*Memory & Cognition*
*2008, 36 (3), 604-616*
*doi: 10.3758/MC.36.3.604*

# Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing

**Andrew C. Butler and Henry L. Roediger III**
*Washington University, St. Louis, Missouri*

Multiple-choice tests are used frequently in higher education without much consideration of the impact this form of assessment has on learning. Multiple-choice testing enhances retention of the material tested (the testing effect); however, unlike other tests, multiple-choice can also be detrimental because it exposes students to misinformation in the form of lures. The selection of lures can lead students to acquire false knowledge (Roediger & Marsh, 2005). The present research investigated whether feedback could be used to boost the positive effects and reduce the negative effects of multiple-choice testing. Subjects studied passages and then received a multiple-choice test with immediate feedback, delayed feedback, or no feedback. In comparison with the no-feedback condition, both immediate and delayed feedback increased the proportion of correct responses and reduced the proportion of intrusions (i.e., lure responses from the initial multiple-choice test) on a delayed cued recall test. Educators should provide feedback when using multiple-choice tests.

The multiple-choice test is a staple of higher education because it provides an efficient and effective measure of student learning (McKeachie, 1999). The popularity of this highly objective testing format has increased over the years, partly due to improvements in technology that make grading multiple-choice tests quick and easy. The multiple-choice test is also highly reliable across scorers, unlike essay tests. For these reasons and others (Frederiksen, 1984), many educators consider the multiple-choice format an optimal method of testing.

Although tests are primarily used as means of assessment, they also affect the knowledge they measure. Taking a test generally improves retention of the material tested—a result commonly referred as the *testing effect* (for a review, see Roediger & Karpicke, 2006a). Multiple-choice tests generally enhance learning as measured on later tests; however, the multiple-choice test presents a unique situation because it exposes students to erroneous information in the form of lure items. By endorsing (or even reading) lure items during the course of taking a multiple-choice test, students may acquire incorrect knowledge (see, e.g., Butler, Marsh, Goode, & Roediger, 2006; Roediger & Marsh, 2005). As a result, the value of using multiple-choice testing as a learning tool will be enhanced to the extent that the positive effects (increased retention) can be maximized and the negative effects (the acquisition of misinformation) can be minimized. Providing feedback after a multiple-choice test may promote optimal learning by helping students to maintain correct responses and correct errors (for a review, see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991).

The present research sought to identify the circumstances under which multiple-choice testing is most beneficial to learning. More specifically, we examined two factors that may influence the positive and negative effects of taking a multiple-choice test: the amount of study prior to the test and the number of lures on the multiple-choice test. Considering the first factor, students vary considerably in their preparation for a test. Obviously, the lack of prior study will result in poor performance on the multiple-choice test, decreasing the positive effects of testing, because students must answer questions correctly in order to benefit from testing. Roediger and Marsh (2005; see too Butler et al., 2006) showed this effect, but they also showed that the negative effects of taking a multiple-choice test were greater when students had studied less, which also makes sense. When students know little and guess, they select a lure, and then (if they are not corrected with feedback) they may believe that they made a correct choice and provide the answer on a later test. The other factor of interest is the number of lures on the test. Instructors often prefer to use multiple lures (three or four lures in addition to the correct answer is typical) in order to drive down the probability of guessing correctly. However, increasing the number of lures produces the same negative effect of acquiring false knowledge, because students are exposed to more erroneous information.

Instructors vary greatly in whether they give feedback on multiple-choice tests. Some do so as a matter of course, but others protect their test banks and do not give students feedback unless they make an appointment to see their exam in the instructor's office. In the present experiment, we were interested in whether providing feedback would influence the magnitude of positive testing effects (we predicted they would) and at the same time overcome the

**A. C. Butler, butler@wustl.edu**

negative effects of such tests (an issue in more doubt). In addition, we were interested in whether the timing of the feedback would have differential effects. Before describing our experiment, we will briefly summarize previous research relevant to this study.

### Testing Benefits Retention

The act of retrieving information from memory serves to modify the memory trace and increase the probability of future retrieval success (see, e.g., Carrier & Pashler, 1992; McDaniel & Masson, 1985; Tulving, 1967; Wheeler & Roediger, 1992). Because of the mnemonic benefit conferred by retrieval, many researchers have argued that tests should be used as learning tools in the classroom (e.g., Bangert-Drowns, Kulik, & Kulik, 1991; Foos & Fisher, 1988; Glover, 1989; Jones, 1923–1924; Roediger & Karpicke, 2006b). Indeed, recent research suggests that testing produces long-lasting benefits for retention of complex, educationally relevant materials (Butler & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007).

With respect to multiple-choice tests in particular, previous research has shown that taking an initial multiple-choice test leads to superior performance on a subsequent test in comparison with not taking an initial test, regardless of whether the final test format is multiple-choice (see, e.g., Duchastel & Nungester, 1982; McDaniel et al., 2007) or cued recall (e.g., Butler & Roediger, 2007; McDaniel & Masson, 1985; Roediger & Marsh, 2005). For example, Butler, Karpicke, and Roediger (2007) had students study prose passages and then take an initial multiple-choice test that covered the material in the passages. On a subsequent cued recall test, subjects produced a higher proportion of correct responses for previously tested items than for a subset of the items that were not tested on the initial multiple-choice test.

