Journal of Experimental Psychology:
Learning, Memory, and Cognition
2008, Vol. 34, No. 4, 918–928

0278-7393/08/$12.00 DOI: 10.1037/0278-7393.34.4.918

Correcting a Metacognitive Error: Feedback Increases Retention of Low-Confidence Correct Responses

Andrew C. Butler
Washington University in St. Louis

Jeffrey D. Karpicke
Purdue University

Henry L. Roediger, III
Washington University in St. Louis

Previous studies investigating posttest feedback have generally conceptualized feedback as a method for correcting erroneous responses, giving virtually no consideration to how feedback might promote learning of correct responses. Here, the authors show that when correct responses are made with low confidence, feedback serves to correct this initial metacognitive error, enhancing retention of low-confidence correct responses. In 2 experiments, subjects took an initial multiple-choice test on general knowledge facts and made a confidence judgment after each response. Feedback was provided for half of the questions, and retention was assessed by a final cued-recall test. Taking the initial test improved retention relative to not testing, and feedback further enhanced performance. Consistent with prior research, feedback improved retention by allowing subjects to correct initially erroneous responses. Of more importance, feedback also doubled the retention of correct low-confidence responses, relative to providing no feedback. The function of feedback is to correct both memory errors and metacognitive errors.

Keywords: feedback, testing, metacognition, confidence

Testing of information can have a powerful positive effect on future retention of the tested material, a phenomenon known as the *testing effect* (Butler & Roediger, 2007; Carpenter & DeLosh, 2006; Carpenter & Pashler, 2007; Chan, McDermott, & Roediger, 2006; Karpicke & Roediger, 2008; Roediger & Karpicke, 2006a, 2006b). Testing often improves later retention even when students are not given feedback following the test, and providing feedback produces even greater gains in long-term retention (Butler & Roediger, 2008; Karpicke & Roediger, 2007; McDaniel & Fisher, 1991). In this article, we examined whether feedback improves long-term retention only of responses that are incorrect on an initial test or whether feedback also improves retention of initially correct responses. A large body of research has investigated factors that determine the effectiveness of feedback, such as how different types and schedules of feedback affect learning (see Azevedo & Bernard, 1995; Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Kulhavy & Stock, 1989; Kulik & Kulik, 1988). The theoretical synthesis of this research has yielded many suggestions for how and when feedback should be given. Although practical recommendations for maximizing the efficacy of feedback vary considerably (and sometimes contradict each other), the majority of these suggestions are derived from research aimed at only one aspect of feedback, viz., that the purpose of feedback is to correct errors. In fact, the current zeitgeist has so sharply shifted toward conceptualizing feedback as an error-correction mechanism that the possible effect of feedback on correct responses is often minimized or completely neglected. For example, some recent investigations of feedback have exclusively dealt with how feedback influences the correction of errors, without examining how feedback affects learning of correct responses (e.g., Butterfield & Metcalfe, 2001, 2006; Meyer, 1986).

The purpose of the current research is to reexamine the effect of feedback on retention of initially correct responses. Of course, we are not arguing against the fact that correcting memory errors is a key purpose of feedback. Instead, we believe that feedback also functions as an error-correction mechanism for correct responses, albeit for a different type of error. When individuals make a correct response but are not confident in the response, there is a discrepancy between the subjective and objective correctness of their answers. In other words, low-confidence correct responses reflect an error of metacognitive monitoring, which in this context refers to the ability to assess the accuracy of one's own performance on a test (Barnes, Nelson, Dunlosky, Mazzoni, & Narens, 1999; Koriat & Goldsmith, 1996; Nelson & Narens, 1990). Feedback that confirms the correctness of low-confidence responses should enable learners to reduce the discrepancy between their perceived and actual performance by allowing them to adjust their subjective assessments of their knowledge. Further, if feedback allows learners to correct initial metacognitive errors, then it should also

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

This research was supported by a Collaborative Activity Award from the James S. McDonnell Foundation and Grant R305H030339 from the Institute of Education Sciences.

Correspondence concerning this article should be addressed to Andrew C. Butler, Department of Psychology, Campus Box 1125, Washington University in St. Louis, One Brookings Drive, St. Louis, MO 63139-4899. E-mail: butler@wustl.edu

enhance long-term retention of the correct responses and improve the accuracy of metacognitive monitoring on subsequent tests. Thus, our hypothesis in this research was that, just as feedback helps correct memory errors, feedback will also help correct metacognitive errors and will improve retention of low-confidence correct responses.

Before describing the current research, we briefly discuss the historical origins of the current zeitgeist focused on the role of feedback in correcting memory errors. We then outline the theoretical basis for our reexamination of the role of feedback after correct responding and review previous research that has investigated the effect of feedback on correct responses.

Feedback as a Mechanism for Correcting Memory Errors

The heavy emphasis on the correction of erroneous responses in feedback research is in large part a product of the effort to dismiss the notion that feedback acts as a reinforcer, an idea popular in earlier literature (e.g., Skinner, 1954). Kulhavy (1977) argued against any reinforcing quality of feedback by demonstrating that feedback did not benefit learning in verbal conditioning paradigms as reinforcement does in other situations. For example, reinforcement increases the future probability of a response and thus should have its greatest effect on correct responses. In addition, reinforcement is most effective when given immediately after the response. From an extensive review of the feedback literature, Kulhavy concluded that there was scant evidence to support the notion that the principles derived from behavioral research on reinforcement apply to the provision of feedback on learning of educational materials in humans. As an example, he cited evidence that delayed feedback has greater positive effects than immediate feedback in some situations (e.g., Brackbill & Kappy, 1962; Surber & Anderson, 1975; Sassenrath & Yonge, 1968). Indeed, Kulhavy argued that the whole idea of conceptualizing the complexities of learning in educational settings purely within an operant conditioning framework may be of limited utility. Nevertheless, in the process of eliminating the idea that feedback may be conceived as a reinforcer, researchers began to overlook how feedback may benefit correct responses, a trend that continues today.

Feedback as a Mechanism for Correcting Metacognitive Errors

The impetus for reexamining the function of feedback after correct responding stems in part from studies that assessed subjects' confidence in their responses followed by self-paced study of feedback. Kulhavy, Yekovich, and Dyer (1979) had subjects complete a program of instruction in which they learned about heart disease and answered multiple-choice questions on each section of the tutorial. After each multiple-choice question, subjects were given feedback and allowed to study that feedback for as long as they wanted. Among other dependent measures, Kulhavy et al. reported feedback study times as a function of initial response outcome (correct or incorrect) and response confidence. After a correct response, subjects studied feedback for a significantly longer period of time if that response was made with low confidence. In addition, feedback study times were roughly equivalent for low-confidence correct responses and low-confidence incorrect responses. Overall, this study and others (for a review, see Kulhavy, 1977; Kulhavy & Stock, 1989) showed that subjects spent a substantial amount of time processing feedback after a low-confidence correct response.

