Journal of Experimental Psychology: Applied
2007, Vol. 13, No. 4, 273–281

1076-898X/07/$12.00 DOI: 10.1037/1076-898X.13.4.273

The Effect of Type and Timing of Feedback on Learning From Multiple-Choice Tests

Andrew C. Butler, Jeffrey D. Karpicke, and Henry L. Roediger III
Washington University in St. Louis, St. Louis, Missouri

Two experiments investigated how the type and timing of feedback influence learning from a multiple-choice test. First, participants read 12 prose passages, which covered various general knowledge topics (e.g., *The Sun*) and ranged between 280 and 300 words in length. Next, they took an initial six-alternative, multiple-choice test on information contained in the passages. Feedback was given immediately for some of the multiple-choice items or after delay for other items. Participants were either shown the correct answer as feedback (standard feedback) or were allowed to keep answering until the correct answer was discovered (answer-until-correct feedback). Learning from the test was assessed on a delayed cued-recall test. The results indicated that delayed feedback led to superior final test performance relative to immediate feedback. However, type of feedback did not matter: discovering the correct answer through answer-until-correct feedback produced equivalent performance relative to standard feedback. This research suggests that delaying the presentation of feedback after a test is beneficial to learning because of the spaced presentation of information.

Keywords: multiple choice, testing, feedback

In 1926, Sidney Pressey introduced a machine that he believed would revolutionize education (Pressey, 1926). Designed for the purpose of administering multiple-choice tests to students, his device featured a mechanism that required students to keep responding to each question until they selected the correct response, referred to as answer-until-correct feedback. More than just an assessment tool, Pressey's machine helped students learn by guiding them to discover the correct response. Importantly, the machine also provided immediate feedback about the accuracy of each response. No longer would students have to wait several days for their teacher to return the corrected exam.

Since Pressey (1926) introduced his teaching machine, other devices that provide immediate, answer-until-correct feedback have appeared in many different forms, including chemically treated answer sheets (Peterson, 1930), punch-boards (Angell & Troyer, 1948), and modified memory drums (Stephens, 1960). Most recently, answer-until-correct feedback has been implemented in a commercial product called the Immediate Feedback Assessment Technique (IF-AT; www.epsteineducation.com). Growing in popularity among educators (DiBattista, 2005), this product consists of a multiple-choice answer sheet with a thin opaque film covering the answer options on which students scratch off the film in order to reveal the correct response. According to the developers of the IF-AT, the efficacy of this technique derives from providing immediate feedback during the test, encouraging the active processing of feedback, and assuring that the last response made for any given question is the correct response (Epstein, Epstein, & Brosvic, 2001).

Our research evaluates how the type of feedback (standard vs. answer-until correct) and the timing of feedback (immediate vs. delayed) affect learning from a multiple-choice test. Most previous research that supports the use of immediate, answer-until-correct feedback has naturally confounded these two variables (e.g., Angell, 1949; Sullivan, Schutz, & Baker, 1971; Dihoff, Brosvic, Epstein, & Cook, 2004; but see Brosvic, Epstein, Cook, & Dihoff, 2005). Many of these studies were designed to compare two practical options for giving feedback in the classroom: the use of a device that provides immediate, answer-until-correct feedback, or the standard method of grading a test and returning it to students after a delay with the correct responses indicated. For example, Dihoff, Brosvic, and Epstein (2003) found a benefit of answer-until-correct feedback compared with standard feedback (i.e., simply presenting students with the correct answer). However, in their study, answer-until-correct feedback was given immediately after each question, whereas standard feedback was given on the following day. In this comparison, the advantage of answer-until-correct feedback could be because of either or both of the two factors that covary in this design: passively reading the answer or actively generating it, and immediate or delayed feedback. Our research was aimed at disentangling these two factors and examining the effects of these feedback conditions on long-term retention.

Standard Versus Answer-Until-Correct Feedback

Across the many different types of feedback that have been studied, the efficacy of feedback in the acquisition of test content

Andrew C. Butler, Jeffrey D. Karpicke, and Henry L. Roediger III, Department of Psychology, Washington University in St. Louis, St. Louis, Missouri.

We thank Tanya Antonini and Patrick Flanagan for their help in collecting data. A Collaborative Activity Award from the James S. McDonnell Foundation and a grant from the Institute of Education Sciences supported this research.

This article was accepted under the editorial term of Wendy A. Rogers.

Correspondence concerning this article should be addressed to Andrew C. Butler, Department of Psychology, Campus Box 1125, Washington University in St. Louis, St. Louis, MO 63130-4899. E-mail: butler@wustl.edu

is largely determined by whether the feedback message contains the correct response (for meta-analyses see Bangert-Drowns, Kulik, C., Kulik, J., & Morgan, 1991; Kluger & DeNisi, 1996). For example, feedback that includes a presentation of the correct response is much more effective than simply indicating that a response is right or wrong (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005). Accordingly, the standard way to present feedback is simply to show the correct answer to students. Answer-until-correct feedback also provides students with the correct response by having them continue responding until they select the correct answer. Thus, the critical difference between these two types of feedback is the process by which students arrive at the correct response: with standard feedback students read the correct answer, whereas with answer-until-correct feedback they must discover it. One way to conceptualize this difference is through active versus passive processing. When students are presented with the correct answers, they are more passive receivers of knowledge. In contrast, when students must discover the correct answer, they are active seekers of knowledge. This difference may be similar to the difference between reading and generating information in the generation effect literature (Jacoby, 1978; Slamecka & Graf, 1978; for review, see Mulligan & Lozito, 2004). If information is generated, it is generally better remembered than if the information is simply read or studied. Although selecting an alternative on a multiple-choice test does not involve the generation of a response in the purest sense, the answer-until-correct feedback procedure does require students to actively reason about why another alternative might be correct. Thus, the process by which students arrive at the correct response when given answer-until-correct feedback may lead to superior retention of that response. Indeed, many educational researchers advocate for the benefits of learning through discovery (e.g., Bruner, 1967).

