Journal of Experimental Psychology:
Learning, Memory, and Cognition
2009, Vol. 35, No. 6, 1563–1569

0278-7393/09/$12.00 DOI: 10.1037/a0017021

Cue Strength as a Moderator of the Testing Effect: The Benefits of Elaborative Retrieval

Shana K. Carpenter
Iowa State University

The current study explored the elaborative retrieval hypothesis as an explanation for the testing effect: the tendency for a memory test to enhance retention more than restudying. In particular, the retrieval process during testing may activate elaborative information related to the target response, thereby increasing the chances that activation of any of this information will facilitate later retrieval of the target. In a test of this view, participants learned cue–target pairs, which were strongly associated (e.g., *Toast: Bread*) or weakly associated (e.g., *Basket: Bread*), through either a cued recall test (*Toast:* _____) or a restudy opportunity (*Toast: Bread*). A final test requiring free recall of the targets revealed that tested items were retained better than restudied items, and although strong cues facilitated recall of tested items initially, items recalled from weak cues were retained better over time, such that this advantage was eliminated or reversed at the time of the final test. Restudied items were retained at similar rates on the final test regardless of the strength of the cue–target relationship. These results indicate that the activation of elaborative information—which would occur to a greater extent during testing than restudying—may be one mechanism that underlies the testing effect.

Keywords: testing effect, retrieval practice, cued recall, desirable difficulties

Many studies have shown that recalling information on a test leads to significant enhancements in memory retention over simply restudying the material (e.g., Karpicke & Roediger, 2008). This phenomenon—that is, the testing effect—has been observed for a wide variety of materials, including word lists (e.g., Carpenter & DeLosh, 2006), general knowledge facts (e.g., Carpenter, Pashler, Wixted, & Vul, 2008), foreign language vocabulary (e.g., Carrier & Pashler, 1992), and text passages (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008). The effect has also been demonstrated with face–name pairings (e.g., Carpenter & DeLosh, 2005; Landauer & Bjork, 1978) and with visuospatial information (Carpenter & Pashler, 2007).

The robust benefits of testing have led many researchers to advocate more frequent use of testing to promote students' retention (e.g., Bjork, 1988; Dempster, 1996; McDaniel, Roediger, & McDermott, 2007; Pashler, Rohrer, Cepeda, & Carpenter, 2007). Recent studies have supported this notion by demonstrating reliable testing effects in simulated classroom environments (e.g., Butler & Roediger, 2007) and in actual classroom environments where retention of course material was improved as a result of learning information through quizzes (e.g., Carpenter, Pashler, & Cepeda, 2009; McDaniel, Anderson, Derbish, & Morrisette, 2007).

The results of these studies converge on the notion that tests can be effective learning tools. How best to capitalize on these advantages, however, will depend on a thorough understanding of the mechanisms responsible for the effect. Research on this topic has not yet revealed a clear explanation for why tests are beneficial, so further studies pursuing the theoretical nature of the testing effect are clearly needed. Many studies have already shown that retrieving information on a test leads to benefits over and above simply rereading the material (e.g., Carpenter & DeLosh, 2005, 2006; Carrier & Pashler, 1992; Roediger & Karpicke, 2006a, 2006b). There is something uniquely beneficial about taking a test, therefore, as opposed to engaging in additional reading.

One explanation that has been offered is based on the notion of transfer-appropriate processing (e.g., Morris, Bransford, & Franks, 1977). By its very nature, an initial test bears more resemblance to a final test than does a restudy opportunity. Perhaps it is the similarity between initial and final tests, therefore, that drives the testing effect. If so, one might expect retention to be greatest under conditions in which the initial test and final test are more similar. Studies that have tested this rather straightforward prediction have so far failed to find strong support for it, however. Carpenter and DeLosh (2006) gave initial and final tests that required recognition, cued recall, or free recall. They arranged the final test conditions such that some items experienced matching initial and final tests (e.g., recognition–recognition) and others experienced mismatching initial and final tests (e.g., recognition-cued recall, free recall–recognition). Contrary to a transfer-appropriate processing view, final test performance was not best under conditions in which the initial test and final test were more similar. Instead, it was best for items that were given a free-recall initial test, regardless of the type of final test. Other studies have replicated this general pattern (e.g., Glover, 1989; Kang et al., 2007; see also Carpenter, Pashler, & Vul, 2006), suggesting that the testing effect is not driven primarily by the similarity between initial and final

Shana K. Carpenter, Department of Psychology, Iowa State University.

