Memory & Cognition
2010, 38 (7), 951-961
doi:10.3758/MC.38.7.951

Scaffolding feedback to maximize long-term error correction

Bridgid Finn
Washington University, St. Louis, Missouri

and

Janet Metcalfe
Columbia University, New York, New York

Scaffolded feedback was tested against three other feedback presentation methods (standard corrective feedback, minimal feedback, and answer-until-correct multiple-choice feedback) over both short- and long-term retention intervals in order to assess which method would produce the most robust gains in error correction. Scaffolded feedback was a method designed to take advantage of the benefits of retrieval practice by providing incremental hints until the correct answer could be self-generated. In Experiments 1 and 3, on an immediate test, final memory for the correct answer was lowest for questions given minimal feedback, moderate for the answer-until-correct condition, and equally high in the scaffolded feedback condition and the standard feedback condition. However, tests of the maintenance of the corrections over a 30-min delay (Experiment 2) and over a 1-day delay (Experiment 3) demonstrated that scaffolded feedback gave rise to the best memory for the correct answers at a delay.

Memory errors are common. People fail to retrieve information that they have learned, and they retrieve flawed or even false information, often judging that it is correct with great confidence (Butterfield & Metcalfe, 2001, 2006; Roediger & McDermott, 1995, 2000). What, then, is the most effective way to correct memory errors, and what method results in the stability of those corrections over time? One approach that has proven successful has been to present corrective feedback following an error. However, the resilience of the feedback over time and the type of feedback that is most effective in correcting performance have not been extensively tested. Our purpose in the present study was to investigate the durability and effectiveness of error correction following corrective feedback and to contrast a variety of formats of providing feedback to surmise whether there is a method that is more effective than simply presenting the correct answer. *Scaffolded* feedback, in which incremental hints were given until the correct answer could be self-generated, was contrasted with other methods, over both short- and long-term retention intervals, to assess which method would produce the most effective, long-lasting gains in error correction.

Feedback has been shown to have considerable positive benefits for memory performance (R. C. Anderson, Kulhavy, & Andre, 1971; Butler & Roediger, 2007, 2008; Lhyle & Kulhavy, 1987; Pashler, Cepeda, Wixted, & Rohrer, 2005). However, the kind of feedback that is given matters. The first and most elementary finding concerning feedback is that it is usually not sufficient to simply tell the learner whether they were right or wrong (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Moreno, 2004; Pashler et al., 2005). Pashler et al. (2005) showed that for feedback to be processed effectively, it is crucial that the correct answer be conveyed. Their results showed that feedback that only relayed a "correct" or "incorrect" message was ineffectual. Only after getting the correct answer as feedback did the participants show an increase in retention. In a similar set of experiments, Hancock, Stock, and Kulhavy (1992) found that participants spent more time processing feedback that relayed the correct answer than they did for feedback simply indicating whether they had been right or wrong. If time spent is a measure of effort, Hancock et al.'s findings suggest that people allocate more effort and processing resources to feedback when it contains the correct response. Together, these results indicate that feedback is more constructive when it relays the correct answer.

When the correct answer is made available after an error, people are able to integrate that information into memory, as is illustrated by an increased probability of answering correctly on a follow-up test (R. C. Anderson et al., 1971). Corrective feedback appears to work well for errors of commission, errors of omission (Metcalfe & Kornell, 2007; Pashler et al., 2005), and high-confidence errors (Butterfield & Metcalfe, 2001, 2006). Feedback can also strengthen correct answers that were given with low confidence (Butler, Karpicke, & Roediger, 2008).

B. Finn, bridgid.finn@wustl.edu

All corrective feedback, however, is not created equal. A few studies have shown that there are differential benefits depending on how the corrective feedback is presented (Butler, Karpicke, & Roediger, 2007; Lhyle & Kulhavy, 1987; Pashler et al., 2005). These studies draw on depth-of-processing research (Craik & Lockhart, 1972; Craik & Tulving, 1975) showing that memory benefits accompany more active, elaborate processing. The rationale behind these studies was that if there is more elaborative processing of the answer during the presentation of the feedback, a stronger representation of the correct answer should result. Lhyle and Kulhavy wrote, "If feedback functions primarily to correct errors, then it follows that any design characteristic that leads students to process, study, or apprehend the feedback more closely should increase the amount of correction that takes place and ultimately improve criterion performance" (p. 320). In their study, participants read a text, answered a multiple-choice question about the content, and were then given feedback about the correct answer. When the feedback was scrambled and the participants were required to unscramble it, more errors were corrected on the subsequent test than in a condition in which the feedback was presented intact. This outcome occurred in only one of two of the experiments reported in the study, however. According to Lhyle and Kulhavy, rearranging the feedback produced better performance because it was effortful, took more time, and made use of semantic processing—all characteristics of deep, elaborate processing.

Another possibility might be to try to promote retrieval practice (Bjork, 1975) or the use of self-generation of the answer (Jacoby, 1978; Slamecka & Graf, 1978) at the time of feedback, rather than to passively present the answer. Retrieval practice is thought by many to be at the core of the *testing effect*, the finding that taking a test has benefits for memory retention that go beyond the gains obtained from mere presentation of the material (for a review, see Roediger & Karpicke, 2006). The testing effect is closely related to the *generation effect* (Jacoby, 1978; Slamecka & Graf, 1978), the parallel finding that self-generating or retrieving a response leads to better retention and recognition performance for the generated item than does a presentation of the same item (see Carrier & Pashler, 1992, for a discussion of how the generation and testing effects can differ). A recent meta-analysis of over 86 generation-effect studies, in which over 17,000 participants were tested, demonstrated that the benefits of generation to memory are robust and consistent (Bertsch, Pesta, Wiscott, & McDaniel, 2007).

How do the benefits of self-generation arise? One possibility is that when items are generated by the participants, the participants are simply given an additional learning opportunity (Thompson, Wenger, & Bartling, 1978)—the so-called *amount-of-processing hypothesis* (Dempster, 1996, 1997; Roediger & Karpicke, 2006). However, much evidence has shown that an additional study presentation does not enhance retention as much as generating, even when processing time is matched (Allen, Mahler, & Estes, 1969; Carrier & Pashler, 1992; Hogan & Kintsch, 1971; Tulving, 1967; Wenger, Thompson, & Bartling, 1980). A more widely accepted proposal is that the process of retrieval itself aids memory. In contrast to reading a response, retrieval of a response from memory requires more effort and may engender deeper processing and more elaborate or variable encoding, and may strengthen or increase the number of semantic cues or routes available for retrieval of the item from memory (Bjork, 1975; Craik & Lockhart, 1972; Craik & Tulving, 1975; Jacoby, 1978; McDaniel & Masson, 1985; Melton, 1967; Whitten & Bjork, 1977).