Why do subjects spend more time processing feedback after a low-confidence than a high-confidence correct response? One potential explanation is that feedback is processed differently when there is a large discrepancy between the subjective assessment and the objective correctness of a response. Consider a test in which subjects are required to respond to every question on the test (a forced-response test; cf. Koriat & Goldsmith, 1994; Roediger & Payne, 1985). For each test question, they retrieve information from memory and monitor the accuracy of that information (assessed as a confidence judgment), but they are also required to make a response to each question. On such a test, an individual's confidence in his or her responses may not correspond well to the correctness of the responses, leading to the production of low-confidence correct responses (Roediger, Wheeler, & Rajaram, 1993) and high-confidence incorrect responses (Butterfield & Metcalfe, 2001; 2006). When subjects become aware of a metacognitive error through feedback, they may attempt to resolve the discrepancy between the subjective assessment and the objective correctness of their response by devoting additional cognitive resources to processing the feedback.

Research that has examined the role of feedback in correcting errors committed with high confidence provides some evidence to support this idea (Butterfield & Metcalfe, 2001, 2006). Butterfield and Metcalfe (2001) had subjects answer general knowledge questions and rate their confidence in their initial answers. The subjects were given feedback following all of their responses on the first test. After a brief delay, the subjects took a second test over the general knowledge questions. Butterfield and Metcalfe (2001) examined the relationship between subjects' initial confidence in incorrect responses and the likelihood they would correct those responses on the final test. They found that subjects were especially likely to correct high-confidence incorrect responses, a result they called the "hypercorrection effect" (see also Kulhavy, Yekovich, & Dyer, 1976). In a follow-up study, Butterfield and Metcalfe (2006) replicated the hypercorrection effect and found that subjects tend to miss or ignore tones presented in a secondary tone detection task while processing feedback after high-confidence errors. They concluded that subjects pay more attention to feedback after high-confidence errors because of the surprise of being highly confident but incorrect, and this additional attention to the correct response produces better retention (Butterfield & Metcalfe, 2006).

Our hypothesis is that feedback serves to correct the metacognitive error inherent in low-confidence correct responses, much like it does for high-confidence errors in the hypercorrection effect. However, we believe that the correction of these two types of metacognitive error probably leads to better retention through different mechanisms. As described above, retention may be enhanced following high-confidence errors because a feeling of surprise causes subjects to pay more attention to feedback (Butterfield & Metcalfe, 2006). In contrast, we think that providing feedback after low-confidence correct responses might enhance retention by enabling learners to strengthen the association between the cue and response and to inhibit any competing responses. We turn now to summarizing previous research on feedback after correct responses; we return to the idea of why

providing feedback after low-confidence correct responses might produce better retention in the General Discussion.

Previous Research on the Effect of Feedback on Correct Responses

Previous researchers have typically treated the effect of feedback on correct responses as a secondary concern rather than as the focus of their research efforts. These researchers have also widely concluded that when a correct response is produced, feedback makes little or no difference for learning (Anderson, Kulhavy, & Andre, 1971; Guthrie, 1971; Kulhavy & Anderson, 1972; Pashler, Cepeda, Wixted, & Rohrer, 2005; Pashler, Rohrer, Cepeda, & Carpenter, 2007). For example, Pashler et al. (2005) had subjects study a list of 20 Luganda–English word pairs twice and then take two consecutive cued-recall tests during an initial learning session. Of importance, subjects were free to withhold responses on the initial tests (a point we elaborate below). Subjects made confidence ratings after each response and were given feedback for some of the items and no feedback for other items. Finally, long-term retention was measured on a final cued-recall test 1 week later. When final test performance was examined as a function of whether responses on the first test were correct or incorrect, Pashler et al. (2005) found that feedback did not enhance the retention of correct responses, even those made with medium or low confidence, using their procedure. They concluded, "When the learner makes a correct response, feedback makes little difference for what can be remembered 1 week later" (Pashler et al., 2005, p. 7; see, too, Pashler et al., 2007).

However, a careful consideration of the methodology used in the experiment by Pashler et al. (2005), and in other studies, raises the question of whether the experiments were capable of properly assessing the potential benefit of providing feedback after correct responses. We argue that previous researchers may have failed to find benefits of feedback for correct responses because few (if any) low-confidence correct responses were made on the initial free-report tests used in prior experiments. On a free-report test, when subjects can choose to volunteer or withhold responses, there is a strong correspondence between subjective confidence (metacognitive monitoring) and the willingness to volunteer a response (metacognitive control). That is, subjects generally volunteer high-confidence responses and withhold low-confidence responses, even if the low-confidence responses are correct (see Barnes et al., 1999; Kelley & Sahakyan, 2003; Koriat & Goldsmith, 1996). In the previous research discussed above, subjects were free to withhold responses to test items, and, therefore, low-confidence responses were likely withheld even if they were in fact correct. Presumably, the previous studies used free-report tests because none of them was designed with the specific intention of examining the effects of feedback on low-confidence correct responses. Although multiple-choice and cued-recall tests are generally forced and free report, respectively, it is important to note that either report option can be used with either test format. Thus, report option is the critical variable, not test format. In the present research, we used a multiple-choice test with a forced-responding procedure in which subjects were required to respond to each test question and to indicate their confidence in their responses. This procedure ensured that subjects would produce low-confidence responses on the initial test.

Experiment 1

In Experiment 1, subjects took a multiple-choice general knowledge test on which they received feedback for half of the questions (test with feedback condition) and no feedback on the other half of the questions (test with no feedback condition). Subjects were required to make a response to each question (a forced-report test), and they then made a confidence judgment after each response (always before receiving any feedback). After a brief distracter task, subjects took a final cued-recall test, in which they answered the general knowledge questions but without the aid of response alternatives. The final test included previously tested items and new items as a no-test control condition. Of specific interest was the effect of feedback on retention of low-confidence correct responses.

Method

Subjects. Thirty undergraduate psychology students at Washington University in St. Louis participated for course credit. All subjects were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" put forth by the American Psychological Association (2002).

Materials and counterbalancing. Stimuli consisted of general knowledge questions that were created from 60 facts taken out of the *World Book Encyclopedia* (World Book, Inc., 2002). Test items were constructed from the facts by forming a question stem and target response (e.g., *What is the longest river in the world?* Answer: *the Nile river*). For the purposes of the multiple-choice test, three plausible lure responses were generated for each test item and paired with the target to form a four-alternative multiple-choice test.

The experiment was counterbalanced in two ways. First, the questions were separated into three groups and each group of questions appeared in each of the three learning conditions (no test, test with no feedback, test with feedback) equally across subjects. To accomplish this counterbalance, the groups of questions were rotated through the learning conditions, creating three versions of the experiment. Second, the position of the correct answer relative to the lures was systematically varied such that the target appeared equally often in each of the four possible positions across all the questions on each version of the multiple-choice test.