Portions of this study were presented at the annual meeting of the Midwestern Psychological Association, Chicago, Illinois, April–May, 2009. I thank Matthew Erdman and Jessica Kloeppel for their assistance with data collection.

Correspondence concerning this article should be addressed to Shana K. Carpenter, Department of Psychology, Iowa State University, W112 Lagomarcino Hall, Ames, IA 50011-3180. E-mail: shacarp@iastate.edu

tests but instead by retrieval processes taking place during the initial test.

Roediger and Karpicke (2006a) discussed two possibilities as to how retrieval during an initial test could benefit retention. One is that retrieval helps differentiate items that are well learned from those that are not. The success or failure of a retrieval attempt reveals which items require further study, and this could lead to better use of encoding strategies that are applied to the items that need it most. Testing could therefore have a mediating benefit on retention by influencing how well learners process the material during their next opportunity to study it (see also Izawa, 1992).

Even in cases where learners do not have subsequent opportunities to study the material after trying to retrieve it, however, testing produces significant benefits. Many studies have shown that retrieving information on a test is significantly more beneficial than restudying it, even when the initial retrieval attempt is not accompanied by corrective feedback or additional exposure to the material (e.g., Carpenter & DeLosh, 2005; Cull, 2000; Kuo & Hirshman, 1996; Roediger & Karpicke, 2006b). Roediger and Karpicke (2006a) referred to this as the direct benefit of testing (i.e., the notion that the retrieval process itself produces significant benefits on retention).

What might be going on during the retrieval process itself that accounts for these benefits? Progress toward answering this question can be made by asking what kind of processing occurs during retrieval that is not likely to occur during restudying. Whereas restudying involves reading of information that is currently presented, the act of retrieval requires a process by which the contents of memory are activated in an effort to search for some target piece of information that is *not* presented (e.g., Anderson & Bower, 1972; Bahrick, 1970). In line with spreading activation theories of memory (e.g., Collins & Quillian, 1972), the information activated during retrieval may spread to other related concepts and eventually activate an elaborative semantic network with multiple pathways leading to the correct target. As an example, Anderson (1976) described how the cue–target pair *dog–chair* can be learned by thinking about a dog who loved his master but also loved to sit on his master's chairs. One day the dog climbed onto his master's black velvet chair and was scolded for leaving his white hairs all over it. As a result of generation of this elaborative structure, activation that would have been confined to *dog* and *chair* is now spread to various other concepts (e.g., *love*, *master*, *sit*, *scold*, *hairs*, *black*, *velvet*), resulting in multiple pathways between *dog* and *chair* (e.g., *dog* → *love* → *master* → *chair*; *dog* → *scold* → *master* → *chair*; *dog* → *sit* → *chair*; *dog* → *climb* → *chair*). The future recall of *chair* is now facilitated through activation of any of these concepts (e.g., *love*, *master*, *scold*, *sit*, *climb*). Generating this elaborative structure is therefore beneficial to future retention because it provides more information that is capable of activating the target (e.g., Anderson, 1983).

Such elaboration seems more likely to occur during retrieval than during restudy (given that *chair* is immediately available during restudy), and this could explain why retrieval is more beneficial. Furthermore, the degree of elaboration during retrieval would seem to be greater under conditions in which the target is less accessible. This could account for the fact that information is often retained better when it was harder to retrieve initially. For example, retention is often superior for information that was initially tested with longer as opposed to shorter time intervals in between the presentation of information and the initial test (e.g., Carpenter & DeLosh, 2005; Karpicke & Roediger, 2007; Whitten & Bjork, 1977), for initial test conditions that promote interference (e.g., Cuddy & Jacoby, 1982), and for initial tests that require recall as opposed to recognition (e.g., Carpenter & DeLosh, 2006; Kang et al., 2007). Recently, Pyc and Rawson (2009) provided more direct evidence that retention of information is best when initial retrieval is difficult but successful.