So, we might expect that getting the person to generate the correct response after they have made an error would be an effective method of presenting feedback. There are two problems with this method. The first is that having just generated the wrong answer, people are unlikely to be able to generate anything, and generating nothing will not help later recall. The second is that if they do generate something, it is likely to be wrong, which may, in turn, result in enhanced memory for the wrong answer. If it were possible to circumvent these two problems, self-generated or active feedback might be more effective than standard feedback, in which the person is simply given the answer.

Accordingly, Butler et al. (2007) explored the possibility that feedback involving active selection of the correct answer might enhance learning. They tested an answer-until-correct feedback format. In this paradigm, originally developed by Pressey (1926), participants answered multiple-choice questions by selecting from the response options until they chose the correct option. An "incorrect" message followed incorrect responses. Final retention was tested 1 day later. Butler et al. (2007) proposed that in comparison to a passive presentation of the feedback, the answer-until-correct feedback format would serve as a kind of self-generation of the response through self-selection and would thereby enhance learning. The participants would also have the benefit of knowing the correct answer by the end of each question trial. However, in contrast to their predictions, Butler et al.'s (2007) results showed no advantage for items for which the answer-until-correct feedback was given over those items for which the standard correct-answer presentation was given.

There are several reasons that the answer-until-correct procedure may not have been the most favorable format to showcase the benefits of generation-enhanced feedback. First, selection of the correct answer may not require the generation of a response from memory. Selection could be based on familiarity and may not engage the deep memory-enhancing retrieval processes that accompany recollection (Jacoby, 1991; Yonelinas, 2002). Second, multiple-choice tests expose people to incorrect information at the same time as the correct information is presented, similarly to the classic A–B A–C interference paradigms (Barnes & Underwood, 1959) and misinformation effect paradigms (Loftus & Palmer, 1974). Interference from the incorrect responses may compromise memory for the correct answer. Though overall a testing benefit has been shown with multiple-choice tests, the selection of the response lures, or indeed, even the mere exposure of the lures before the correct answer is ultimately chosen, can interfere with memory for the final, correct item (Butler, Marsh, Goode, & Roediger, 2006; Huelser & Marsh, 2006; Marsh, Roediger, Bjork, & Bjork, 2007; Roediger & Marsh, 2005;

Schooler, Foster, & Loftus, 1988). Roediger and Marsh, for example, showed that multiple-choice lures are frequently offered as answers on a follow-up cued-recall test if the lures were chosen during the initial test. These findings suggest that the most advantageous format for presenting feedback may not be multiple choice.

Our interest lay in exploiting the fact that self-generating an answer, in contrast to either reading the answer or selecting it from a set of alternatives, might provide a considerable boost to memory. We wanted to contrast a method for presenting feedback that would make use of the benefits of self-generation with other methods that required less elaborative processing. Accordingly, we tested a method that we call *scaffolded feedback*. Our intention was to use a method that required self-generation, while still ensuring that the correct answer would be produced. With scaffolded feedback, participants made retrieval attempts that were guided by incremental hints. For example, a participant might be asked, "What was the crime committed by those in Dante's lowest level of hell in 'The Inferno'?" If they could not answer the question or if they provided the wrong response, they were given another opportunity. If the next response that they gave was incorrect, they were given the first letter of the answer (e.g., B) and another chance to answer. If they still could not answer, they were given the next letter (e.g., E), and so on, until they answered correctly or the whole answer (e.g., BETRAYAL) had been revealed. Because the participants could use the hints to manage their own memory retrieval, we hoped to engage more active retrieval processes and better attendant memory than those that would be utilized either during an answer-until-correct procedure or during the standard correct-answer presentation.

The scaffolded feedback method borrows from the domain of educational psychology, in which the scaffolding approach has long been regarded as an effective support of learning. In general, scaffolding involves a process of helping students reach goals and solve problems that they could not work out independently but that with some assistance, they are often able to (Wood, Bruner, & Ross, 1976). Typically, the scaffolding process involves more than just presenting the correct solution to a problem. Instead, students may be given hints about the correct response or a new suggestion about how to think about the problem, with the ultimate goal of solving the problem correctly themselves.

Carpenter and DeLosh (2006) used a procedure similar to our scaffolding condition to explore the question of how intervening tests, as compared to repeated study of a list, benefited later free recall tests. Single words were given an initial study, followed by an intervening test or an intervening study trial and then an immediate final free recall test. In the test condition that was similar to our scaffolding condition, the participants were asked to recall the studied items (on the intervening test) and were prompted with one, two, three, or four of the item's first letters. Final free recall was better with test rather than with study as the intervening task. Furthermore, fewer rather than more letter prompts resulted in better final recall. A procedure similar to the scaffolding that we used has also been used to enhance learning in special populations. In the studies of the method of vanishing cues, Glisky and Schacter (1989) and Glisky, Schacter, and Tulving (1986) had amnesic patients first learn new computer terms by seeing the definition (e.g., *to store a program on a disk*) with a fragment of the target (e.g., S____). Increasing letters were given until the patient was able to guess the correct term (e.g., SAVE). On later trials, after they were able to generate correctly with all letter cues, the letters vanished one by one from the target item (so long as the patient maintained perfect performance) until the patient could retrieve the answer with the cue alone. These studies showed that amnesiac patients could learn and retain new information if the number of errors that they produced was minimized.

A possibly unfavorable consequence of scaffolding feedback, as we will use it, is the possibility of unsuccessful retrieval attempts, which could make the conditions of learning errorful, rather than errorless. There are advantages to errorless learning in individuals with memory impairments or learning disabilities (N. D. Anderson & Craik, 2006; Baddeley & Wilson, 1994; Hayman, Macdonald, & Tulving, 1993; Jones & Eayrs, 1992; Sidman & Stoddard, 1967; and see Clare & Jones, 2008, and Kessels & de Haan, 2003, for reviews). But it is not clear that such errorful learning conditions have a detrimental effect in healthy young participants (Kornell, Hays, & Bjork, 2009; Metcalfe & Kornell, 2007; Pashler, Zarow, & Triplett, 2003). Several recent studies indicate that as long as the participant ultimately receives feedback of the correct answer, unsuccessful attempts at retrieving may not harm memory (Kornell et al., 2009; Richland, Kornell, & Kao, 2009).