However, one possible drawback to answer-until-correct feedback is the potential for students to make more than one incorrect response before the correct response is discovered. Previous research has shown that incorrectly selecting lures on a multiple-choice test often leads students to learn and retain those lures on later memory tests (Butler & Roediger, 2007b; Roediger & Marsh, 2005). The selection of two or more lure responses in the course of answering a question may inflate this negative effect by leading students to learn multiple incorrect responses. In addition, selecting multiple responses on a multiple-choice test may increase the number of answers associated to the question in memory, and competition among several associated answers may interfere with the ability to remember the correct response (Anderson, 1974).

Immediate Versus Delayed Feedback

As described earlier, differences between standard and answer-until-correct feedback have typically been confounded with whether feedback occurred immediately or after a delay. In most prior experiments, answer-until-correct feedback is given immediately during the test, while standard feedback is given after a delay. The length of the feedback delay in previous studies has ranged from short delays lasting a few minutes to longer delays of up to several days. Previous research on the optimal timing of feedback has yielded a conflicting body of literature (for review see Kulik & Kulik, 1988), but the main point of contention centers on how the timing of feedback enables learners to correct errors. Originating from behaviorist theories of reinforcement, one idea is that feedback must be given immediately in order to eliminate incorrect responses and reinforce correct responses (e.g., Skinner, 1954). This position predicts that the efficacy of feedback will decrease substantially as the delay before the presentation of feedback increases.

In contrast, proponents of delayed feedback argue that incorrect responses must be allowed to dissipate or they will interfere with the learning of correct responses (e.g., Kulhavy, 1977). In addition, delaying feedback may also benefit the retention of correct responses because of spaced presentation. Information is better retained when learned through repeated presentations that are spaced (or distributed) as opposed to massed (Dempster, 1989; Schmidt & Bjork, 1992). After a correct response, delayed feedback would represent a spaced presentation of the information, whereas immediate feedback would represent a massed presentation. Feedback after correct responses may be very important to learning, a point we return to in the discussion (see also Butler, Karpicke, & Roediger, 2007).

Our Experiments

The goal of our experiments was to disentangle the effects of the type of feedback (standard vs. answer-until-correct) and the timing of feedback (immediate vs. delayed) on long-term retention assessed on a delayed criterial test. In both experiments, participants studied a series of brief prose passages about a variety of educational topics and then took an initial multiple-choice test. In the standard feedback conditions, participants were shown the correct answer, while in the answer-until correct conditions they were required to respond repeatedly until they answered the question correctly. In the immediate feedback conditions, participants received feedback immediately after each question, whereas in the delayed feedback conditions they were given feedback after a delay (10 min in Experiment 1 and 1 day in Experiment 2). As a control, no feedback was given for another subset of questions (test with no feedback), and an additional subset of questions was not tested on the initial multiple-choice test (no test). Participants returned after a delay for a final cued-recall test (1 day in Experiment 1 and 1 week in Experiment 2).

We predicted that taking an initial test would lead to better performance on the final test relative to not taking a test. This phenomenon, known as the *testing effect*, has received considerable renewed interest in recent research (see Butler & Roediger, 2007a; Carrier & Pashler, 1992; Chan, McDermott, & Roediger, 2006; Roediger & Karpicke, 2006a; Roediger & Karpicke, 2006b). We also predicted that testing with feedback would confer even greater benefit than testing without feedback. However, of central importance, we hypothesized that if the process of discovering the correct answer is indeed similar to the act of generating to-be-remembered information, answer-until-correct feedback should enhance long-term retention more than standard feedback. Finally, we expected a benefit of delayed feedback compared with immediate feedback because of the spaced presentation of the feedback.

Experiment 1

Method

Participants. Forty-eight undergraduate students from Washington University participated for either course credit or a payment of $20 and were tested in groups of one to four people.

Design. The experiment consisted of a 2 (timing of feedback: immediate, delayed) x 2 (type of feedback: standard, answer-until-correct) mixed factorial design. Timing of feedback was manipulated between-participants, and type of feedback was manipulated within-participants, between-materials. We also included two control conditions: (1) a subset of items was not tested on the initial multiple-choice test (no test), and (2) another subset of items was tested, but no feedback was given (test with no feedback).

Materials and counterbalancing. Materials consisted of 12 passages taken from GRE, TOEFL, and SAT study guides. Each passage consisted of 280–300 words of text organized into four paragraphs. Four facts were identified in each passage (one fact per paragraph) to serve as the to-be-tested information. For testing purposes, each fact was transformed into a question and correct response (e.g., Question: *In which state was Dorothea Dix born?* Answer: *Maine*), and five plausible incorrect responses were developed to serve as multiple-choice lures. Individual PCs running E-Prime software (Schneider, Eschman, & Zuccolotto, 2002) were used to present all the materials and collect responses.