The current study explored whether the testing effect can be explained by this elaborative retrieval hypothesis. A cued recall paradigm was used in which feedback was not provided after initial recall. This ensured that any effects would be the direct result of retrieval and not the result of postretrieval encoding strategies. Participants learned cue–target pairs (e.g., *Toast: Bread*) through either a cued recall test (e.g., *Toast:* _____) or a restudy opportunity (e.g., *Toast: Bread*). The likelihood of successfully recalling the target—and hence, of activating more elaborative information during retrieval—was manipulated by varying the strength of the cue–target association. Participants learned items that were either strongly associated (e.g., *Toast: Bread*) or weakly associated (e.g., *Basket: Bread*). Initial retrieval was expected to be higher for targets recalled from strong cues than from weak cues. However, because initial retrieval is less direct with weak cues, recalling a target from a weak cue is more likely to involve the activation of more elaborative information (e.g., *Basket* → *Eggs* → *Flour* → *Bread*) than is recalling a target from a strong cue (e.g., *Toast* → *Bread*). If this information is beneficial for retention, items recalled from weak cues should be retained better over time, such that the initial advantage for strong cues is eliminated or reversed at the time of the final test.

Method

Two experiments were conducted in which participants learned strongly associated versus weakly associated cue–target pairs through either testing or restudying and then received a final free-recall test over the target items. Experiment 1 manipulated trial type (test vs. study) and cue strength (strong vs. weak) within subjects and ensured that initial recall would be relatively easy for both strong and weak cues. Experiment 2 replicated the same design under conditions in which both variables were manipulated between subjects and initial recall was more difficult. Regardless of whether variables are manipulated within or between subjects and whether the overall rate of recall is initially high or relatively lower, the elaborative retrieval hypothesis predicts that targets initially recalled from weak cues should be better retained over time than targets initially recalled from strong cues. If elaborative activation is not as likely to occur during restudy, final test recall of items presented for restudy would not be expected to vary in either experiment as a function of cue–target relatedness.

Participants

A total of 212 undergraduate students participated to fulfill partial course requirements for introductory psychology courses at Iowa State University. There were 60 participants in Experiment 1, with 15 each randomly assigned to one of four counterbalancing conditions. There were 152 participants in Experiment 2, with 38

each randomly assigned to one of four experimental conditions. Participants were tested individually on personal computers.

Materials

Wilson's (1988) database was used to select 48 English nouns with a frequency of greater than 20 per million and concreteness ratings between 500 and 700. Each noun had between five and seven letters and between one and three syllables. The norms of Nelson, McEvoy, and Schreiber (1998) were then used to obtain two different cues for each of the 48 target items. Strong cues had an average cue-to-target strength (i.e., the probability of producing the target given the cue) of .33, and weak cues had an average cue-to-target strength of .01. A complete list of the target items, cues, and cue-to-target strength is reported in the Appendix. In Experiment 1 the items were arranged into six lists containing eight items each, whereas in Experiment 2 all 48 items appeared in a single list.

Design and Procedure

Both experiments began with an encoding phase in which participants were presented with each cue–target pair and were asked to rate the degree of relatedness between the two words. The cue and target appeared in separate boxes in the center of the computer screen, with the cue on the left and the target on the right. The target always appeared in bold, underlined font. Under these boxes, a 5-point rating scale appeared in which 1 indicated "not related," 3 indicated "somewhat related," and 5 indicated "highly related." Participants were asked to type in a number indicating the degree of relatedness between the two words and to use this rating to help them remember the underlined word. As soon as a number was pressed, the next cue–target pair appeared, along with the same rating scale.