In the present set of experiments, we contrasted four methods of presenting feedback: (1) standard feedback, in which the correct answer was presented immediately following an error; (2) scaffolded feedback, in which participants were given increasing hints until they answered the question correctly; (3) answer-until-correct multiple-choice feedback; and (4) minimal feedback, in which participants knew the answer was wrong, and they were also given one additional chance to provide the correct answer. We measured error correction over short- and long-term test delays. We tested short-term recall performance with an immediate test in Experiment 1 and longer-term recall performance with a 30-min test delay in Experiment 2. In Experiment 3, we sought to extend and replicate our findings by comparing results from an immediate and a 1-day delayed test, in a within-participants design.

EXPERIMENT 1

Method

Participants. The participants were 24 undergraduates at Columbia University and Barnard College. They participated for course credit or cash.

Materials. The questions were 191 general information questions (e.g., "What is the name of the unit of measure that refers to a six-foot depth of water?"; answer, *fathom*). These items were composed of a subset of the published questions from Nelson and Narens (1980). A number of questions that were in the original pool were no longer relevant or correct and were eliminated from the pool. All correct answers were a single word.

Procedure. The participants were tested individually on computers. The experiment had two test phases: an initial test and a surprise final recall test. During both test phases, the participants answered general information questions. During the final recall test, the participants only answered questions that they had answered incorrectly during the initial test. At the beginning of the experiment, the participants were instructed that they would answer general information questions and indicate their confidence in their answer. They were encouraged to guess if they did not know the answer. The participants were not told about the retest that was to follow their initial test and confidence ratings. In both test phases, a general information question was presented, and the participants typed in their response. There were no restrictions on the amount of time that they could take to answer each question.

During the initial test phase, the participants entered their response and were then asked to indicate their confidence in their response by using a horizontal slider that ranged from *very unsure* on the left end to *very sure* on the right end. The slider bar was set to the middle of the slider at the onset of each question. Confidence ratings were coded along a scale from 0 to 100, with 0 indicating a selection of the lowest limit of the slider, at the *very unsure* end, and 100 indicating a selection of the highest limit, at the *very sure* end.

When the participants' answer was correct, a chime would sound, and the next general information question was presented. If their answer was incorrect, there was no chime, and one of four feedback conditions immediately occurred. The set of four feedback conditions were rerandomized after every four incorrect answers. The four conditions were as follows: standard feedback, scaffolded feedback, answer-until-correct multiple-choice feedback, and minimal feedback. The standard feedback was a presentation of the correct response immediately following the error. The participants could study the feedback for as long as they liked. In the scaffolded feedback condition, the participants were given an opportunity to provide another answer. If the new answer that they provided was not correct, the first letter of the correct answer was presented, and they were given another opportunity to enter the correct answer. This process continued, with one additional letter of the answer presented after each answer attempt, until the participants answered correctly. In the answer-until-correct multiple-choice feedback condition, an array of six options, including the correct answer, was presented, and the participants could choose a new response. The experiment program randomly selected the six options from a set of nine potential options. If the participants' original error was included in the list of six options, that option was replaced with one of the remaining three options, ensuring that there were always six novel alternatives. Upon selection, if the item was incorrect, it turned red for 500 msec, and the participants were asked to try again. All six options remained on the screen. When the correct answer was selected, it turned green for 500 msec, and the experiment moved on. In the minimal feedback condition, after making an error, the participants were given one opportunity to provide another answer. A chime sounded if they answered correctly. They then moved on to the next question.

Immediately following each feedback response, the participants used a slider to specify whether they knew the answer all along. The slider ranged from *That's new to me* on the left end to *I actually knew it all along* on the right end. These judgments were not of focal interest for the present investigation and will not be discussed further. After the participants made this judgment, the next general knowledge question was presented. This process continued until the participants had answered 36 questions incorrectly and received feedback in one of the four conditions. Then, for the final recall phase, those 36 incorrect questions were randomized by the computer, and each cue was presented for test.

Results

Basic data. The participants' mean recall performance (proportion correct) was .28 (SE = .02) during the initial test. The initial confidence in the answers was .36 (SE = .03). Because the feedback condition was determined only after the incorrect response had been made, there was no possible effect of feedback condition on initial test performance, in this or in the experiments that follow. The participants' confidence ratings were postdictive of their initial test performance. The mean gamma correlation between initial confidence ratings and initial recall performance was γ = .77 (SE = .03), which was significantly greater than 0 [$t(23) = 26.76$, $p < .05$].

Performance at feedback. Without corrective feedback, in the minimal feedback condition, error correction rarely occurred. When the participants were asked to supply a new answer after having been told that their answer was incorrect, but without any additional support, they were able to correct only a few of the errors of their own accord, resulting in a proportion of .09 (SE = .02) of errors corrected at feedback in this condition. This was, however, significantly greater than 0 [$t(23) = 4.39$, $p < .01$]. In the scaffolded condition, 82% (SE = 3) of the items were answered correctly before the entire answer had been revealed with the successive letter hints. To examine the average number of hints needed to answer each item correctly for each participant, we computed the average proportion of the answer revealed, since the answers were made up of different numbers of letters. On average, the participants needed a proportion of .56 (SE = .03) of the word revealed before they were able to answer correctly. In the answer-until-correct condition, a proportion of .88 (SE = .03) of the correct answers were selected before they were the only item not yet selected. On average, the participants selected 2.88 (SE = 0.14) incorrect items from the six alternatives presented before they picked the correct answer.

Final test performance. There was a significant effect of feedback condition on final test performance [$F(3,69) = 111.39$, $MS_e = 0.02$, $p < .05$, $\eta_p^2 = .83$]. As is shown in Figure 1, final test performance was best in the scaffolded condition (M = .77, SE = .03) and in the standard feedback condition (M = .73, SE = .04), followed by

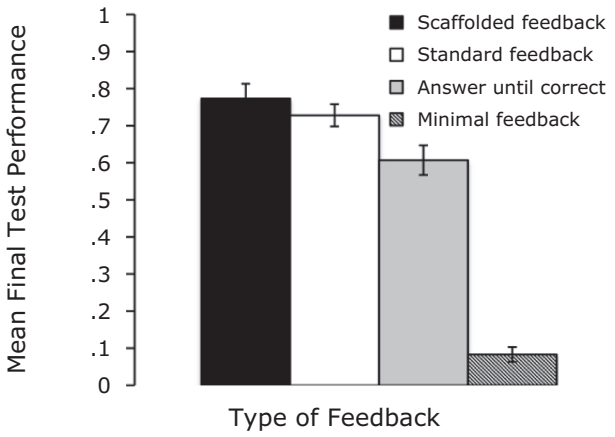

Figure 1. Mean final test performance on an immediate test as a function of type of feedback given to original errors in Experiment 1.