To counterbalance the materials, different versions of the experiment were created in which the 12 passages were rotated through the four conditions that each participant experienced: the two type of feedback conditions (standard and answer-until-correct) as well as the no test and test with no feedback control conditions. In each version of the experiment, four passages were assigned to each of the type of feedback conditions, and two passages were assigned to the control conditions. In addition, the position of the target relative to lures on the multiple-choice test was counterbalanced such that the target appeared in each of the six possible positions equally often across all the conditions.

Procedure. The experiment consisted of two 1-hour-long sessions on consecutive days. On Day 1, groups of participants were randomly assigned to a between-participants condition (immediate or delayed feedback) on arrival. They studied the 12 passages in a random order determined by the computer. Each passage was presented for 50 seconds (pilot testing showed this amount of time was sufficient to read the entire passage once). Next, they engaged in a distractor task, playing a *Pac-Man* video game, for 5 minutes. The purpose of the distractor task was to ensure that material from the passages could not be recalled from working memory when the multiple-choice test was given. Then, participants took a 6-alternative multiple-choice test that covered the material in the passages. The test included a total of 40 questions: 8 questions for the test with no-feedback condition, 16 for the standard feedback condition, and 16 for the answer-until-correct feedback condition. The additional 8 items were not initially tested because they constituted the no test condition.

Feedback was presented either immediately after a multiple-choice alternative was selected (immediate feedback condition) or after a 10-min distractor task (delayed feedback condition). Participants received standard feedback for some questions and answer-until-correct feedback for other questions. Standard feedback consisted of a 10-s representation of the question along with the response outcome ("correct" or "incorrect"), the selected response, and the correct response. Participants received this feedback message regardless of whether their response was correct or incorrect. Answer-until-correct feedback consisted of a recursive loop in which participants continued to answer the multiple-choice question until they selected the correct response. Whenever the correct response was selected, they received the same feedback message as in the standard condition: a 10-s re-presentation of the question along with the response outcome ("correct"), the selected response, and the correct response. However, when an incorrect response was selected, they received a feedback message with the question, the selected response, and the word "incorrect" for four seconds, after which the initial question screen was represented. For questions in the no-feedback condition, a 10-s filler message ("please wait while the next question loads") followed each response to equate for overall time in the different conditions.

On completion of the multiple-choice test, all participants engaged in another distractor task for 10 minutes (recall of newsworthy events from 2005). For participants in the delayed feedback condition, the feedback was presented after the second distractor task. Delayed feedback was presented in exactly the same manner as the immediate feedback. Finally, the participants were dismissed and asked to return the next day.

On Day 2, participants returned to take a final cued-recall test that contained all 48 questions (40 from the multiple-choice test and 8 not previously tested). The questions were presented one at a time in a random order and participants were required to type a response to each question (either a word or a short phrase), guessing when necessary. This forced report procedure closely resembles the way in which testing is conducted in educational settings. In the classroom, students are rarely penalized for guessing, so they produce a response to every question because a guess may turn out to be the correct response. After finishing the test, participants were debriefed and dismissed.

Results

All results deemed significant were reliable at the .05 level of confidence unless otherwise noted. Pairwise comparisons were Bonferroni-corrected to the .05 level. The data were initially analyzed with counterbalancing condition included as a between-subjects factor; however, this variable did not produce any main effects nor did it interact with any of the variables of interest and was therefore omitted from further analyses. On the initial multiple-choice test, responses were scored as either correct or incorrect. On the final cued-recall test, the potential response outcomes were correct, incorrect-lure (an incorrect answer that was a lure on the initial multiple-choice test), and incorrect-other (all other incorrect responses).

Initial multiple-choice test. On average, participants responded correctly to .64 of the items on the initial multiple-choice test and performance was roughly equivalent in all conditions. A 2×2 repeated-measures analysis of variance (ANOVA) did not reveal any significant differences [type of feedback: $F(1, 46) = .00$, $MSE = .010$, $p = 1.00$; timing of feedback: $F(1, 46) = .18$, $MSE = .029$, $p = .68$; type x timing: $F(1, 46) = .99$, $MSE = .010$, $p = .32$]. In the answer-until-correct condition, .63 of the multiple-choice questions were answered correctly on the first attempt, .12 on the second attempt, .10 on third attempt, .05 on the fourth attempt, .05 on the fifth attempt, and .05 on the sixth attempt.

Final cued recall test: correct recall. Figure 1 shows the proportion of correct responses on the final cued-recall test given one day after learning as a function of timing and type of feedback as well as for the two control conditions (test with no feedback and no test). First, there was a testing effect: Taking