After the last item on the list was encoded (i.e., the eighth item in Experiment 1 and the 48th item in Experiment 2), participants completed a distractor task that involved adding together single-digit numbers presented at a rate of one per second. This task lasted approximately fifteen seconds in Experiment 1 and two minutes in Experiment 2. At the end of this time period, participants were asked to type the sum total of the numbers onto a blank screen. Then, these items were presented again in the form of a test or study trial. During test, the cue appeared in a box in the center of the computer screen, with a blank box immediately to its right in which participants were asked to type in the underlined word that was previously paired with it. Participants were given unlimited time to enter their response, and feedback was not provided. As soon as participants entered their response and pressed the *ENTER* key, the cue for the next item appeared. During study, each cue–target pair was presented again and participants rated the relatedness between the two words, just as they had during encoding. Each item was presented once during encoding and then once again as either a test or study trial. Practice lists were provided at the beginning of both experiments to demonstrate the procedure.

In Experiment 1, three of the six experimental lists consisted of test items and three consisted of study items. Half of the cues on each list were weak, and half were strong. Four counterbalancing conditions were created to ensure that each target item occurred equally often as test or study and equally often with a strong versus weak cue. The order in which items were presented within a list was always randomized and different for each participant. The order of the lists themselves was also pseudorandomized, such that each participant saw the lists in the same relative order but always began at a random starting point (e.g., one participant might see Lists 2, 3, 4, 5, 6, and 1, whereas another participant might see Lists 4, 5, 6, 1, 2, and 3). In Experiment 2, participants encountered a single, randomly ordered list of all 48 word pairs that contained test items with strong cues (Group 1), test items with weak cues (Group 2), study items with strong cues (Group 3), or study items with weak cues (Group 4).

After all items had been encoded and then presented as either test or study trials, participants saw a blank screen numbered 1–50 and were asked to type in as many names of U.S. states as they could. After 5 min, participants were shown a new screen of instructions asking them to type in all of the underlined words they could remember from the entire experiment. Participants typed each word onto the computer screen, and after they pressed the *ENTER* key, the response disappeared from the screen to allow another response to be entered. Participants were given 10 min to complete the final test, and feedback was not provided. After 10 min, participants were debriefed and dismissed. The entire procedure for both experiments lasted approximately 35 min.

Results

Initial Ratings

The initial 1–5 ratings given to each cue–target pair during encoding were calculated as a function of trial type (test vs. study) and cue strength (strong vs. weak). The average ratings across all conditions in both experiments can be found in Table 1. Participants gave higher ratings for strong cues than for weak cues, and these ratings were consistent across both test and study items (see Table 1 for analyses).

Table 1
Mean Relatedness Ratings (on a 1–5 Scale) Given to Cue–Target Pairs as a Function of Trial Type and Cue Strength

	Trial type		
Cue strength	Test	Study (first rating)	Study (second rating)
Experiment 1			
Strong	4.26 (0.53)	4.31 (0.44)	4.33 (0.45)
Weak	3.58 (0.66)	3.62 (0.67)	3.62 (0.68)
Experiment 2			
Strong	4.02 (0.48)	4.04 (0.61)	4.04 (0.60)
Weak	3.48 (0.43)	3.52 (0.61)	3.55 (0.63)

Note. Standard deviations are given in parentheses. The only significant effect emerging from a 2 × 2 (Trial Type × Cue Strength) analysis of variance was a main effect of cue strength (for the within-subject analysis in Experiment 1, $F[1, 59] = 150.22$, $p < .001$, $MSE = 0.18$, and for the between-subjects analysis in Experiment 2, $F[1, 148] = 37.20$, $p < .001$, $MSE = 0.29$), confirming that strong cues received higher ratings than weak cues. Average ratings for study items did not differ between initial encoding (first rating) and the restudy opportunity (second rating) in either experiment, and during restudy, participants continued to give higher ratings for strong cues than for weak cues (in Experiment 1, $t[59] = 11.85$, $p < .001$, and in Experiment 2, $t[74] = 3.51$, $p < .01$).

Initial Test Performance

Accuracy of recall on the initial test was calculated as a function of whether targets were recalled from strong or weak cues. In Experiment 1, targets recalled from strong cues (M = .96, SD = .09) were at a significant advantage over targets recalled from weak cues (M = .91, SD = .14), $t(59) = 4.87$, $p < .001$. The overall rate of recall was relatively lower in Experiment 2, but the advantage of strong cues (M = .88, SD = .17) over weak cues (M = .75, SD = .17) was significant, $t(74) = 3.23$, $p < .01$.