Table 1
Mean Final Test Performance As a Function of Percentage of Answer Revealed in Scaffolded Condition

Percentage of Answer Revealed	Experiment 1		Experiment 2		Experiment 3			
					Immediate		Delayed	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
0	.92	.08	.81	.13	.83	.17	1.00	.00
20	.95	.05	.99	.01	1.00	.00	1.00	.00
40	.92	.05	.85	.09	.89	.11	.78	.15
60	.90	.04	.75	.07	.97	.03	.86	.08
80	.68	.08	.73	.10	.71	.11	.46	.14
100	.44	.09	.45	.07	.60	.11	.46	.11

the answer-until-correct condition ($M = .61$, $SE = .04$). The worst performance was found in the minimal feedback condition ($M = .08$, $SE = .02$). All feedback conditions were significantly different from 0 (all ts > 1, all ps $< .05$). Post hoc pairwise comparisons (which in this and subsequent experiments were Bonferroni corrected) showed that final test performance results for the scaffolded and standard feedback conditions were not significantly different from one another ($t < 1$). Performance in the scaffolded condition was significantly better than that in the answer-until-correct condition [$t(23) = 3.37$, $p < .05$]. The performance difference between the standard feedback and the answer-until-correct conditions was marginally significant [$t(23) = 2.71$, $p = .07$]. Finally, performance in the minimal feedback condition was significantly worse than performance in each of the three other conditions (all ps $< .05$).

Final test performance as a function of the amount of the cue revealed in the scaffolded condition. With the following analyses, we explored the relationship between final retention and the amount of cue revealed. It is possible that performance differences favoring fewer letters could reflect the fact that those items that are recalled with the need for fewer cues during the scaffolding procedure were the easier items, and hence, differences in final recall might be due to that fact alone. (Note that in Carpenter & DeLosh's [2006] third experiment, they varied the number of intervening test cues, rather than allowing them to be participant controlled, and still found an advantage in final free recall for items given fewer cues. An item-selection effect cannot account for this result.) On the other hand, in our experiment, items that required more letters to be revealed were almost certainly studied considerably longer—although we did not formally measure the time—than those that required only a few letters to be remembered correctly, since the letter cues were given one at a time. This study time factor would predict that memory should be better for items that required many cues. In Carpenter and DeLosh's experiment, study time probably followed the opposite pattern, since the participants were asked to retrieve the correct answer when given one, two, three, or four cues, and the effort to retrieve almost certainly took longer with one than with four cues. Thus, time on in their experiment (unlike in ours) would predict a final recall advantage for fewer cues—which is what they found.

Final test performance for items in which the whole word had to be revealed before it was answered correctly ($M = .41$, $SE = .10$) was lower than was final test performance for items that were correctly answered with only part of the word having been revealed ($M = .86$, $SE = .03$) [$t(18) = 4.84$, $p < .01$]. The degrees of freedom in this and subsequent analyses may differ from the total number of participants, because there were some who always answered before the whole word hint had been revealed. Having the whole word revealed in the scaffolded condition resulted in much lower final recall than did seeing the whole word in the standard feedback condition ($M = .71$, $SE = .04$) [$t(19) = 2.95$, $p < .01$]—a result undoubtedly due to the fact that items that required the whole word to be revealed in the scaffolded condition were much more difficult than the random selection of items given whole word feedback in the standard feedback condition.

In a second analysis, similar to that given by Carpenter and DeLosh (2006), items were split into bins of 0%, 20%, 40%, 60%, 80%, and 100% on the basis of the proportion of the answer that had been revealed before a correct guess resulted. Although there were too few participants with a value in each bin to conduct a one-way repeated measures ANOVA, in this or in the analyses that follow, the means are presented in Table 1 for archival purposes. Final test performance showed the first drop when 80% of the word had been revealed. The steepest drop occurred when the entire word had been revealed. Our data, like those of Carpenter and DeLosh, revealed the worst performance when the greatest number of hints were given.

Final performance as a function of the number of multiple-choice options needed. Would a similar pattern to that in the scaffolded condition appear in the answer-until-correct multiple-choice condition—namely, that selection of a correct option, before it was the only remaining option not yet selected would lead to better final test performance than selection of the correct option when it was the last possible option? Again, we present these results for archival purposes and emphasize that the results of this analysis should be interpreted with caution, given the possible item-selection artifacts, as items answered correctly only once all the other items had been chosen were undoubtedly more difficult, a priori, than those items that could be answered before they were the only item not yet selected. The mean final recall performance for questions in which the correct answer had been selected last was .37

Table 2
Mean Final Test Performance As a Function of Order of Correct Selection in Multiple Choice

Ordinal Position	Experiment 1		Experiment 2		Experiment 3			
					Immediate		Delayed	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
First	.80	.07	.65	.07	.83	.09	.76	.08
Second	.66	.10	.48	.09	.96	.04	.46	.14
Third	.45	.11	.55	.10	.80	.13	.29	.16
Fourth	.48	.11	.36	.10	.62	.24	.50	.17
Fifth	.56	.11	.57	.14	.44	.18	.39	.16
Last	.37	.11	.51	.12	.44	.15	.33	.21

($SE = .11$), in comparison to .63 ($SE = .04$) for questions in which the correct answer had been selected before it was the only remaining item. This difference was significant [$t(16) = 2.11, p = .05$]. Again, we created bins on the basis of the number of items that were selected until the correct item had been chosen, giving first, second, third, fourth, fifth, and sixth selection bins. As can be seen in Table 1, the benefit of selecting the correct item appeared to be confined to its having been selected early. Performance dropped to around 45% on and after the third selection (see Table 2).

Discussion

The results of this experiment indicate that providing corrective feedback is important. When no corrections were given, but the participants simply had to try again to come up with the answers, their eventual performance was very poor. Performance was better when the participants received the correct answer by successively guessing in a multiple-choice test, until they got the right answer. But the benefits of this answer-until-correct procedure were not as great as those either when the participants were simply given the answer or when their self-retrieval of the correct answer was scaffolded. However, scaffolding and simply being given the answer did not result in different performance levels. When feedback was scaffolded, the time needed to present the feedback probably increased, as did the effort needed to do the task. But the result was not better performance on the immediate final test. Given that the same proportion of errors were corrected following scaffolded and standard feedback, if the test is immediate, these results indicate that there is no advantage to using the more laborious scaffolding methodology.

