied information.

References

Allen, G. A., & Arbak, C. J. (1976). The priority effect in the A-B, A-C paradigm and subjects' expectations. *Journal of Verbal Learning and Verbal Behavior, 15,* 381–385. doi:10.1016/S0022-5371(76)90033-5

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 and Verbal Behavior, 8,* 463–470. doi:10.1016/S0022-5371(69)80090-3

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. doi:10.1037/0278-7393.20.5.1063

Anderson, R. C., & Myrow, D. L. (1971). Retroactive inhibition of meaningful discourse. *Journal of Educational Psychology, 62,* 81–94. doi:10.1037/h0030774

Arkes, H. R., & Lyons, D. J. (1979). A mediational explanation of the priority effect. *Journal of Verbal Learning and Verbal Behavior, 18,* 721–731. doi:10.1016/S0022-5371(79)90425-0

Auble, P. M., & Franks, J. J. (1978). The effects of effort toward comprehension on recall. *Memory & Cognition, 6,* 20–25. doi:10.3758/BF03197424

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127,* 55–68. doi:10.1037/0096-3445.127.1.55

Birnbaum, I. M., & Eichner, J. T. (1971). Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 10,* 516–521. doi:10.1016/S0022-5371(71)80023-3

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

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

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

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., Hofacker, C., & Burns, M. J. (1981, November). *An "effectiveness-ratio" measure of tests as learning events.* Paper presented at the meeting of the Psychonomic Society, Philadelphia, PA.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18,* 49–57. doi:10.1080/09658210903405737

Darley, C. F., & Murdock, B. B., Jr. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology, 91,* 66–73. doi:10.1037/h0031836

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Human memory* (pp. 197–236). San Diego, CA: Academic Press.

deWinstanley, P. A., Bjork, E. L., & Bjork, R. A. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory, 4,* 31–48. doi:10.1080/741940667

Duchastel, P. C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology, 6,* 217–226. doi:10.1016/0361-476X(81)90002-3

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research, 75,* 309–313.

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

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216. doi:10.3758/BF03198098

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40).

Gelfand, H., Bjork, R. A., & Kovacs, K. E. (1983, November). *Retrieval as a recognition-memory modifier: A distribution-based theory.* Paper presented at the meeting of the Psychonomic Society, San Diego, CA.

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

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review, 17,* 797–801.

Hofacker, C. F. (1982). *Some measurement properties of human memory* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (AAT 8306054)

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

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

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667. doi:10.1016/S0022-5371(78)90393-6

Jacoby, L. L. (1996). Dissociating automatic and consciously controlled effects of study/test compatibility. *Journal of Memory and Language, 35,* 32–52. doi:10.1006/jmla.1996.0002

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology, 19,* 528–558. doi:10.1080/09541440601056620

Kornell, N., Bjork, R. A., & Garcia, M. (2010). *Why tests appear to prevent forgetting: A distribution-based bifurcation model.* Unpublished manuscript.

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

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109,* 451–464. doi:10.2307/1423016

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. doi:10.1037/0278-7393.11.2.371

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

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

Robbins, D., & Irvin, J. R. (1976). The priority effect: Test effects on negative transfer and control lists. *Bulletin of the Psychonomic Society, 8,* 167–168.

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., & Blaxton, T. A. (1987). Retrieval modes produce

dissociations in memory for surface information. In D. Gorfein & R. R. Hoffman (Eds.), *Memory and cognitive process: The Ebbinghaus Centennial Conference* (pp. 349–379). Hillsdale, NJ: Erlbaum.

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

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11,* 641–650. doi:10.3758/BF03198289

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217. doi:10.1111/j.1467-9280.1992.tb00029.x

Szpunar, K. K., McDermott, K. B., & Roediger, H. L., III (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1392–1399. doi:10.1037/a0013082

Thomas, A. K., & McDaniel, M. A. (2007). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition, 35,* 668–678. doi:10.3758/BF03193305

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 210–221. doi:10.1037/0278-7393.4.3.210

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: Questions and answers. *Experimental Psychology, 56,* 252–257. doi:10.1027/1618-3169.56.4.252

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6,* 175–184. doi:10.1016/S0022-5371(67)80092-6

Tulving, E., & Watkins, M. J. (1974). On negative transfer: Effects of testing one list on the recall of another. *Journal of Verbal Learning and Verbal Behavior, 13,* 181–193. doi:10.1016/S0022-5371(74)80043-5

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory, 11,* 571–580. doi:10.1080/09658210244000414

Wheeler, M. A., & Roediger, H. L., III. (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

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38,* 995–1008.

Received August 24, 2010
Revision received January 31, 2011
Accepted February 7, 2011 ■