Procedure. Subjects were tested in groups of 1 to 5 people. The stimuli were presented, and responses were collected individually on a PC using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002). First, a self-paced multiple-choice test was given in which 40 questions were presented sequentially in a random order determined by the computer. Each question was displayed on the top of the screen with four alternative answers below it. Subjects were required to respond to each question (i.e., forced report) by pushing the button of the number that corresponded to the correct answer (*1*, *2*, *3*, or *4*). After responding to the question, subjects were prompted to rate their confidence in the response on the following 4-point scale: 1 = *guess*, 2 = *low confidence*, 3 = *medium confidence*, and 4 = *high confidence*. Immediately after choosing an answer and rating their confidence, subjects either received feedback on their answer or were presented with a screen that instructed them to wait for the next question. Feedback consisted of a re-presentation of the multiple-

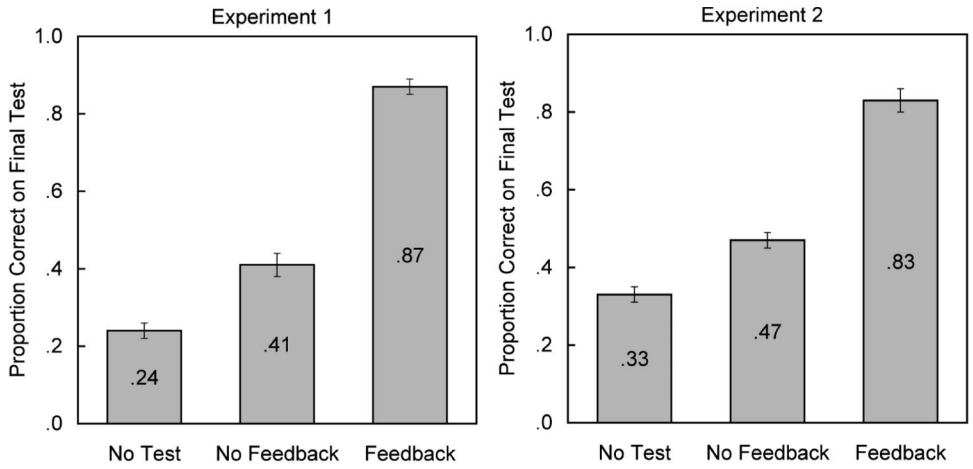

Figure 1. Proportion of correct responses on the final cued-recall test as a function of initial learning condition for Experiment 1 (left panel) and Experiment 2 (right panel). Error bars indicate the standard error of the mean.

choice question stem along with the correct response. The feedback and wait instructions were both displayed for 4 s, so that total time spent on each question was equated in the test with feedback and test with no feedback conditions. After the multiple-choice test, subjects played a computer game for 5 min as a filler task. Finally, a cued-recall test was given in which the complete set of 60 questions (40 from the initial multiple-choice test plus 20 untested questions) were tested. Again, the questions were presented in a random order determined by the computer and answering was self-paced. Subjects were told to type in the correct answer to each question but were warned to respond only if they were reasonably sure that the answer was correct. Thus, the final test was free report, not forced report. If they did not know the correct answer, then they were instructed to push the *Enter* key to skip that question. After the cued-recall test was complete, subjects were debriefed and dismissed.

Results

All results, unless otherwise stated, were significant at the .05 level. Pairwise comparisons were Bonferroni-corrected to the .05 level. In the analysis of repeated measures, the Geisser–Greenhouse epsilon correction was used for violations of the sphericity assumption (Geisser & Greenhouse, 1958).

Initial multiple-choice test. Performance on the initial multiple-choice test was equivalent in the test with no feedback and test with feedback conditions (.52 vs. .55; $t < 1$), which was expected because no manipulation had been introduced yet.

Final cued-recall test. There were large effects of testing and feedback on recall on the final test (see the left panel of Figure 1). The test with feedback condition produced a greater proportion of correct responses on the final test than the test with no feedback condition (.87 vs. .41), $t(29) = 19.1$, $SEM = .024$, $d = 1.77$, $p_{rep} = 1.00$ (p_{rep} is an estimate of the probability of replicating the direction of an effect; see Killeen, 2005), which, in turn, led to a greater proportion of correct responses than the no-test condition (.41 vs. .24), $t(29) = 7.5$, $SEM = .023$, $d = 1.06$, $p_{rep} = 1.00$. A one-way repeated-measures analysis of variance (ANOVA) revealed a significant difference among the three learning conditions, $F(2, 58) = 355.5$, $MSE = .009$, $\eta_p^2 = .93$. Thus, we observed a strong testing effect even without feedback, but it was greatly enhanced when feedback was given.

Conditional analyses. Conditional analyses were performed to examine performance on the final cued-recall test as a function of (a) the response outcome on the initial multiple-choice test (correct or incorrect), (b) the presence or absence of feedback, and (c) the level of confidence in the initial response. For each subject, performance on the final cued-recall test in the test with no feedback and test with feedback conditions was broken down as a function of response outcome on the initial multiple-choice test. The left panel of Figure 2 shows the proportion of correct responses on the final cued-recall test as a function of initial response outcome. As expected, initially incorrect responses benefited substantially from feedback. When feedback was provided, most of the initially incorrect responses were corrected, whereas few initially incorrect responses were (spontaneously) corrected on the final test without feedback (.82 vs. .03), $t(29) = 34.4$, $SEM = .023$, $d = 6.08$, $p_{rep} = 1.00$. It is important to note that feedback also benefited correct responses: A greater proportion of initially correct responses were reproduced on the final test when they had been followed by feedback than when they had been followed by no feedback (.93 vs. .79), $t(29) = 4.4$, $SEM = .032$, $d = 0.76$, $p_{rep} = .99$.

The conditionalized results were further partitioned as a function of the confidence rating ("guess," "low confidence," "medium confidence," "high confidence") given after each question on the initial multiple-choice test. On the basis of the literature reviewed above, one concern was whether the forced-report procedure we used would produce a good distribution of confidence responses and, in particular, a sufficient number of low-confidence correct responses. Table 1 shows the proportion of items (averaged across subjects) that were assigned to each confidence rating as a function of response outcome (correct or incorrect) and initial learning condition (test with no feedback or test with feedback). Overall, the items were well distributed across the four confidence levels, with 38% and 40% of correct responses assigned ratings of "guess" or "low confidence" in the no feedback and feedback conditions, respectively.