Table 2 displays the response times during the initial test for strong versus weak cues. As expected, participants took longer to recall targets from weak cues than from strong cues. This difference was significant in Experiment 1 whether response times were based on all items, $t(59) = 4.58$, $p < .001$, or just items correctly recalled, $t(59) = 3.67$, $p < .01$. In Experiment 2, this difference was significant when response times were based on all items, $t(74) = 2.09$, $p < .05$, but did not reach significance when response times were based on correct items ($t = 0.81$).

Final Test Performance

Accuracy of recall on the final test was calculated for target items that had been recalled from strong versus weak cues. Figure 1 displays initial test accuracy and final test accuracy for these items in Experiment 1 (upper panel) and Experiment 2 (lower panel). Across both experiments, strong cues led to an initial advantage in recall, but items recalled from weak cues were actually retained better over time, such that the advantage for strong cues was eliminated or reversed at the time of the final test. This interaction was significant in Experiment 1, $F(1, 56) = 36.13$, $p < .001$, $MSE = 0.01$, and in Experiment 2, $F(1, 74) = 15.16$, $p < .001$, $MSE = 0.01$.

Both experiments revealed a robust testing effect. In Experiment 1, only 17% of items learned through study were correctly recalled on the final test, whereas 36% of items learned through test—more than double that of study—were correctly recalled. Similarly, in Experiment 2, 18% of items learned through study were correctly recalled, compared to 32% of items learned through test. Table 3 displays the mean proportion of target items recalled on the final test for all conditions in both experiments.

A 2 × 2 × 4 (Trial Type × Cue Strength × Counterbalancing Condition) mixed analysis of variance (with counterbalancing condition as the between-subjects factor) revealed that the testing benefit was significant in Experiment 1, $F(1, 56) = 112.88$, $p < .001$, $MSE = 0.02$, and a 2 × 2 (Trial Type × Cue Strength) between-subjects analysis of variance revealed the same effect for Experiment 2, $F(1, 148) = 53.94$, $p < .001$, $MSE = 0.02$. In both experiments, items learned through study were retained at approximately equal rates on the final test whether they were learned in the context of strong or weak cues (in Experiment 1, 17% vs. 17%, respectively, and in Experiment 2, 19% vs. 17%, respectively). Although the final test data for Experiment 1 revealed a significant Trial Type × Cue Strength interaction, such that weak cues led to significantly better recall of test items on the final test, $F(1, 56) = 10.24$, $p < .01$, $MSE = 0.01$, this interaction was not significant in Experiment 2 ($F = 0.82$).[1] Counterbalancing condition did not affect final test recall in Experiment 1, nor did it interact with any variables (Fs < 2).

Table 2
Mean Response Times (in Milliseconds) During Initial Recall of Test Items as a Function of Cue Strength

	Type of response	
Cue strength	All responses	Correct responses
Experiment 1		
Strong	3,136.83 (1,076.06)	3,070.76 (1,079.9)
Weak	3,600.38 (1,388.22)	3,384.30 (1,162.73)
Experiment 2		
Strong	3,346.75 (1,215.99)	3,159.97 (1,147.44)
Weak	3,844.91 (823.50)	3,330.02 (595.35)

Note. Standard deviations are given in parentheses.

Discussion

The present study demonstrated significant benefits of testing on memory retention. This finding replicates reports of similar benefits from many previous studies (e.g., Roediger & Karpicke, 2006b). Even when compared to an equivalent number of restudy opportunities, recalling information on a test produced superior retention later on. This finding is consistent with a number of studies that have demonstrated significant testing benefits even after controlling for the amount of exposure to the material by including a restudy condition (e.g., Carpenter & DeLosh, 2005, 2006; Carrier & Pashler, 1992; Roediger & Karpicke, 2006a).