One caveat to this conclusion is that the similarity in the effectiveness of the scaffolded and standard feedback conditions might be constrained to an immediate test. One method or the other might have differential long-term consequences. If there were a longer term advantage to the scaffolded method, that might provide a compelling reason to switch to the more intensive method.

There are some indications, from the literature on testing effects, that memorial advantages of retrieval practice may not be different from direct study or simply being presented in the immediate term but may have large effects when testing is delayed. Although being tested rather than restudying can make no difference or can even produce worse performance on an immediate follow-up test, being tested has large beneficial effects on long-term tests (Carpenter, Pashler, Wixted, & Vul, 2008; Carrier & Pashler, 1992; Cull, 2000; Roediger & Karpicke, 2006; Thompson et al., 1978; Wenger et al., 1980; Wheeler, Ewers, & Buonanno, 2003), consistent with what Bjork (1994) called *desirable difficulties*. The retrieval practice involved in testing may make items more resistant to forgetting (Carpenter et al., 2008; Hogan & Kintsch, 1971; Roediger & Karpicke, 2006). Carpenter et al. compared a test with feedback with a study presentation over a range of retention intervals ranging from 5 min to 42 days. They found that the rate of forgetting was less following testing than following restudy. If the retrieval practice in scaffolded feedback is similar to retrieval practice that may be operative in testing, we might find performance benefits to scaffolded feedback over standard feedback, when the criterion test is delayed rather than immediate.

EXPERIMENT 2

In Experiment 2, we investigated test performance at a delay. This allowed us to explore feedback-related differences in the maintenance of the correct information over a longer retention interval. The hypothesis was that scaffolded feedback would benefit retention more at a delay than would standard feedback.

Method

The participants were 25 undergraduates at Columbia University and Barnard College. They participated for course credit or cash. The design, materials, and procedure in Experiment 2 were identical those in Experiment 1, except that the final test came after a half-hour delay instead of immediately following questions and feedback. The half-hour delay was filled with an unrelated experiment.

Results

Basic data. The participants' mean recall performance was a proportion of .22 ($SE = .03$) correct during the initial test, and their initial confidence in their answers was .37 ($SE = .03$). The participants' confidence ratings were postdictive of their initial test performance. The mean gamma correlation between initial confidence ratings and initial recall performance was $\gamma = .72$ ($SE = .04$) and was significantly greater than 0 [$t(23) = 18.56, p < .05$].

Performance at feedback. As in Experiment 1, in the minimal feedback condition, error correction was rare, resulting in a proportion of .04 ($SE = .02$) of errors cor-

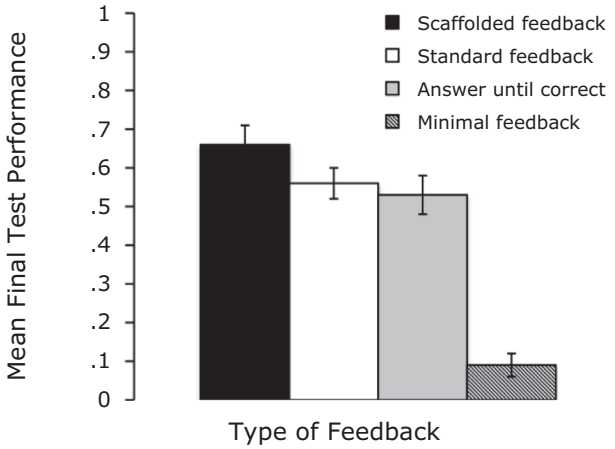

Figure 2. Mean final test performance on a test delayed by half an hour as a function of type of feedback given to original errors in Experiment 2.

rected at the time of feedback ($t > 1$, $p < .01$). In the scaffolded condition, a proportion of .59 ($SE = .05$) of the items were answered correctly before the entire answer had been revealed with hints. On average, the participants needed a proportion of .68 ($SE = .03$) of the word revealed before they were able to answer correctly. In the answer-until-correct multiple-choice condition, a proportion of .91 ($SE = .05$) of the correct answers were selected before they were the only item not yet selected. On average, the participants selected 2.68 ($SE = 0.13$) incorrect items until they picked the correct answer.

Final test performance. There was a significant effect of feedback condition on final test performance [$F(3,72) = 49.46$, $MS_e = 0.03$, $p < .05$, $\eta_p^2 = .67$]. The mean final test performance for each of the conditions was as follows and is shown in Figure 2: The scaffolded condition ($M = .66$, $SE = .05$) was followed by the answer-until-correct condition ($M = .56$, $SE = .04$), then the standard feedback condition ($M = .53$, $SE = .05$), and finally the minimal feedback condition ($M = .09$, $SE = .03$). Planned comparisons revealed significant differences between the scaffolded and the standard feedback conditions [$t(24) = 2.46$, $p < .05$] and between the scaffolded and the answer-until-correct conditions [$t(24) = 2.31$, $p < .05$], with the scaffolded condition showing superior performance across the delay. Performance following answer-until-correct and standard feedback conditions was equivalent ($t < 1$). All feedback condition comparisons with the minimal feedback condition showed significant differences (all ps $< .05$). The performance scores in all of the feedback conditions were significantly different from 0 (all ps $< .05$).

Final test performance as a function of the amount of the cue revealed in the scaffolded condition. When the participants were given the entire word, performance on the delayed test was significantly lower ($M = .47$, $SE = .07$) than when they were able to answer it with only partial cues ($M = .72$, $SE = .06$) [$t(23) = 2.99$, $p < .05$]. Delayed final test performance for items given the whole word answer was not significantly worse than the performance for items given standard feedback ($M = .53$; $t < 1$, $p > .05$). Performance showed the largest drop after 100% of the word had been revealed (see Table 1).

Final performance as a function of the number of multiple-choice options needed. Recall performance was the same for questions in which the correct answer had been selected last ($M = .51$, $SE = .07$) and for questions in which the correct answer had been selected before it was the only remaining item ($M = .51$, $SE = .12$) ($t < 1$). An analysis of performance using bins created on the basis of the number of items that had been selected until the correct item had been chosen showed that there appeared to be no performance benefit on the delayed test for selecting the correct item early (see Table 2).

Discussion

When a follow-up recall test was given at a delay, performance differences between the standard feedback condition and the scaffolded condition emerged. There were more items answered correctly following the scaffolded feedback than following the standard feedback, which did not show significantly better performance than the answer-until-correct format. As in Experiment 1, the minimal feedback condition showed the lowest rate of error correction.