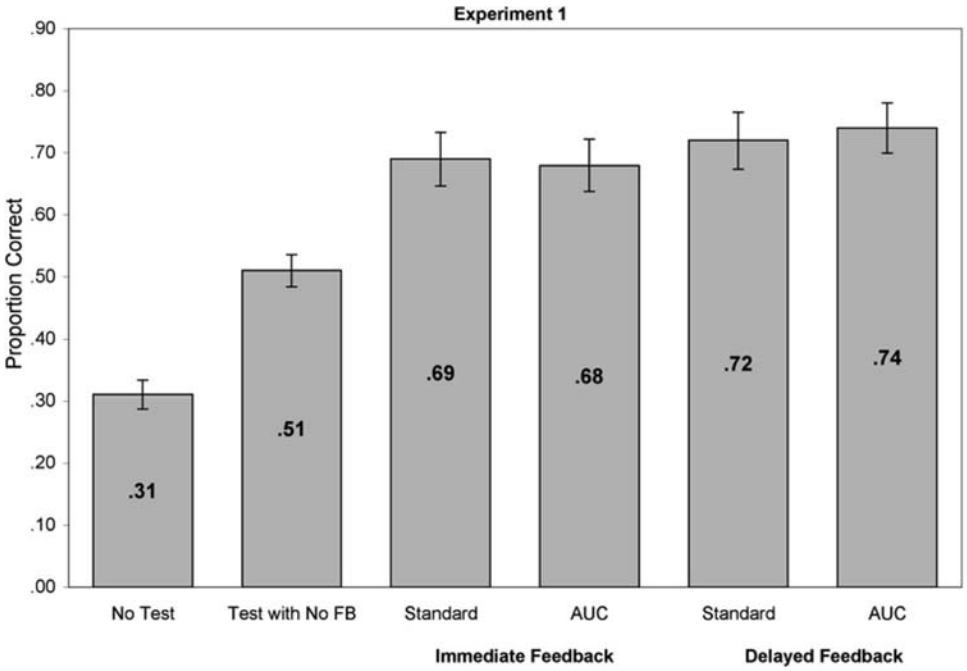

Figure 1. Proportion correct on the final cued recall in Experiment 1 as a function of timing and type of feedback as well as for the two control conditions (test with no feedback and no test). AUC denotes the answer-until-correct feedback condition. Error bars represent standard error of the mean. The abbreviation *FB* refers to feedback.

the initial multiple-choice test enhanced recall on the final test relative to reading the passages but not testing [.51 vs. .31; $t(47) = 13.96$, $SEM = .028$, $d = .77$, $p_{rep} = 1.00$]. (p_{rep} is an estimate of the probability of replicating the direction of an effect; see Killeen, 2005.) In addition, there was a positive effect of feedback on long-term retention. Collapsed across the four feedback conditions, feedback enhanced retention more than testing without feedback [.70 vs. .51; $t(47) = 6.46$, $SEM = .030$, $d = .91$, $p_{rep} = 1.00$]. Finally, there was no effect of the type of feedback: performance in the standard and answer-until-correct feedback conditions was identical (.71 vs. .71). Delayed feedback produced a higher proportion of correct responses than immediate feedback (.73 vs. .68; $d = .28$, $p_{rep} = .61$), and this advantage was apparent in both of the type of feedback groups; however, the result was not statistically significant. A 2×2 repeated-measures ANOVA did not reveal any significant differences [type of feedback: $F(1, 46) = .61$, $MSE = .010$, $p = .44$; timing of feedback: $F(1, 46) = .92$, $MSE = .05$, $p = .34$; type x timing: $F(1, 46) = .07$, $MSE = .010$, $p = .80$].

Final cued recall test: production of incorrect lures. Taking the initial multiple-choice test (without feedback) led to the production of a slightly higher proportion of incorrect-lure responses on the final test relative to not taking the test (.20 vs. .18). Collapsing across feedback conditions, providing feedback after the test substantially reduced the proportion of incorrect-lure responses relative to both the test with no feedback [.09 vs. .20; $t(47) = 5.33$, $SEM = .022$; $d = .72$, $p_{rep} = 1.00$] and the no test [.09 vs. .18; $t(47) = 4.71$, $SEM = .019$, $d = .67$, $p_{rep} = 1.00$] conditions. However, there was virtually no difference between the two types of feedback conditions in reducing errors: standard and answer-until-correct (.08 vs. .09). Likewise, timing of feedback did not have a differential effect as immediate and delayed feedback produced roughly the same proportion of incorrect-lure responses (.08 vs. .09). A 2×2 repeated-measures ANOVA confirmed these observations [type of feedback: $F(1, 46) = .58$, $MSE = .004$, $p = .45$; timing of feedback: $F(1, 46) = .47$, $MSE = .010$, $p = .50$; type x timing: $F(1, 46) = 1.42$, $MSE = .003$, $p = .24$].

Conditional analyses. Final cued recall performance was analyzed as a function of response outcome (correct/incorrect) on the initial multiple-choice test. The conditional analyses were conducted on the aggregated data (i.e., across participants) rather than the alternative method of computing conditionalized means for each individual participant. This method was used to avoid the problem of how to replace or estimate a mean for individual participants when they did not produce any observations in one of the conditionalized cells (e.g., "correct on final cued recall given incorrect on initial multiple-choice").

The critical procedural difference between the answer-until-correct and standard feedback occurred after incorrect responses (feedback was identical in both conditions after correct responses). Thus, any difference in the efficacy of the two types of feedback would be expected to emerge when looking at final test performance for initially incorrect responses. However, the proportion of correct responses on the final cued-recall test given an incorrect response on the multiple-choice test (i.e., wrong answers that were corrected) was nearly equivalent in the standard and answer-until-correct conditions (.53 vs. .52).