Typically, a restudy condition involves presenting the material again for participants to read, without any overt response required. In such a condition, it can be difficult to know the degree to which participants are processing the material or even paying attention to it at all. The current study helps rule out this potential problem by including a restudy condition that required participants to judge the relatedness between two words in a pair. The consistency of the ratings across the encoding and restudy trials indicates that participants were processing the items well enough to make relevant and accurate judgments that were in line with what would be expected.

[1] This result was expected, given that initial test recall was rather low for weak cues in Experiment 2. It has been well documented that the benefits of testing generally apply only to those items that were successfully retrieved on the initial test (e.g., Bjork, 1988; Carrier & Pashler, 1992; Kuo & Hirshman, 1996; Runquist, 1983), and so the advantage of weak cues over strong cues at the time of the final test was probably underestimated because fewer items recalled from weak cues (75%) than from strong cues (88%) would benefit from testing. Some confirmation for this notion is provided when the data are reanalyzed according to a method, advocated by Runquist (1983), that involves calculating the proportion of items correctly recalled on the final test out of those that were recalled at the time of the initial test. When the final test data from both experiments are reanalyzed in this way, both experiments reveal an interaction whereby test items are recalled better after having been learned in the context of weak cues compared to strong cues (44% vs. 32% in Experiment 1 and 41% vs. 35% in Experiment 2), and study items are recalled at about the same rate regardless of whether they were learned in the context of weak cues or strong cues (17% vs. 17% in Experiment 1 and 17% vs. 19% in Experiment 2). This interaction was significant in Experiment 1, $F(1, 56) = 13.31$, $p < .01$, $MSE = 0.02$, and in Experiment 2, $F(1, 148) = 4.06$, $p < .05$, $MSE = 0.02$.

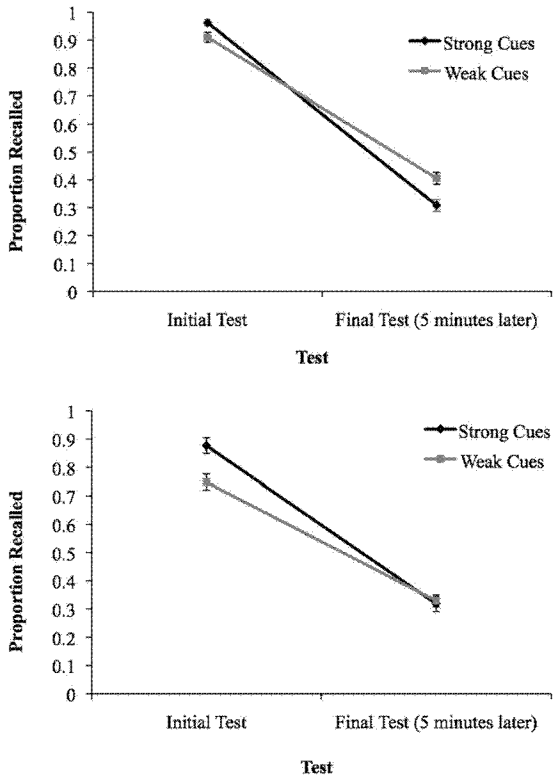

Figure 1. Proportion of target items recalled on the initial test and final test in Experiment 1 (upper panel) and Experiment 2 (lower panel). Target items (e.g., *Bread*) were presented in the context of a strong cue (*Toast: Bread*) or a weak cue (*Basket: Bread*) and given an initial cued recall test, followed by a 5-min delayed free-recall test. Error bars represent standard errors.

This indicates that even when participants actively process and respond to items during restudy, those items are not retained as well as items that had been recalled on a test.

Both experiments revealed an interaction whereby strong cues were advantageous for initial recall, but items recalled from weak cues were retained better over time. This pattern held regardless of whether the overall rate of initial recall was high or relatively lower and whether variables were manipulated within or between subjects (Experiments 1 and 2, respectively). One reason for this may be that strong cues do not encourage the activation of elaborative information that benefits later retrieval through spreading activation (e.g., Anderson, 1983; Collins & Quillian, 1972). In the case of a strong cue (*Toast:* _____), a target item (*Bread*) comes to mind quickly and easily, as evidenced by the faster and more accurate recall of targets from strong cues than from weak cues in both experiments. This quick and easy access may produce detrimental effects later on, however, by reducing the likelihood of activating more elaborative information that is helpful for retention. The likelihood of activating this elaborative information seems more likely with weak cues than strong cues, and that could account for the fact that items recalled from weak cues were retained better over time.