Because Experiments 1 and 2 were run on separate groups of participants at different times in the academic year, it was not appropriate to contrast results from the immediate and delayed tests. To directly compare participants' results from an immediate and a delayed test and to expand our results by using an extended delay, we conducted a final experiment. In Experiment 3, we used a within-participants design to contrast the magnitude of error correction following each of the feedback conditions on an immediate test and on a 1-day delayed test.

EXPERIMENT 3

Method

The participants were 18 undergraduates at Columbia University and Barnard College. They participated for course credit or cash. The experiment was a 2 (test delay: immediate vs. delayed test) × 4 (feedback condition) within-participants design. The materials and procedure of Experiment 3 were identical to those in Experiments 1 and 2, except for two procedural changes that were implemented so that we could test both immediately and at a delay. The first difference was that all of the participants were tested immediately on half of the items and came back either 1 or 2 days later for a delayed test on the remaining half. The mean delay between feedback and the final delayed test was 1.22 days. The items were assigned randomly in equal numbers into the immediate and delayed test conditions. The second change was that the participants answered questions until they had attained 40 incorrect answers.

Results

Basic data. The participants' mean recall performance on the initial test was a proportion of a proportion of .30 ($SE = .02$) correct during the initial test. Initial confidence in answers given was .39 ($SE = .03$). The participants' confidence ratings were postdictive of their initial test performance. The mean gamma correlation between ini-

tial confidence ratings and initial recall performance was $\gamma = .76$ ($SE = .03$) and was significantly greater than 0 [$t(17) = 29.18$, $p < .05$].

Performance at feedback. As in Experiments 1 and 2, error correction at feedback was rare in the minimal feedback condition ($M = .09$, $SE = .02$; $t > 1$, $p < .05$). There were no significant differences between the immediate and delayed conditions in any of the following feedback performance analyses (ps $> .05$), which was as expected, since the test delay manipulation had not yet been introduced at the time of feedback. In the scaffolded condition, a proportion of .68 ($SE = .04$) of the items were answered correctly before the entire answer had been revealed with hints. On average, the participants needed a proportion of .61 ($SE = .03$) of the word revealed before they were able to answer correctly. In the answer-until-correct multiple-choice condition, a proportion of .90 ($SE = .02$) of the correct answers were selected before they were the only item not yet selected. On average, the participants selected 2.73 ($SE = 0.12$) incorrect items until they picked the correct answer.

Final test performance on the immediate and delayed tests. Mean test performance for each condition on the immediate and delayed tests can be seen in Figure 3. There was a main effect of time of test [$F(1,17) = 25.30$, $MS_e = 0.03$, $p < .05$, $\eta_p^2 = .60$] showing an expected delay-related drop in performance (immediate, $M = .60$, $SE = .04$; delayed, $M = .45$, $SE = .04$). There was also a main effect of feedback condition [$F(3,51) = 68.19$, $MS_e = 0.04$, $p < .05$, $\eta_p^2 = .80$]. The lowest performance was shown in the minimal feedback condition ($M = .11$, $SE = .02$), which was significantly different from that in all other conditions (all ts > 1, all ps $< .05$). There was no overall test performance difference between the other feedback conditions (scaffolded, $M = .72$, $SE = .05$; answer-until-correct, $M = .65$, $SE = .04$; standard feedback, $M = .63$, $SE = .05$) (all ts < 1, all ps $> .05$).

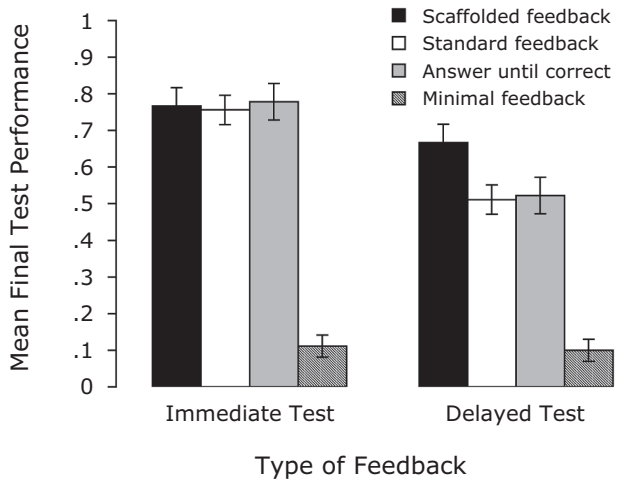

Figure 3. Mean final test performance on an immediate test and on a test delayed by 1 day as a function of type of feedback given to original errors in Experiment 3.

Performance for all feedback conditions was significantly different from 0 (all ps $< .05$).

The main result of interest was the significant time of test $\times$ feedback condition interaction [$F(3,51) = 4.85$, $MS_e = 0.03$, $p < .05$, $\eta_p^2 = .22$]. Performance following the standard feedback and the scaffolded conditions was our central focus. Post hoc tests revealed that there was no significant difference between the standard feedback ($M = .76$, $SE = .07$) and scaffolded ($M = .77$, $SE = .06$) conditions ($t < 1$, $p > .05$) on the immediate test. Importantly, however, the benefits of scaffolded feedback over standard feedback were shown at the delay. Performance on the delayed test showed a significant ($M = .16$) performance advantage for items given scaffolded feedback ($M = .67$, $SE = .06$) over items given standard feedback ($M = .51$, $SE = .06$) [$t(17) = 2.61$, $p < .05$].

On the immediate test, performance following answer-until-correct feedback ($M = .78$, $SE = .04$) was not different from performance following either the standard or the scaffolded feedback (all ts < 1, all ps $> .05$). This result differed from those of Experiment 1, in which performance in the answer-until-correct condition was different from that in the standard and scaffolded conditions. On the delayed test, performance in the answer-until-correct condition ($M = .52$, $SE = .06$) was not different from that in the standard feedback condition ($t < 1$, $p > .05$). The difference between performance in the answer-until-correct condition and that in the scaffolded condition on the delayed test was at significance [$t(17) = 2.06$, $p = .05$]. Performance following the minimal feedback condition was the worst and was significantly different from that in all other conditions on both the immediate and delayed tests (all ts > 1, all ps $< .05$).

Final test performance as a function of the amount of the cue revealed in the scaffolded condition. Performance on the final test was lower ($M = .52$, $SE = .11$) when the whole word was revealed than when only partial cues were revealed ($M = .83$, $SE = .05$) [$F(1,13) = 8.56$, $MS_e = 0.03$, $p < .05$, $\eta_p^2 = .40$]. Neither the effect of time of test nor the time of test $\times$ hint amount interaction was significant ($p > .05$). Items given standard feedback were recalled significantly better than items given the whole word answer in scaffolded feedback ($t > 1$, $p < .05$). Final test performance for the scaffolded condition showed the first drop when about 80% of the word had been revealed (see Table 1).