Of additional interest was whether the timing of feedback had a differential effect on correct and incorrect responses. The top panel of Figure 2 displays the proportion correct on the final cued-recall

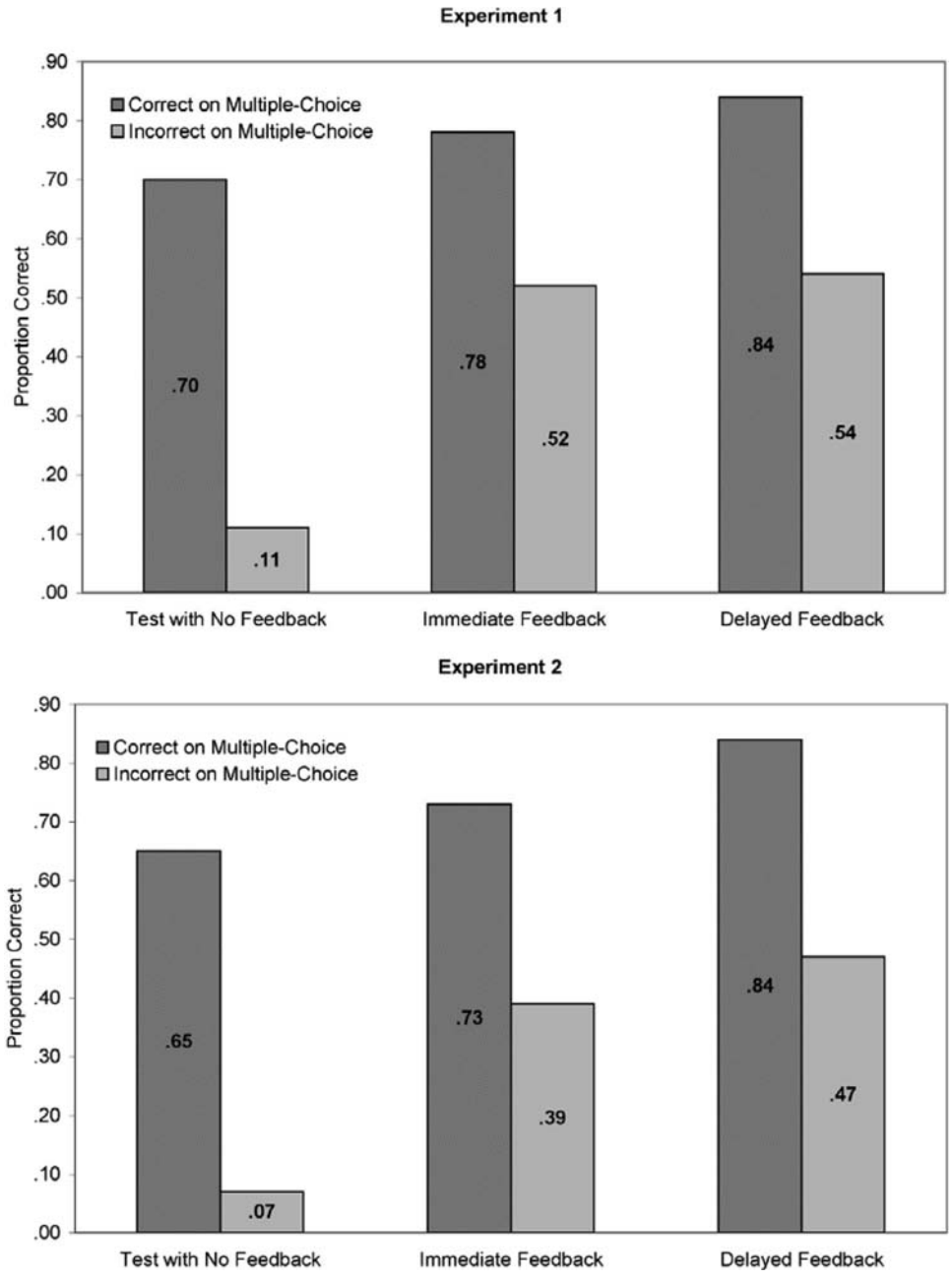

Figure 2. Conditional analysis: Proportion correct on the final cued-recall test as a function of initial test outcome (correct/incorrect) and feedback condition for Experiment 1 (top panel) and Experiment 2 (bottom panel).

test as a function of initial test outcome (correct/incorrect) and feedback condition. Providing feedback led to the correction of roughly half of the initially incorrect responses, whereas only a small proportion of errors was spontaneously corrected without feedback (.11). However, the timing of feedback did not affect the correction of errors. Interestingly, feedback also benefited initially correct responses, perhaps because it helped to confirm the accuracy of some correct guesses that were not maintained in the test with no-feedback condition (Butler, Karpicke, & Roediger, 2007). In addition, delaying feedback boosted performance even higher, which might be thought of as a type of spacing effect. Thus, the .05 difference between delayed and immediate feedback in the main results may be due to the benefit of delayed feedback after correct responses.

Discussion

Experiment 1 showed clear effects of testing and of providing feedback. There was also a numerical advantage of delaying feedback rather than providing it immediately after each item, although

this advantage was small and not statistically significant. However, we did not find a benefit of answer-until-correct feedback compared with standard feedback. The last result is contrary to the findings of previous studies that answer-until-correct feedback represents a superior testing technique.

Experiment 2

Experiment 2 was aimed at replicating the effects observed in Experiment 1 and extending them by lengthening the delay in the delayed feedback condition and by examining long-term retention one week after learning. One potential explanation for the results of Experiment 1 is that the retention interval (1 day) might not have been sufficient to allow any effects of type or timing of feedback to emerge. Recent research has shown that longer intervals of retention are required to reveal some memory effects, such as the testing effect (Roediger & Karpicke, 2006a). To address this issue, Experiment 2 featured a longer retention interval (1 week) as well as longer feedback delay (1 day). The only other change to the procedure involved a switch in the between-participant and within-participant aspects of the design. Type of feedback was manipulated between-participants, and timing of feedback was manipulated within-participants for Experiment 2. This change was made because some of the studies showing a benefit of answer-until-correct feedback used a between-participants design (e.g., Angell, 1949) as well as for the purpose of generalizability. Otherwise, the procedure used in Experiment 2 was the same as that used in Experiment 1.