Consistent with this idea, previous studies have shown that memory retention often benefits by activating more elaborative information that is related to the target. In a study of semantic generation, for example, Soraci et al. (1994) asked participants to generate an item using a congruent cue (e.g., *An article of clothing: C _ P*) versus an incongruent cue (e.g., *NOT a policeman: C _ P*). Later memory for the correct response (*CAP*) was better when it had been generated from an incongruent cue, even though this was more likely to involve activation of the wrong response first (*COP*). Along similar lines, Hirshman and Bjork (1988) found that retention of items generated from semantic associates was better if the item (e.g., *Hammer*) had been generated from a third-order associate (*Tool:* _____) rather than a first-order associate (*Nail:* _____). Soraci et al. suggested that extraneous information that is activated during generation of a target response can provide retrieval cues to aid in the future recall of that response.

Prior studies on the testing effect are also consistent with the notion that information activated during retrieval can facilitate later retention. For example, Chan, McDermott, and Roediger (2006) found that an initial test can produce significant benefits on retention of related information that never appeared on the test. Another study (Carpenter & DeLosh, 2006) found that retention of target items (e.g., *street*) recalled through an initial cued recall test was best when the test provided only one letter cue (*s _ _ _ _ _*), followed by two letter cues (*s t _ _ _ _*), three letter cues (*s t r _ _ _*), and then four letter cues (*s t r e _ _*). One-letter cues would bring to mind the largest number of potential retrieval candidates (e.g., "*s _ _ _ _ _*" activates a larger number of candidates than "*s t r e _ _*"). The fact that retention was superior in the one-letter cue condition suggests that the activation of additional target candidates could play a beneficial role in retention of the target.

According to the elaborative retrieval hypothesis, the activation of such information is less likely to occur during restudy and so final test recall of restudied items would not be expected to benefit as a function of manipulations that encourage elaborative processing during retrieval. The current results confirm this prediction by demonstrating that final test recall of study items was unaffected by the strength of the cue–target relationship. These results point to elaborative activation during retrieval as one mechanism that may help to explain why retrieval is beneficial. Future research would benefit from further explorations of elaboration during retrieval and how it contributes to later retention.

Table 3
Proportion of Items Correctly Recalled on Final Test as a Function of Trial Type and Cue Strength

	Trial type		
Cue strength	Test	Study	Total
Experiment 1			
Strong	.31 (.16)	.17 (.11)	.24 (.11)
Weak	.41 (.17)	.17 (.14)	.29 (.12)
Total	.36 (.14)	.17 (.11)	
Experiment 2			
Strong	.32 (.16)	.19 (.09)	.25 (.15)
Weak	.33 (.13)	.17 (.10)	.25 (.14)
Total	.32 (.12)	.18 (.12)	

Note. Standard deviations are given in parentheses.

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22,* 861–876.

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22,* 261–295.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79,* 97–123.

Bahrick, H. A. (1970). Two-phase model for prompted recall. *Psychological Review, 77,* 215–222.

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* (pp. 397–401). New York, NY: Academic Press.

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

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

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34,* 268–276.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review, 14,* 474–478.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th-grade students' retention of U.S. history facts. *Applied Cognitive Psychology, 23,* 760–771.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review, 13,* 826–830.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36,* 438–448.

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

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135,* 553–571.

Collins, A. M., & Quillian, M. R. (1972). Experiments on semantic memory and language comprehension. In L. Gregg (Ed.), *Cognition and learning* (pp. 117–138). New York, NY: Wiley.

Cuddy, L. J., & Jacoby, L. L. (1982). When forgetting helps memory: An analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior, 21,* 451–467.

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

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

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

Hirshman, E., & Bjork, R. A. (1988). The generation effect: Support for a two-factor theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 484–494.