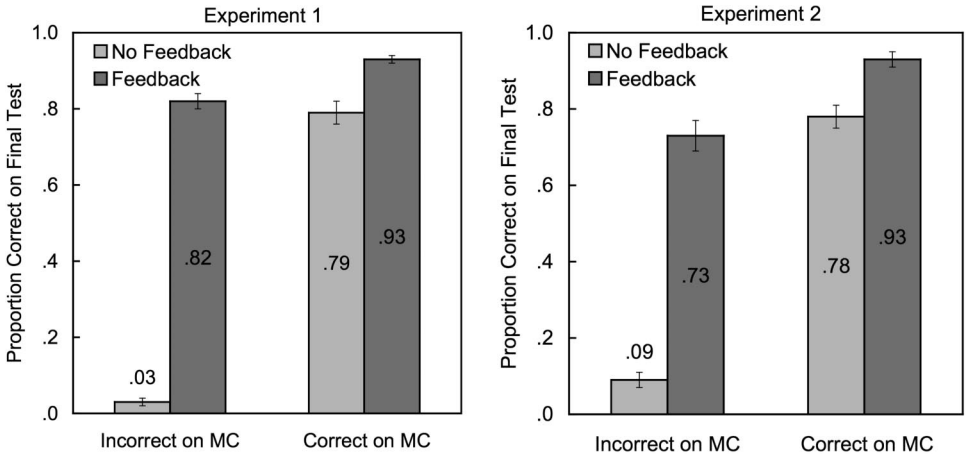

Figure 2. Proportion of correct responses on the final cued-recall test as a function of response outcome (correct or incorrect) on the initial multiple-choice (MC) test and learning condition for Experiment 1 (left panel) and Experiment 2 (right panel). Error bars indicate the standard error of the mean.

The key results of Experiment 1 are shown in Figure 3, which depicts the proportion correct on the final cued-recall test for initially correct responses as a function of initial response confidence, learning condition, and initial response outcome. Focusing first on the initially incorrect responses (the left panel of Figure 3), the proportion of correct responses on the final test generally did not differ as a function of response confidence, regardless of whether feedback was provided. There was, however, one exception: When feedback was given, a significantly greater proportion of high-confidence incorrect responses were corrected relative to all other confidence levels (.93 vs. .80), $t(29) = 3.4$, $SEM = .035$, $d = 0.59$, $p_{rep} = .99$, replicating the hypercorrection effect that has been found in previous research (Butterfield & Metcalfe, 2001, 2006; see also Kulhavy et al., 1976).

A different pattern of results emerged for initially correct responses (the right side of Figure 3). As described above, a greater proportion of initially correct responses were maintained to the final cued-recall test when feedback was provided relative to no feedback. Figure 3 also shows that a greater proportion of correct responses were maintained as the initial response confidence increased in both the test with no feedback and test with feedback conditions. In addition, these two factors interacted such that the test with feedback condition produced a greater proportion of correct responses than the test with no feedback condition at every confidence level except high confidence (which approached ceiling). The difference between the two learning conditions increased as response confidence decreased. For initial responses labeled as a "guess" (the lowest level of confidence), .85 were maintained with feedback whereas only .40 were maintained without feedback. A 4 × 2 repeated-measures ANOVA confirmed these observations: There were significant main effects of initial learning condition, $F(1, 29) = 42.4$, $MSE = .054$, $\eta_p^2 = .59$, and response confidence, $F(3, 87) = 27.4$, $MSE = .059$, $\eta_p^2 = .49$, as well as

Table 1
Proportion of Items (With Total Number of Items in Parentheses) Assigned Each Confidence Rating on the Initial Multiple-Choice Test as a Function Response Outcome and Initial Learning Condition

	Correct on multiple choice		Incorrect on multiple choice	
Experiment and confidence	Test with no feedback	Test with feedback	Test with no feedback	Test with feedback
Experiment 1				
Guess	.18 (52)	.20 (66)	.44 (133)	.41 (116)
Low	.20 (61)	.20 (62)	.26 (79)	.34 (97)
Medium	.24 (73)	.22 (69)	.20 (60)	.19 (55)
High	.38 (115)	.38 (120)	.10 (27)	.06 (15)
Total	1.00 (301)	1.00 (317)	1.00 (299)	1.00 (283)
Experiment 2				
Guess	.10 (33)	.11 (33)	.32 (91)	.29 (87)
Low	.25 (80)	.22 (66)	.39 (112)	.49 (145)
Medium	.13 (39)	.18 (55)	.18 (51)	.14 (41)
High	.52 (164)	.49 (149)	.11 (30)	.08 (24)
Total	1.00 (316)	1.00 (303)	1.00 (284)	1.00 (297)

Note. Confidence responses are binned for Experiment 2 (.25 = *guess*, .26–.50 = *low confidence*, .51–.75 = *medium confidence*, and .76–1.00 = *high confidence*).

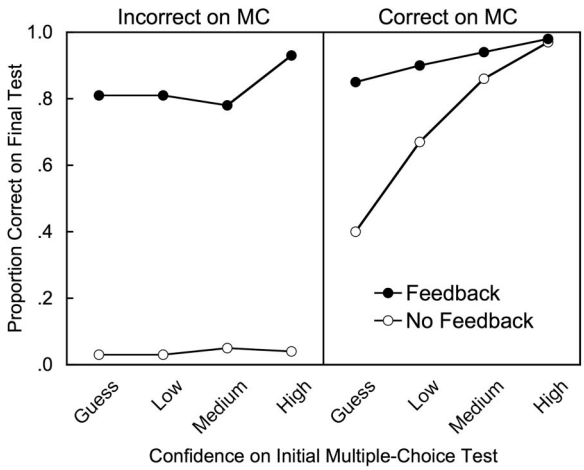

Figure 3. Proportion of correct responses on the final cued-recall test in Experiment 1 as a function of response confidence and learning condition for responses that were incorrect (left side) and correct (right side) on the initial multiple-choice (MC) test.

a significant interaction between learning condition and response confidence, $F(3, 87) = 17.9$, $MSE = .042$, $\eta_p^2 = .38$.

We carried out one final analysis to examine the effect of feedback on the relationship between confidence and memory performance. Specifically, we asked how confidence on the initial multiple-choice test is related to the production of answers on the final cued-recall test and how providing feedback affects this relationship. To address these questions, we computed the within-subject Goodman–Kruskal gamma correlations between (a) multiple-choice performance and initial confidence and (b) initial confidence and final cued recall. Not surprisingly, the gamma correlations between initial multiple-choice performance and initial confidence were nearly identical in the test with feedback and test with no feedback conditions (.58 vs. .55; $t < 1$, $SEM = .063$, $p = .63$). In contrast, providing feedback significantly reduced the gamma correlation between initial confidence and final cued recall, relative to the test with no feedback condition (.70 vs. .40), $t(25) = 2.9$, $SEM = .101$, $d = 0.75$, $p_{rep} = .97$.[1] This result indicates that when subjects did not receive feedback after the initial multiple-choice test, final test performance corresponded well with initial confidence. However, providing feedback allowed subjects to correct erroneous responses and maintain correct responses, thereby reducing the relationship between initial response confidence and final cued recall.