Final performance as a function of the number of multiple-choice options needed. We could not compute the full 2 (time of test: immediate vs. delayed) $\times$ 2 (hint: partial vs. whole) repeated measures ANOVA for the answer-until-correct multiple-choice condition, because there were only a few participants who had selected the correct answer as their last possible selection in the immediate and delayed test conditions. Collapsing over time of test, we found that recall performance was worse for items in which the correct answer had been selected last ($M = .42$, $SE = .12$) than for questions in which the correct answer had been selected before it was the only remaining item ($M = .72$, $SE = .16$) ($p < .05$). There appeared to

be some benefit at final test for selecting the correct item early (see Table 2).

Discussion

In Experiment 3, we replicated and extended the findings of Experiments 1 and 2. Final test performance was not different between the standard and scaffolded conditions on an immediate test. When the test was delayed over 24 h, however, scaffolded feedback produced more long-lasting gains in error correction than did standard feedback.

GENERAL DISCUSSION

In the set of experiments presented here, we contrasted four methods of presenting corrective feedback. Scaffolding feedback, by giving successive hints but requiring that the participant generate the answer him or herself, took advantage of the benefits of retrieval practice and generation. This method was designed to utilize the deep retrieval processes that are engaged in the intentional generation of a response from memory (Jacoby, 1991; Yonelinas, 2002). In contrast to standard feedback, scaffolded feedback capitalized on the benefits of retrieval attempts on memory, while making certain that the correct answer would be produced. We found on a test conducted immediately after study that errors in the scaffolded feedback condition were corrected at a rate equally high as that in the standard feedback condition, in which the answers were simply provided to partipants. However, when the test was delayed for either a half hour or slightly more than a day, scaffolded feedback led to greater recall than did standard feedback or the answer-until-correct feedback.

With scaffolding, feedback can be flexible and dynamic, allowing calibration of feedback to the knowledge and skills of each student. Student A may not need to be exposed to as many clues as Student B to answer a particular question or to retrieve relevant information from memory. Each student will have different memories, experiences, and domain knowledge, and therefore, each will have different feedback requirements, which can be dynamically adjusted on the basis of the student's current state of knowledge. Items just at the boundary of what the person knows (or what Metcalfe and colleagues have called the *region of proximal learning*; Metcalfe, 2002, in press; Metcalfe & Kornell, 2003, 2005) might be those items that were initially answered incorrectly but, with the benefit of a small amount of scaffolding, could be self-generated correctly. Effective learning can occur because the instructor (even when that instructor is a computer) and student are coordinated (Pea, 2004; Wood et al., 1976).

One limitation of this particular instantiation of scaffolded feedback was the narrow scope of our hints, which only displayed additional letters of the correct answer. A more sophisticated scaffolding system using semantic cues, for example, might produce even better results, although this needs to be tested. Even the very simple scaffolding system that we used, which is very easy to implement, had considerable favorable effects.

Much research has shown that providing students with corrective feedback can improve performance on a follow-up test. Here, we showed that scaffolded corrective feedback resulted in corrections that were more resilient to a delay interval than corrections following either the standard feedback or answer-until-correct multiple-choice formats. Standard feedback is the most effective and efficient method to use if a student only has a few minutes to correct their errors before a test. However, if the goal is long-term knowledge retention, the results presented here indicate that the student will be best served by scaffolded feedback.

AUTHOR NOTE

This research was supported by NIMH Grant RO1MH60637 and by Grant 220020166 from the James S. McDonnell Foundation. We thank the scholars from MetaLab for their help and comments. Correspondence concerning this article should be addressed to B. Finn, Department of Psychology, Washington University, St. Louis, MO 63130 (e-mail: bridgid.finn@wustl.edu).

REFERENCES

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning & Verbal Behavior*, **8**, 463-470. doi:10.1016/S0022-5371(69)80090-3

Anderson, N. D., & Craik, F. I. M. (2006). The mnemonic mechanisms of errorless learning. *Neuropsychologia*, **44**, 2806-2813. doi:10.1016/j.neuropsychologia.2006.05.026

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Research*, **62**, 148-156. doi:10.1037/h0030766

Baddeley, A., & Wilson, B. A. (1994). When implicit learning fails: Amnesia and the problem of error elimination. *Neuropsychologia*, **32**, 53-68. doi:10.1016/0028-3932(94)90068-X

Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, **61**, 213-238.

Barnes, J., & Underwood, B. (1959). "Fate" of first learned associations in transfer theory. *Journal of Experimental Psychology*, **58**, 97-105.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, **35**, 201-210.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Erlbaum.

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.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, **13**, 273-281.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III (2008). Correcting a metacognitive error: Feedback enhances retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **34**, 918-928. doi:10.1037/0278-7393.34.4.918

Butler, A. C., Marsh, E. J., Goode, M. K., & Roediger, H. L., III (2006). When additional multiple-choice lures aid versus hinder later memory. *Applied Cognitive Psychology*, **20**, 941-956. doi:10.1002/acp.1239

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. doi:10.1080/09541440701326097

BUTLER, A. C., & ROEDIGER, H. L., III (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, **36**, 604-616. doi:10.3758/MC.36.3.604

BUTTERFIELD, B., & METCALFE, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **27**, 1491-1494. doi:10.1037/0278-7393.27.6.1491

BUTTERFIELD, B., & METCALFE, J. (2006). The correction of errors committed with high confidence. *Metacognition & Learning*, **1**, 1556-1623. doi:10.1007/s11409-006-6894-z

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., WIXTED, J. T., & VUL, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, **36**, 438-448. doi:10.3758/MC.36.2.438

CARRIER, M., & PASHLER, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, **20**, 632-642.

CLARE, L., & JONES, R. S. P. (2008). Errorless learning in the rehabilitation of memory impairment: A critical review. *Neuropsychology Review*, **18**, 1-23. doi:10.1007/s11065-008-9051-4

CRAIK, F. I. M., & LOCKHART, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning & Verbal Behavior*, **11**, 671-684. doi:10.1016/S0022-5371(72)80001-X

CRAIK, F. I. M., & TULVING, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, **104**, 268-294.