Method

Participants. Forty Washington University undergraduate students participated for either course credit or a payment of $20 and were tested in groups of one to four people.

Design and materials. The second experiment consisted of the same design as Experiment 1, except that type of feedback was manipulated between-participants and timing of feedback was manipulated within-participants, but between-materials. The same materials were used in Experiment 2.

Procedure. The procedure was the same as in Experiment 1 except for the following changes. First, participants were randomly assigned to one of the two types of feedback conditions (standard or answer-until-correct). Immediate feedback was given for some questions, whereas delayed feedback was given for other questions. Second, delayed feedback was given in a second session that occurred 1 day later. Third, the final test was given in a third session that occurred 1 week after the initial session.

Results

Initial multiple-choice test. Overall, participants responded correctly to .61 of the questions, similar to the initial multiple-choice performance in Experiment 1. Again, a 2 × 2 repeated-measures ANOVA indicated no differences among the conditions [type of feedback: $F(1, 38) = .18$, $MSE = .030$, $p = .68$; timing of feedback: $F(1, 38) = .59$, $MSE = .040$, $p = .45$; type x timing: $F(1, 38) = .43$, $MSE = .030$, $p = .52$]. In the answer-until-correct condition, .64 of the multiple-choice questions were answered correctly on the first attempt, .15 on the second attempt, .07 on the third attempt, .07 on the fourth attempt, .04 on the fifth attempt, and .03 on the sixth attempt.

Final cued recall test: correct recall. Figure 3 shows the proportion of correct responses on the final cued-recall test as a function of timing and type of feedback as well as for the two control conditions (test with no feedback and no test). Again, there was a superiority of previous testing compared with items not tested [.42 vs. .26; $t(39) = 4.99$, $SEM = .032$, $d = .80$, $p_{rep} = 1.00$], and a benefit of providing feedback relative to taking a test without feedback [.65 vs. .42; $t(39) = 7.07$, $SEM = .032$; $d = 1.13$, $p_{rep} = 1.00$]. In addition, delayed feedback led to higher proportion of correct responses on the final test relative to immediate feedback (.70 vs. .60; $d = .47$, $p_{rep} = .93$). However, as in Experiment 1, type of feedback had no effect on the proportion of correct responses produced on the final test (.64 vs. .65). A 2 × 2 repeated-measures ANOVA confirmed these results [type of feedback: $F(1, 38) = .02$, $MSE = .067$, $p = .89$; timing of feedback: $F(1, 38) = 9.14$, $MSE = .022$; type x timing: $F(1, 38) = .14$, $MSE = .022$, $p = .71$].

Final cued recall test: incorrect lures. As in Experiment 1, taking a multiple-choice test with no feedback led to higher proportion of incorrect-lure responses on the final test than were spontaneously produced in the no test condition [.26 vs. .18; $t(39) = 2.25$, $SEM = .033$; $d = .38$, $p_{rep} = .91$]. Providing feedback significantly reduced incorrect-lure responses relative to taking a test with no feedback [.13 vs. .26; $t(39) = 4.66$, $SEM = .027$; $d = .77$, $p_{rep} = 1.00$]. However, neither type nor timing of feedback had an effect on the proportion of incorrect-lure responses produced on the final test with each feedback condition producing roughly the same proportion: answer-until-correct (.15), standard (.12), immediate (.14), and delayed (.13) feedback. A 2 × 2 repeated-measures ANOVA confirmed this observation [type of feedback: $F(1, 38) = 2.49$, $MSE = .013$, $p = .12$; timing of feedback: $F(1, 38) = .33$, $MSE = .013$, $p = .57$; type x timing: $F(1, 38) = 1.53$, $MSE = .013$, $p = .22$].

Conditional analyses. Final cued recall performance was again analyzed with respect to response outcome (correct/incorrect) on the initial multiple-choice test. As in Experiment 1, the proportion of correct responses on the final test given an incorrect response on the initial test (i.e., wrong answers that were corrected) was approximately equal in the standard and answer-until-correct conditions (.43 vs. .41). The analysis of how the timing of feedback affected initially correct and incorrect responses revealed a similar pattern of results. The bottom panel of Figure 2 displays the proportion correct on the final cued-recall test as a function of initial test outcome (correct/incorrect) and feedback condition. Replicating the results of Experiment 1, providing feedback benefited both correct and incorrect responses. However, unlike Experiment 1, delayed feedback benefited both correct and incorrect responses relative to immediate feedback.

Discussion

Experiment 2 showed robust effects of testing and providing feedback, replicating the results of Experiment 1. A significant advantage of delayed feedback compared with immediate feedback was found, and the magnitude of the advantage was increased relative to Experiment 1, probably because of the longer interval of retention. However, we again found no difference in performance