Discussion

The results of Experiment 1 show that taking an initial multiple-choice test led to better performance on the final cued-recall test relative to not taking the test and that providing feedback after the initial test substantially increased the benefit of prior testing. Of more importance for present purposes, the conditional analyses revealed that both initially incorrect and correct responses benefited from feedback. After making an incorrect response, subjects used feedback to learn the correct response. When feedback was not provided after an incorrect response, the response almost always remained incorrect on the final test. This pattern of results did not differ as a function of the level of initial response confidence, except for the high-confidence incorrect responses, which were hypercorrected when feedback was provided.

The novel result of Experiment 1 was that feedback increased retention of initially correct responses. When feedback was not provided, initially correct responses were more likely to be changed or omitted on the final test. The level of response confidence on the initial multiple-choice test modulated this pattern of results. Almost all high-confidence correct responses were maintained to the final test, regardless of whether feedback was provided. However, as the level of initial response confidence decreased, feedback became increasingly important for maintaining correct responses on the final test. Providing feedback doubled retention of initially correct "guess" responses, relative to when feedback was not provided (.85 vs. .40).

Experiment 2

In contrast with previous studies, the results of Experiment 1 showed that feedback is critical for retention of low-confidence correct responses. Experiment 2 was conducted to replicate the results of Experiment 1 and to further investigate whether feedback benefits low-confidence correct responses. As described in the introduction, our hypothesis is that feedback enables subjects to correct a metacognitive error that occurs when the subjective correctness of a response (assessed by a confidence rating) does not correspond with its objective correctness. On the basis of our hypothesis, we predicted that feedback should not only increase retention of low-confidence correct responses but also improve the accuracy of confidence judgments made during a final delayed test. Thus, in Experiment 2, subjects were required to respond to each question on the final cued-recall test and to make confidence judgments for each response. Our prediction was that feedback would enhance the accuracy of metacognitive monitoring during the final test.

The procedure was the same as in Experiment 1 except for the following changes. First, on the final cued-recall test, subjects were required to respond to each test item and to make a confidence judgment for each item. Second, subjects made their confidence judgments on scales using cardinal values. On the initial four-alternative multiple-choice test, subjects rated their confidence on a scale from 25–100%, and on the final cued-recall test, subjects rated their confidence on a scale from 0–100%. This was done so that we could assess calibration (the absolute correspondence between test performance and confidence) in addition to resolution (the relative correspondence between performance and confidence; for elaboration, see Koriat & Goldsmith, 1996; Nelson, 1984; Nelson & Dunlosky, 1991). Finally, the retention interval before the final test was lengthened to 2 days so that we could generalize the results we observed on a relatively immediate test in Experiment 1 to a delayed final test in Experiment 2.

Method

Subjects. Thirty undergraduate psychology students at Washington University in St. Louis participated for course credit.

[1] Four subjects were excluded from this analysis because a gamma correlation could not be calculated for one of the two feedback conditions.

Materials and counterbalancing. The materials and counterbalancing scheme from Experiment 1 were used.

Procedure. The procedure was identical to that used in Experiment 1 with the following exceptions. First, the final cued recall was changed to a forced-report test, in which subjects produced a response and made a confidence rating for every question. Second, subjects made their confidence judgments on the initial multiple-choice test on a scale from 25–100%, where 25% represented no confidence and 100% represented complete confidence. Subjects were told that 25% represented guessing because chance probability of a correct response in a four-alternative multiple-choice test is 25%. On the final cued-recall test, subjects made their confidence judgments on a scale from 0–100%, where 0% represented guessing. Finally, subjects were dismissed after completing the initial multiple-choice test and returned 2 days later for the final recall test.

Results

Initial multiple-choice test. Performance on the initial multiple-choice test was virtually identical in the test with no feedback and test with feedback conditions (.53 vs. .51; $t < 1$).

Final cued-recall test. As in Experiment 1, testing improved retention relative to not taking an initial test, and testing with feedback produced better retention than testing without feedback (see the right panel of Figure 1). A one-way repeated-measures ANOVA showed a significant difference among the learning conditions, $F(2, 58) = 207.1$, $MSE = .009$, $\eta_p^2 = .88$. As in Experiment 1, the test with feedback condition produced a greater proportion of correct responses relative to the test with no feedback condition (.83 vs. .47), $t(29) = 15.9$, $SEM = .023$, $d = 2.89$, $p_{rep} = 1.00$, which in turn produced a greater proportion of correct responses than the no-test condition (.47 vs. .33), $t(29) = 5.5$, $SEM = .026$, $d = 1.00$, $p_{rep} = .99$.

Conditional analyses. Performance on the final cued-recall test in the test with no feedback and test with feedback conditions was broken down as a function of response outcome on the initial multiple-choice test for each subject. The right panel of Figure 2 shows the proportion of correct responses on the final cued-recall test as a function of initial response outcome. The pattern of results was the same as in Experiment 1. A substantially greater proportion of initially incorrect responses were corrected with feedback relative to without feedback (.73 vs. .09), $t(29) = 14.9$ $SEM = .043$, $d = 2.70$, $p_{rep} = 1.00$, and a greater proportion of initially correct responses were maintained with feedback than without feedback (.93 vs. .78), $t(29) = 5.3$, $SEM = .028$, $d = 0.99$, $p_{rep} = .995$.

The conditionalized data were further analyzed as a function of initial response confidence. For the purpose of data presentation and some analyses, the confidence responses have been separated into four bins that roughly corresponded to the categories used in Experiment 1 (.25 = *guess*, .26–.50 = *low confidence*, .51–.75 = *medium confidence*, and .76–1.00 = *high confidence*). Table 1 shows the proportion of items (averaged across subjects) that were assigned to each confidence rating as a function of response outcome (correct or incorrect) and initial learning condition (test with no feedback or test with feedback). As in Experiment 1, the items were well distributed across the four feedback levels and at least a third of the correct responses were assigned a rating of "guess" (.25) and "low confidence" (.26–.50).

Figure 4 shows the proportion of correct responses on the final cued-recall test as a function of initial response confidence, learning condition, and initial response outcome. Despite a number of differences in the procedures of Experiments 1 and 2, including the forced-response procedure on the final test and the 2-day delay used in Experiment 2, the overall pattern of results was similar to that of Experiment 1. When a response was answered incorrectly on the initial multiple-choice test and feedback was not given, it was unlikely for that response to be (spontaneously) corrected on the final cued-recall test. The one exception was the incorrect responses that were made with a confidence of .25 ("guess"). These responses were more likely to be corrected on the final test relative to responses in the other three confidence bins (.22 vs. .04), $t(29) = 3.3$, $SEM = .055$, $d = 1.00$, $p_{rep} = .98$. When forced to respond on the final test, subjects may have decided to switch their response to another alternative (whereas they might have omitted the response rather than switch in Experiment 1). In contrast to the hypercorrection effect observed in Experiment 1, initial response confidence did not influence the correction of errors when feedback was provided in Experiment 2, $F(3, 87) = 1.2$, $MSE = .076$, $p = .31$. Even when the analysis was restricted to initially incorrect items that were given 100% confidence judgment, no hypercorrection effect emerged. In fact, as Figure 4 shows, if anything, a hypocorrection effect seems to appear after 2 days because a somewhat lower proportion of high-confidence incorrect responses were corrected relative to all other confidence levels; however, this difference was not significant (.62 vs. .73), $t(29) = 1.7$, $SEM = .065$, $p = .10$. We address this failure to find the hypercorrection effect and the possibility of a hypocorrection effect below.