Izawa, C. (1992). Test trial contributions to optimization of learning processes: Study/test trials interactions. In A. F. Healy & S. M. Kosslyn (Eds.), *Essays in honor of William K. Estes: From learning theory to connectionist theory* (Vol. 1, pp. 1–33). Hillsdale, NJ: Erlbaum.

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

Karpicke, J. D., & Roediger, H. L., III. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 704–719.

Karpicke, J. D., & Roediger, H. L., III. (2008, June 27). The critical importance of retrieval for learning. *Science, 319,* 966–968.

Kuo, T., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology, 109,* 451–464.

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

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

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

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519–533.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* Retrieved from http://w3.usf.edu/FreeAssociation/

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14,* 187–193.

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.

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

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

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

Soraci, S. A., Jr., Franks, J. J., Bransford, J. D., Chechile, R. A., Belli, R. F., Carr, M., & Carlin, M. (1994). Incongruous item generation effects: A multiple-cue perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 67–78.

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

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioural Research Methods, Instruments, and Computers, 20,* 6–11.

Appendix

Table A1 shows experimental items used in the current study. Cue-to-target strength according to the norms of Nelson et al. (1998) is indicated in parentheses next to each cue. In Experiment 1, items were presented in six 8-item lists, as shown below. In Experiment 2, items were presented in a single list of all 48 items.

Table A1
Experimental Items Used in the Current Study

Strong cue	Weak cue	Target
List 1		
Toast (.364)	Basket (.014)	Bread
Jury (.250)	Rights (.015)	Court
Valentine (.423)	Rib (.014)	Heart
Rodeo (.477)	Camel (.014)	Horse
Stereo (.333)	Theater (.014)	Music
Neptune (.399)	Comet (.014)	Planet
Chimney (.24)	Fire (.018)	Smoke
Dentist (.459)	Lips (.014)	Teeth
List 2		
Chunk (.054)	Chisel (.01)	Block
Adolescent (.262)	Mitten (.011)	Child
Tea (.369)	Steam (.014)	Coffee
Patient (.365)	Virus (.013)	Doctor
Mow (.275)	Picnic (.014)	Grass
Fork (.37)	Kitchen (.015)	Knife
Cost (.418)	Contest (.015)	Money
Station (.083)	Airplane (.01)	Train
List 3		
Vein (.384)	Bruise (.013)	Blood
Sphere (.258)	Hole (.016)	Circle
Skirt (.295)	Maid (.011)	Dress
Blossom (.441)	Leaf (.012)	Flower
Suite (.356)	Lounge (.014)	Hotel
Adventure (.295)	Desert (.015)	Island
Switch (.459)	Morning (.014)	Light
Education (.315)	Pupil (.016)	School
List 4		
Bristle (.397)	Flick (.013)	Brush
Beverage (.493)	Ice (.016)	Drink
Mop (.244)	Elevator (.014)	Floor
Home (.333)	Barn (.016)	House
Video (.258)	Television (.014)	Movie
Folder (.322)	Scissors (.012)	Paper
Stream (.321)	Bay (.013)	River
Main (.315)	Directions (.013)	Street
List 5		
Shell (.25)	Raft (.011)	Beach
Couch (.288)	Hammock (.013)	Chair
Alarm (.388)	Dial (.019)	Clock
Prom (.221)	Spin (.013)	Dance
Wilderness (.264)	Meadow (.014)	Forest
Shatter (.412)	Bead (.016)	Glass
Saliva (.262)	Speak (.017)	Mouth
Rock (.269)	Building (.017)	Stone
List 6		
Minister (.349)	Soul (.014)	Church
Lunch (.269)	Manners (.014)	Dinner
Picket (.384)	Barrier (.014)	Fence
Plum (.299)	Seed (.011)	Fruit
Note (.299)	Print (.014)	Letter
Call (.378)	Ear (.014)	Phone
Cuff (.247)	Jacket (.013)	Shirt
Hose (.473)	Mist (.013)	Water

Received December 8, 2008
Revision received May 27, 2009
Accepted May 29, 2009 ■