CULL, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, **14**, 215-235. doi:10.1002/(SICI)1099-0720(200005/06)14:3<215::AID-ACP640>3.3.CO;2-T

DEMPSTER, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. C. Carterette & M. P. Friedman (Series Eds.) & E. L. Bjork & R. A. Bjork (Vol. Eds.), *Handbook of perception and cognition: Vol. 10. Memory* (2nd ed., pp. 317-344). San Diego: Academic Press.

DEMPSTER, F. N. (1997). Using tests to promote classroom learning. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 332-346). Westport, CT: Greenwood Press.

GLISKY, E. L., & SCHACTER, D. L. (1989). Extending the limits of complex learning in organic amnesia: Computer training in a vocational domain. *Neuropsychologia*, **27**, 107-120. doi:10.1016/0028-3932(89)90093-6

GLISKY, E. L., SCHACTER, D. L., & TULVING, E. (1986). Computer learning by memory-impaired patients: Acquisition and retention of complex knowledge. *Neuropsychologia*, **24**, 313-328. doi:10.1016/0028-3932(86)90017-5

HANCOCK, T. E., STOCK, W. A., & KULHAVY, R. W. (1992). Predicting feedback effects from response-certitude estimates. *Bulletin of the Psychonomic Society*, **30**, 173-176.

HAYMAN, C. A. G., MACDONALD, C. A., & TULVING, E. (1993). The role of repetition and associative interference in new semantic learning in amnesia: A case experiment. *Journal of Cognitive Neuroscience*, **5**, 375-389. doi:10.1162/jocn.1993.5.4.375

HOGAN, R. M., & KINTSCH, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning & Verbal Behavior*, **10**, 562-567. doi:10.1016/S0022-5371(71)80029-4

HUELSER, B. J., & MARSH, E. J. (2006, November). *Does guessing on a multiple-choice test affect later cued recall?* Poster presented at the 47th Annual Meeting of the Psychonomic Society, Houston.

JACOBY, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning & Verbal Behavior*, **17**, 649-668. doi:10.1016/S0022-5371(78)90393-6

JACOBY, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, **30**, 513-541. doi:10.1016/0749-596X(91)90025-F

JONES, R. S. P., & EAYRS, C. B. (1992). The use of errorless learning procedures in teaching people with a learning disability: A critical review. *Mental Handicap Research*, **5**, 204-212.

KESSELS, R. P. C., & DE HAAN, E. H. F. (2003). Implicit learning in memory rehabilitation: A meta-analysis on errorless learning and vanishing cues methods. *Journal of Clinical & Experimental Neuropsychology*, **25**, 805-814. doi:10.1076/jcen.25.6.805.16474

KORNELL, N., HAYS, M. J., & BJORK, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **35**, 989-998. doi:10.1037/a0015729

LHYLE, K. G., & KULHAVY, R. W. (1987). Feedback processing and error correction. *Journal of Educational Psychology*, **79**, 320-322. doi:10.1037/0022-0663.79.3.320

LOFTUS, E. F., & PALMER, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning & Verbal Behavior*, **13**, 585-589. doi:10.1016/S0022-5371(74)80011-3

MARSH, E. J., ROEDIGER, H. L., III, BJORK, R. A., & BJORK, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, **14**, 194-199.

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

MELTON, A. W. (1967). Repetition and retrieval from memory. *Science*, **158**, 532. doi:10.1126/science.158.3800.532-b

METCALFE, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, **131**, 349-363. doi:10.1037/0096-3445.131.3.349

METCALFE, J. (in press). Desirable difficulties and study in the region of proximal learning. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork*. New York: Psychology Press.

METCALFE, J., & KORNELL, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, **132**, 530-542. doi:10.1037/0096-3445.132.4.530

METCALFE, J., & KORNELL, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory & Language*, **52**, 463-477. doi:10.1016/j.jml.2004.12.001

METCALFE, J., & KORNELL, N. (2007). Principles of cognitive science in education: The effects of generation, errors and feedback. *Psychonomic Bulletin & Review*, **14**, 225-229.

MORENO, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, **32**, 99-113. doi:10.1023/B:TRUC.0000021811.66966.1d

NELSON, T. O., & NARENS, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning & Verbal Behavior*, **19**, 338-368. doi:10.1016/S0022-5371(80)90266-2

PASHLER, H., CEPEDA, N. J., WIXTED, J. T., & ROHRER, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **31**, 3-8. doi:10.1037/0278-7393.31.1.3

PASHLER, H., ZAROW, G., & TRIPLETT, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1051-1057.

PEA, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *Journal of the Learning Sciences*, **13**, 423-451. doi:10.1207/s15327809jls1303_6

PRESSEY, S. L. (1926). A simple apparatus which gives tests and scores and teaches. *School & Society*, **23**, 373-376.

RICHLAND, L. E., KORNELL, N., & KAO, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, **15**, 243-257. doi:10.1037/a0016496

ROEDIGER, H. L., III, & KARPICKE, J. D. (2006). 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., III, & MARSH, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimen-*

tal Psychology: Learning, Memory, & Cognition, **31**, 1155-1159. doi:10.1037/0278-7393.31.5.1155

ROEDIGER, H. L., III, & MCDERMOTT, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 803-814. doi:10.1037/0278-7393.21.4.803

ROEDIGER, H. L., III, & MCDERMOTT, K. B. (2000). Distortions of memory. In F. I. M. Craik & E. Tulving (Eds.), *The Oxford handbook of memory* (pp. 149-164). Oxford: Oxford University Press.

SCHOOLER, J. W., FOSTER, R. A., & LOFTUS, E. E. (1988). Some deleterious consequences of the act of recollection. *Memory & Cognition*, **16**, 243-251.

SIDMAN, M., & STODDARD, L. T. (1967). The effectiveness of fading in programming during a simultaneous form discrimination for retarded children. *Journal of the Experimental Analysis of Behavior*, **10**, 3-15.

SLAMECKA, N. J., & GRAF, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning & Memory*, **4**, 592-604. doi:10.1037/0278-7393.4.6.592

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

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

WENGER, S. K., THOMPSON, C. P., & BARTLING, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 135-144. doi:10.1037/0278-7393.6.2.135

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

WHITTEN, W. B., II, & BJORK, R. A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning & Verbal Behavior*, **16**, 465-478. doi:10.1016/S0022-5371(77)80040-6

WOOD, D., BRUNER, J. S., & ROSS, G. (1976). The role of tutoring and problem solving. *Journal of Child Psychology & Psychiatry*, **17**, 89-100. doi:10.1111/j.1469-7610.1976.tb00381.x

YONELINAS, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory & Language*, **46**, 441-517.

(Manuscript received December 9, 2009;
revision accepted for publication March 7, 2010.)