When a response was answered correctly on the initial multiple-choice test and feedback was not given, the pattern of performance

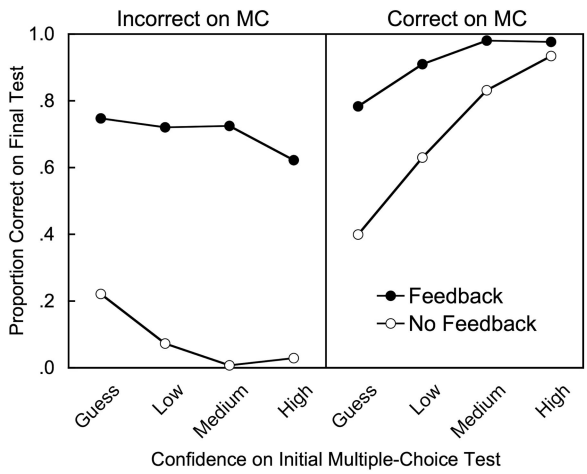

Figure 4. Proportion of correct responses on the final cued-recall test in Experiment 2 as a function of response confidence and learning condition for responses that were incorrect (left side) and correct (right side) on the initial multiple-choice (MC) test. Confidence responses have been separated into four bins that roughly corresponded to the categories used in Experiment 1 (.25 = *guess*, .26–.50 = *low confidence*, .51–.75 = *medium confidence*, and .76–1.00 = *high confidence*).

depended on the level of initial response confidence. As the level of confidence increased, a greater proportion of initially correct responses were maintained on the final test. In contrast, when feedback was provided after a correct response, that response tended to be maintained regardless of initial response confidence. Thus, the conditionalized data for initially correct responses exhibited the same sort of interaction as in Experiment 1. To confirm these observations, we conducted a 4 × 2 repeated-measures ANOVA. The ANOVA confirmed significant main effects of initial learning condition, $F(1, 29) = 57.4$, $MSE = .044$, $\eta_p^2 = .66$, and response confidence, $F(3, 87) = 29.1$, $MSE = .051$, $\eta_p^2 = .50$, as well as a significant interaction between learning condition and response confidence, $F(3, 87) = 6.3$, $MSE = .064$, $\eta_p^2 = .18$.

The effect of feedback on the relation between confidence and memory performance. The final set of analyses examined the effects of feedback on (a) the accuracy of initial confidence judgments relative to initial multiple-choice performance, (b) the relation between initial confidence and final test performance, and (c) the accuracy of confidence judgments made during the final cued-recall test.

As in Experiment 1, we first examined the relationship between initial confidence judgment and initial multiple-choice test performance (see Figure 5, Panel A). On the initial multiple-choice test, as expected, the mean gamma correlation (or resolution) was roughly equivalent in the test with no feedback and test with feedback conditions (.55 vs. .59), $t(29) = 0.6$, $SEM = .063$, $p = .54$. Subjects exhibited overconfidence in both the test with no feedback ($M_{Confidence} = .61$, $M_{Correct} = .53$) and the test with feedback ($M_{Confidence} = .59$, $M_{Correct} = .51$) conditions, but there was no difference in overconfidence between these two conditions. A 2 × 2 ANOVA confirmed that absolute confidence judgments were significantly higher than multiple-choice accuracy (.60 vs. .52), $F(1, 29) = 28.6$, $MSE = .007$, $\eta_p^2 = .50$, but there was no interaction ($F < 1$, $MSE = .005$, $p = .82$).

Gamma correlations were again computed to investigate the relationship between initial confidence judgments and final cued-recall test performance (see Figure 5, Panel B). As in Experiment 1, the no-feedback condition produced a significantly higher gamma correlation than did the feedback condition (.57 vs. .38), $t(25) = 1.86$, $SEM = .104$, $d = 0.47$, $p_{rep} = .89$.[2]

However, on the final cued-recall test (see Figure 5, Panel C), a different pattern emerged: Resolution was significantly better in the test with feedback condition relative to both the test with no feedback (.94 vs. .67), $t(28) = 5.1$, $SEM = .054$, $d = 0.92$, $p_{rep} = .99$, and no-test (.94 vs. .69), $t(28) = 6.3$, $SEM = .041$, $d = 0.98$, $p_{rep} = 1.00$, conditions.[3] On the final cued-recall test, there was a difference among the conditions: Subjects in the test with no feedback ($M_{Confidence} = .56$, $M_{Correct} = .47$) and the no-test ($M_{Confidence} = .38$, $M_{Correct} = .33$) conditions both showed overconfidence, whereas those in the test with feedback condition ($M_{Confidence} = .83$, $M_{Correct} = .82$) were almost perfectly calibrated. A 3 × 2 ANOVA revealed a significant interaction, $F(1, 58) = 10.1$, $MSE = .003$, $\eta_p^2 = .26$.

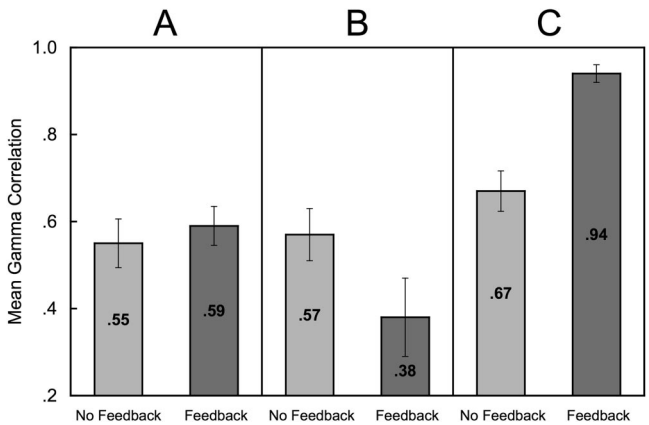

Figure 5. Mean gamma correlations between initial confidence judgments and initial multiple-choice performance (A), initial confidence judgments and final cued-recall performance (B), and final confidence judgments and final cued-recall performance (C). All results are from Experiment 2. Error bars indicate the standard error of the mean.