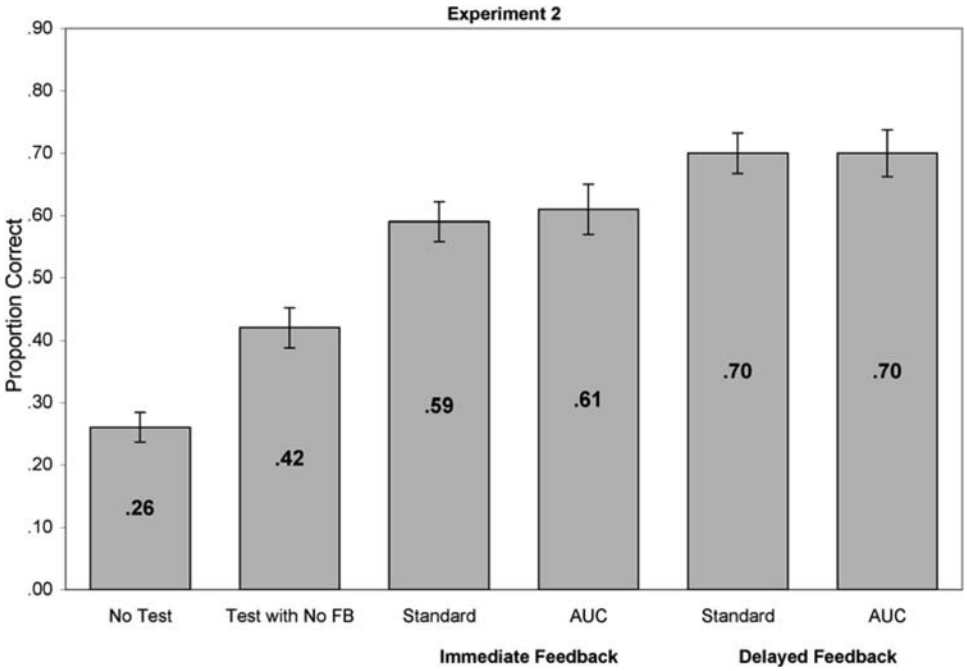

Figure 3. Proportion correct on the final cued recall in Experiment 2 as a function of timing and type of feedback as well as for the two control conditions (test with no feedback and no test). AUC denotes the answer-until-correct feedback condition. Error bars represent standard error of the mean. The abbreviation *FB* refers to feedback.

between the two types of feedback. Answer-until-correct feedback did not provide any increase in retention relative to standard feedback.

General Discussion

Our experiments investigated how the type and timing of feedback influence learning from a multiple-choice test. Delayed feedback produced better long-term retention than immediate feedback. However, there was no difference between the two types of feedback: answer-until-correct and standard feedback produced almost identical performance on the final cued-recall test. Feedback increased the proportion of correct responses and reduced the proportion of incorrect-lure responses relative to taking a test with no feedback. A testing effect was also observed: participants showed better performance for items initially tested than those not tested.

In both experiments, the type of feedback provided after the multiple-choice test did not affect final performance on the test. As discussed in the introduction, previous studies that found a superiority of answer-until-correct feedback have often naturally confounded type and timing of feedback (e.g., Angell, 1949; Dihoff et al., 2003; Dihoff et al., 2004; Sullivan et al., 1971). When these two variables were disentangled in our experiment, answer-until-correct feedback was found to be no more effective than standard feedback. One potential explanation for this similarity in effectiveness is the controlled manner in which feedback was presented in our experiment. In both the answer-until-correct and standard feedback conditions, participants received feedback for a set amount of time, regardless of whether they were correct or incorrect. As described earlier, classroom studies (e.g., Dihoff et al., 2003) generally provide standard feedback by returning a graded test to students and allowing them to process feedback however they please. In such circumstances, students may not fully process the feedback; for example, they may choose to concentrate on feedback for incorrect responses. In contrast to this standard method of providing feedback, the answer-until-correct feedback ensures that students will process feedback for all of their responses. In situations where feedback processing is not controlled, such a difference in the processing required by the feedback task may give rise to the superiority of answer-until-correct feedback. Another possible explanation is that the iterative process of responding in answer-until-correct feedback may contain a potential detriment to learning: the exposure to misinformation in the form of multiple-choice lures (Butler & Roediger, 2007b; Roediger & Marsh, 2005). The more responses needed to select the correct response, the more lures are selected and scrutinized. This process may lead to poorer retention of the correct response, but more research is needed to investigate this possibility. Nevertheless, the results of both experiments clearly show that, overall, the answer-until-correct procedure does not provide any additional benefit relative to standard feedback.

Delaying feedback led to better performance on the final cued-recall test relative to immediate feedback, and the magnitude of the effect grew as the retention interval increased (from .05 after one day to .10 after one week). This result suggests that, like many other memory effects, the benefit of delayed feedback may require longer periods of time to emerge. Theoretically, the superiority of the delayed feedback result can be best explained by invoking two

different, but compatible ideas (see also Butler & Roediger, 2007b). The conditional analyses indicated that delayed feedback benefited both initially correct and incorrect responses. Presumably, the benefit of delayed feedback after a correct response is a type of spacing effect (Dempster, 1989; Schmidt & Bjork, 1992): after a correct response, delayed feedback represents a spaced presentation of the material, whereas immediate feedback constitutes a massed presentation. However, a different explanation is required for initially incorrect responses. The benefit of providing delayed feedback after incorrect responses may be the result of a decrease in response competition between the incorrect and correct responses (see also Kulhavy, 1977). Delaying feedback may allow the accessibility of the incorrect response to dissipate, which facilitates learning of the correct response.