Discussion

Overall, Experiment 2 replicated and extended the main results of Experiment 1 after a 2-day retention interval, providing generality to the feedback effect along this dimension. Taking a prior test led to better performance on the final test relative to the no-test control, and feedback increased the benefit of prior testing. Again, the benefit of feedback stemmed from both the correction of erroneous responses and the confirmation of low-confidence correct responses. Experiment 2 also showed that when feedback was provided on the initial test, subjects were better able to discriminate between correct and incorrect responses on the final test. The improvement in the accuracy of metacognitive judgments in the feedback condition supports the idea that feedback helps to eliminate the discrepancy between perceived and actual correctness for low-confidence correct responses.

General Discussion

In summary, the results of both experiments show that taking an initial multiple-choice test produced superior performance on a subsequent cued-recall test and that this benefit of prior testing was further enhanced when feedback was provided. When an incorrect response was given on the initial multiple-choice test, it was unlikely to be corrected spontaneously on the final cued-recall test without the presentation of feedback. Thus, consistent with considerable prior research, we found that feedback helps learners correct memory errors. Of more importance, the results of the present experiments demonstrate that correct responses benefited from feedback, and this positive effect of feedback was greatest for low-confidence correct responses. Thus, feedback also helps learners correct the metacognitive error that occurs when they are correct on an initial test but lack confidence in their response, resulting in enhanced retention of low-confidence correct responses. Experiment 2 also showed that feedback produced a metacognitive benefit on the final test by improving resolution and calibration of confidence judgments made on the final test.

[2] Four subjects were excluded from this analysis because a gamma correlation could not be calculated for one of the two feedback conditions.

[3] One subject was excluded from this analysis because a gamma correlation could not be calculated for one of the two feedback conditions.

Overall, the results obtained in these two experiments confirm several points uncovered in previous research. Many studies have found that taking a prior test improves performance on a future test (e.g., Carpenter & DeLosh, 2006; Carpenter & Pashler, 2007; Roediger & Karpicke, 2006b; Wheeler & Roediger, 1992; for a review, see Roediger & Karpicke, 2006a) and that providing feedback after an initial test enhances subsequent test performance (e.g., Butler, Karpicke, & Roediger, 2007; Butler & Roediger, 2008; Karpicke & Roediger, 2007; McDaniel & Fisher, 1991; Pashler et al., 2005). In addition, several studies have found that providing feedback after incorrect responses is critical to correcting errors (e.g., Lhyle & Kulhavy, 1987; see Bangert-Drowns et al., 1991), especially those errors committed with high confidence (Butterfield & Metcalfe, 2001, 2006). This previous research has largely emphasized the role of feedback in correcting erroneous responses.

The present research provides two novel findings. First, providing feedback for low-confidence correct responses on an initial multiple-choice improved recall on a final test given either a few minutes or 2 days later. As explained in the introduction, this outcome differs from the findings reported in previous studies (e.g., Pashler et al., 2005), which have led to the conclusion that providing feedback after correct responses has no effect. There are many methodological differences between previous studies and the present research, but we argue that the key factor is whether subjects are required to respond to every item on an initial test (forced responding). When the initial test is free report, as was the case in many prior studies, subjects are likely to withhold low-confidence responses, even if they are correct (cf. Barnes et al., 1999; Koriat & Goldsmith, 1996). Therefore, when subjects are free to withhold low-confidence responses, no effect of feedback should be observed. In addition to forced responding, our procedure included many other features that increased the number of low-confidence correct responses. For example, the four-alternative multiple-choice format gave subjects at least a 25% chance of guessing the correct response. This procedure succeeded in producing a relatively large proportion of low-confidence correct responses (at least one third of all responses in each experiment). The finding that low-confidence correct responses benefit from feedback is also consistent with studies in which feedback study time is allowed to vary. For example, test takers generally view feedback on low-confidence correct responses for more time than they do after high-confidence correct responses (e.g., Kulhavy et al., 1976, 1979; Webb, Pridemore, Stock, Kulhavy, & Henning, 1997). Taken as a whole, these findings show that learners utilize feedback after low-confidence correct responses to improve subsequent retention.

Why does feedback enhance retention for low-confidence correct responses? As we briefly described in the introduction, providing feedback after low-confidence correct responses might enhance retention in two ways: (a) strengthening the association between the cue and response and (b) inhibiting competing responses. To understand why this might be true, it helps to consider why low-confidence correct responses are produced. Some researchers consider all low-confidence responses, whether correct or incorrect, to represent a situation in which the subject has insufficient knowledge of the material and thus would benefit more from further instruction than from feedback (Kulhavy, 1977; Kulhavy & Stock, 1989). Certainly, low-confidence correct responses can be lucky guesses, especially on a multiple-choice test where the chance of selecting the correct response is often 20% or greater. However, there are at least two other possible causes. First, subjects might produce a correct response based on partial knowledge and/or familiarity but not be confident that it is the correct response. In this situation, the correct response is already known, and therefore what is needed is for the association between the response and the cue to be strengthened. Second, subjects might give a correct response a low-confidence judgment because they had trouble choosing between two equally attractive responses but happened to volunteer the correct one. In this situation, the association between the cue and correct response must be strengthened, but the competing response must be also inhibited. In both situations, feedback is critical because it first corrects the metacognitive error and then enables the subject to engage the appropriate mechanisms to enhance retention.

The second novel finding from our experiments was that providing feedback after the initial multiple-choice test enhanced the accuracy of confidence judgments on the final test. Subjects were better able to discriminate between correct and incorrect responses on the final test if they had been given feedback on the prior test. This improvement in metacognitive monitoring was evident in both global (overall mean proportion of correct responses and confidence judgments) and relative (item-by-item correspondence between confidence judgments and the proportion correct) assessments of metacognitive accuracy. Previous studies that have investigated the effect of feedback on subsequent confidence judgments have found that feedback can improve both calibration (Lichtenstein & Fischoff, 1980) and resolution (Sharp, Cutler, & Penrod, 1988). However, these studies differ from ours in that feedback was provided in the form of a global assessment of performance (e.g., overall proportion correct) rather than for each item. Presumably, such global feedback might lead subjects to change their overall pattern of responding on future tests (e.g., being more conservative in their confidence judgments). Although feedback on individual responses might have produced overall bias in future metacognitive judgments in our study, it seems more likely that the improvement in metacognitive accuracy is the result of eliminating any discrepancy between perceived and actual correctness of responses.

Finally, it is interesting to note that in Experiment 1, we observed the hypercorrection effect (Butterfield & Metcalfe, 2001, 2006), wherein high-confidence errors were more likely to be corrected on a final test than low-confidence errors. However, we did not observe the effect in Experiment 2, and there was a numerical trend toward a hypocorrection effect in which high-confidence errors were less likely to be corrected relative to low-confidence errors. This result suggests that the hypercorrection effect may be relatively transient. All the studies that have reported a hypercorrection effect have used brief retention intervals (e.g., 5 min; Butterfield & Metcalfe, 2001). Studies that have used longer retention intervals (e.g., 1 week; Pashler et al., 2005) have failed to find the effect, as did we after a 2-day retention interval (but see Kulhavy et al., 1976). The transience of the hypercorrection effect may be the result of the gradual recovery of the original error response as the retention interval increases, similar to the recovery of the A–B pair (and extinction of the A–C pair) in the classic retroactive inference paradigm (Briggs, 1954). Nevertheless, relatively few high-confidence errors were produced

in the current experiment and thus this finding should be interpreted with caution. Of course, in both experiments we showed that feedback increased retention of low-confidence correct responses, regardless of the retention interval before the final test.

The present results have implications for the importance of providing feedback in educational settings. Many of the methods commonly used in education undermine the potential benefits of testing. A prime example is the variable sorts of feedback provided after classroom tests. As shown in many studies, feedback is a critical aspect to learning, but instructors' policies in providing it vary considerably, ranging from comprehensive feedback after each testing occasion to little or no feedback at all. The latter situation is increasingly prevalent in university settings, where large class sizes and repeated teaching responsibilities lead educators to retain completed examinations to guard their test banks. Nevertheless, when feedback is given (i.e., other than a grade or numerical score), the focus is generally on incorrect answers. Students tend to look for the red ink and concentrate on figuring out why they got the answer wrong. The current research suggests that educators should try to give comprehensive feedback (i.e., on both correct and incorrect answers) whenever possible. Such feedback may be particularly important after multiple-choice tests that expose test takers to incorrect information in the form of lures. Attractive alternatives can lead test takers to change their response on a later test (Higham & Gerrard, 2005) and to endorse a lure on an initial test often resulting in it being produced on a subsequent test (Butler, Marsh, Goode, & Roediger, 2006; Butler & Roediger, 2008; Roediger & Marsh, 2005).

Finally, it is worth noting that the tests used in most educational settings are essentially forced report, regardless of test format (e.g., multiple-choice, short answer, and so forth). There is usually no penalty for an incorrect response, which encourages students to provide an answer to every question, even if they have to guess, to maximize their score on the test. (An exception is standardized tests, like the SAT, which penalize students for incorrect responses by deducting points.) The use of such a strategy by students is likely to be even more prevalent in multiple-choice testing, where there is a good chance of guessing the correct answer. Thus, one could argue that the use of an initial forced-report test in the present research is more consistent with the methods used in education than the initial free-report tests used in most of the previous studies that have investigated the effect of feedback on initially correct responses.

In conclusion, the current experiments provide clear evidence that low-confidence correct responses do benefit from feedback and that feedback improves students' metacognitive judgments about their knowledge. Taken together, the two novel findings support the idea that a low-confidence correct response represents an error in metacognitive monitoring that can be corrected through feedback. Providing feedback after low-confidence correct responses enables learners to eliminate the discrepancy between perceived and actual correctness of the response. Feedback after both correct and incorrect responses on tests is a critical aspect of learning.

References

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct.* Retrieved June 1, 2003, from http://www.apa.org/ethics/code2002.html

Anderson, R. C., Kulhavy, R. W., & Andre, T. (1971). Feedback procedures in programmed instruction. *Journal of Educational Psychology, 62,* 148–156.

Azevedo, R., & Bernard, R. M. (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13,* 111–127.

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.

Barnes, A. E., Nelson, T. O., Dunlosky, J., Mazzoni, G., & Narens, L. (1999). An integrative system of metamemory components involved in retrieval. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 287–313). Cambridge, MA: MIT Press.

Brackbill, Y., & Kappy, M. S. (1962). Delay of reinforcement and retention. *Journal of Comparative and Physiological Psychology, 55,* 14–18.

Briggs, G. E. (1954). Acquisition, extinction, and recovery functions in retroactive inhibition. *Journal of Experimental Psychology, 47,* 285–293.

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., 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.

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.

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.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 1491–1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning, 1,* 69–84.

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

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

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

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of F distribution in multivariate analysis. *Annals of Mathematical Statistics, 29,* 885–891.

Guthrie, J. T. (1971). Feedback and sentence learning. *Journal of Verbal Learning and Verbal Behavior, 10,* 23–28.

Higham, P. A., & Gerrard, C. (2005). Not all errors are created equal: Metacognition and changing answers on multiple-choice tests. *Canadian Journal of Experimental Psychology, 59,* 28–34.

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

Karpicke, J. D., & Roediger, H. L. (2008, February 15). The critical importance of retrieval for learning. *Science, 319,* 966–968.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language, 48,* 704–721.

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

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity oriented approaches to memory assessment. *Journal of Experimental Psychology: General, 123,* 297–315.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in strategic regulation of memory accuracy. *Psychological Review, 103,* 490–517.

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

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63,* 505–512.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1,* 279–308.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1976). Feedback and response confidence. *Journal of Educational Psychology, 68,* 522–528.

Kulhavy, R. W., Yekovich, F. R., & Dyer, J. W. (1979). Feedback and content review in programmed instruction. *Contemporary Educational Psychology, 4,* 91–98.

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

Lhyle, K. G., & Kulhavy, R. W. (1987). Feedback processing and error correction. *Journal of Educational Psychology, 79,* 320–322.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26,* 149–171.

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16,* 192–201.

Meyer, L. A. (1986). Strategies for correcting students' wrong responses. *The Elementary School Journal, 87,* 227–241.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95,* 109–133.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL-effect." *Psychological Science, 5,* 207–213.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–141). New York: Academic Press.

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,* 3–8.

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

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

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

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

Roediger, H. L., III, & Payne, D. G. (1985). Recall criterion does not affect recall level or hypermnesia: A puzzle for generate/recognize theories. *Memory & Cognition, 13,* 1–7.

Roediger, H. L., Wheeler, M. A., & Rajaram, S. (1993). Remembering, knowing and reconstructing the past. In D. L. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 30., pp. 97–134). New York: Academic Press.

Sassenrath, J. M., & Yonge, G. D. (1968). Delayed information feedback, feedback cues, retention set, and delayed retention. *Journal of Educational Psychology, 59,* 69–73.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh, PA: Psychology Software Tools, Inc.

Sharp, G. L., Cutler, B. L., & Penrod, S. D. (1988). Performance feedback improves the resolution of confidence judgments. *Organizational Behavior and Human Performance, 42,* 271–283.

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

Surber, J. R., & Anderson, R. C. (1975). Delay-retention effect in natural classroom settings. *Journal of Educational Psychology, 67,* 170–173.

Webb, J. M., Pridemore, D. R., Stock, W. A., Kulhavy, R. W., & Henning, J. E. (1997). Remembering responses and cognitive estimates of knowing: The effects of instructions, retrieval sequences, and feedback. *Contemporary Educational Psychology, 22,* 147–164.

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245.

World Book, Inc. (2002). *The 2002 world book encyclopedia* (Vols. 1–25). Chicago: Author.

Received November 16, 2007
Revision received February 25, 2008
Accepted February 29, 2008 ■