Although our study clearly shows the superiority of delayed feedback, many previous studies have found a benefit of immediate feedback (e.g., Bourne, 1957; Paige, 1966; Brosvic, Epstein, Cook, & Dihoff, 2005). The most recent meta-analysis of research in feedback timing (Kulik & Kulik, 1988) concluded that delayed feedback is generally found to be superior in laboratory studies, whereas immediate feedback is often shown to be more effective in applied studies in actual classroom settings. However, the laboratory versus applied distinction is an unsatisfactory explanation for the variety of findings on this topic. A more viable explanation for the superiority of immediate feedback in some studies is that students sometimes may not fully process feedback after a delay unless required to do so (as in laboratory studies). This hypothesis may also help to explain the laboratory/applied distinction. Laboratory studies generally exercise a greater degree of control of feedback processing after a delay. For example, in our study, participants were required to look at feedback for each responses for a set amount of time, regardless of whether or not the response was correct. In applied studies (e.g., Dihoff et al., 2003), delayed feedback usually consists of a self-paced review of a graded test, and students may attend primarily to incorrect responses. These students may process feedback for the correct response quickly or may skip correct items entirely. In the same studies, immediate feedback is generally given after each response, which is more likely to lead students to fully process feedback after both correct and incorrect responses. Given the benefit of providing feedback after correct responses in this study and others (Butler & Roediger, 2007b; Butler, Karpicke, & Roediger, 2007), the processing of feedback after correct responses is quite important, and conditions should be arranged so as to maximize it.

The results of this study have several implications for educational practice. First, the "standard" method of returning a graded examination seems to be just as beneficial to learning as the use of methods of testing that provide students with answer-until-correct feedback. However, it is critical that students fully process the feedback (i.e., review both correct and incorrect responses). In the classroom, testing with answer-until-correct feedback may be a very practical way to accomplish this goal because it ensures full processing of the feedback. Nevertheless, we also found that delaying feedback benefits retention relative to immediate feedback. Thus, it might be more beneficial to provide feedback after a delay with some incentive to actively process it. For example, the instructor could give students back a photocopy of their ungraded test (keeping the originals to grade) at the next class meeting with the instructions to grade and correct the test for partial credit. Such a procedure would ensure that the feedback is delivered after a delay and that students are motivated to fully process the feedback (also, importantly, it would not take away from class time). As long as students process the feedback for both correct and incorrect answers, educators should not worry about taking a few days to return a graded examination. The optimal delay of feedback is likely to depend on many situational factors (e.g., subsequent retention interval, etc.), and more research is needed to investigate whether the efficacy of delayed feedback diminishes at longer delays (e.g., 1 month or more). It is possible that delaying feedback for too long will result in students being unmotivated to look at anything more than their grade. In practice, individual educators must make a decision about how best to give feedback in their classroom. Nevertheless, our experiments clearly show the benefit of providing feedback after a test. Delayed feedback is preferred, if circumstances require students to attend carefully to the feedback.

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology, 6,* 451–474.

Angell, G. W. (1949). The effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. *Journal of Educational Research, 42,* 391–394.

Angell, G. W., & Troyer, M. E. (1948). A new self-scoring test device for improving instruction. *School and Society, 67,* 84–85.

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

Bourne, L. E., Jr. (1957). Effect of information feedback and task complexity on the identification of concepts. *Journal of Experimental Psychology, 54,* 201–207.

Brosvic, G. M., Epstein, M. L., Cook, M. J., & Dihoff, R. E. (2005). Efficacy of error for the correction of initially incorrect assumptions and of feedback for the affirmation of correct responding: Learning in the classroom. *The Psychological Record, 55,* 401–418.

Bruner, J. S. (1967). On knowing: Essays for the left hand. Cambridge, MA: Harvard University Press.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III (2007). *A matter of confidence: Correct responses benefit from feedback.* Manuscript under review.

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

Butler, A. C., & Roediger, H. L., III (2007b). *Low confident correct responses benefit from feedback.* Manuscript under review.

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

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

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review, 1,* 309–330.

DiBattista, D. (2005). The Immediate Feedback Assessment Technique: A learner-centered multiple-choice response form. *Canadian Journal of Higher Education, 35,* 111–131.

Dihoff, R. E., Brosvic, G. M., & Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record, 53,* 533–548.

Dihoff, R. E., Brosvic, G. M., Epstein, M. L., & Cook, M. J. (2004). Provision of feedback during preparation for academic testing: Learning

is enhanced by immediate but not delayed feedback. *The Psychological Record, 54,* 207–231.

Epstein, M. L., Epstein, B. B., & Brosvic, G. M. (2001). Immediate feedback during academic testing. *Psychological Reports, 88,* 889–894.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119,* 254–284.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47*(1), 211–232.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58,* 79–97.

Mulligan, N. W., & Lozito, J. P. (2004). Self-generation and memory. In B. Ross (Ed.), Psychology of learning and motivation (Vol. 45, pp. 175–214). San Diego, CA: Elsevier Academic Press.

Paige, D. D. (1966). Learning while testing. *The Journal of Educational Research, 59,* 276–277.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31(1),* 3–8.

Peterson, J. C. (1930). A new device for use in teaching, testing and research in learning. *Transactions of the Kansas Academy of Science, 33,* 41–47.

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

Roediger, H. L., III, & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17(3),* 249–255.

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

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 1155–1159.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime Reference Guide. Pittsburgh: Psychology Software Tools, Inc.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24,* 86–97.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 592–604.

Stephens, A. L. (1960). Certain special factors involved in the law of effect. In A. A. Lumsdaine & R. Glaser, Eds. *Teaching machines and programmed learning: A source book,* pp. 89–93. Washington, DC: National Education Association.

Sullivan, H. J., Schutz, R. E., & Baker, R. L. (1971). Effects of systematic variations in reinforcement contingencies on learner performance. *American Educational Research Journal, 8,* 135–141.

Received January 29, 2007
Revision received June 25, 2007
Accepted July 9, 2007 ■