### Exposing Students to Misinformation

Although testing generally enhances retention of the material, studies that utilize multiple-choice tests have also revealed negative consequences of exposing students to incorrect information. Taking a multiple-choice test leads subjects to assign higher "truth" values to false statements that appeared on the earlier multiple-choice test than to novel false statements (Toppino & Luipersbeck, 1993). Similarly, research has shown that exposure to incorrect spellings (Brown, 1988; Jacoby & Hollingshead, 1990) or false facts embedded within a passage (Marsh, Meade, & Roediger, 2003) can interfere with memory for correct spellings and facts, respectively. In addition, exposure to incorrect information can have a negative effect on subsequent test performance, even when the exposure occurs after the initial test (see, e.g., Brown, Schilling, & Hockensmith, 1999). However, the most detrimental effect of multiple-choice testing probably occurs when students endorse a lure, believing it to be the correct response. After selecting a lure on an initial multiple-choice test, students tend to produce that lure when prompted with the same question on a subsequent cued recall test (Butler et al., 2006; Roediger & Marsh, 2005). Moreover, Schooler, Foster, and Loftus (1988) found an impairing effect of endorsing a lure even when the endorsed lure was not included on the final test, indicating that the negative effects of committing an error on a multiple-choice test are not completely due to a bias for maintaining the same response. In essence, the persistence of incorrect responses indicates that students are acquiring false knowledge through multiple-choice testing—an outcome that is especially troubling given the power of testing to enhance retention (of incorrect facts, in this case).

A primary determinant of the magnitude of these negative effects is the level of performance on the multiple-choice test: As students commit more errors, the opportunities for acquiring false knowledge grow. One factor that influences the level of performance is test difficulty. Although test difficulty can be operationalized in many ways, a simple and systematic method for manipulating test difficulty is varying the number of multiple-choice alternatives. For example, Roediger and Marsh (2005) had subjects read prose passages and then take a multiple-choice test that contained equal numbers of two-, four-, and six-alternative questions. As the number of alternatives on the initial multiple-choice test increased, the proportion of correct responses on the multiple-choice test decreased. Then, after a delay, subjects took a comprehensive cued recall test. Increasing the number of lures on the multiple-choice test led to a decrease in the proportion of correct responses and an increase in the proportion of lures produced on the later cued recall tests. As was noted previously, Roediger and Marsh showed that the amount of prior study affected performance on an initial multiple-choice test and, as a result, on the final cued recall test as well. When students were not given the opportunity to read the passages and took the initial test "cold," they performed worse on both the initial multiple-choice test and the final cued recall test than they did when they studied the material.

### Feedback Boosts Retention and Corrects Errors

One potential method for increasing the benefits of testing and reducing the negative effects of exposing students to misinformation is to provide feedback after testing. Feedback allows students to correct errors (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991) and maintain correct responses (Butler, Karpicke, & Roediger, in press), resulting in superior performance on a subsequent test in comparison with no feedback (McDaniel & Fisher, 1991). The type of feedback provided can range from a simple indication of whether the response is correct or incorrect (see, e.g., Schroth, 1977) to an elaborate explanation of why a certain response is correct (e.g., Tait, Hartley, & Anderson, 1973), to a full re-presentation of the original study materials that allows students to determine the accuracy of their responses (Agarwal, Karpicke, Kang, Roediger, & McDermott, in press). Perhaps the most critical piece of information in the feedback message is the correct response, which permits students to both evaluate the accuracy of their knowledge and encode the correct response, if necessary. Consequently, providing the correct response is more effective than simply indicating whether the response is correct or incorrect (e.g., Gilman, 1969;

Pashler, Cepeda, Wixted, & Rohrer, 2005; for a meta-analysis see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991). The feedback message may also include other information, such as a re-presentation of the question and/or the student's prior response, both of which help to re-establish the original context and permit the student to fully process the feedback. Such contextual reinstatement is especially critical when feedback is given after a delay.

Another important consideration is when to deliver the feedback to the student. In contrast with the general consensus about the types of feedback that work best, there is substantial disagreement about the optimal timing of feedback (for a review, see Kulik & Kulik, 1988). Motivated by behavioral theories of reinforcement, some researchers have argued that feedback should be given as soon as possible after an error in order to eliminate incorrect responses (see, e.g., Skinner, 1954), a position supported by numerous studies that have conceptualized feedback as reinforcement (e.g., Angell, 1949; Bourne, 1957; Paige, 1966; Sullivan, Schutz, & Baker, 1971). However, others have contended that feedback is functionally different from reinforcement and that delayed feedback is more effective because it gives errors a chance to dissipate, making the process of learning the correct response easier (e.g., Kulhavy, 1977; Kulhavy & Anderson, 1972; Kulhavy & Stock, 1989). Indeed, many studies have found delayed feedback to be more beneficial to the learning and retention of information than immediate feedback (e.g., Brackbill, Bravos, & Starr, 1962; Butler et al., 2007; Sturges, 1969; Surber & Anderson, 1975). For the most part, these disparate results have yet to be reconciled, feeding the debate about the optimal timing of feedback.

### Report Option: Responding and Belief in Correctness

In most testing situations in the classroom, students are not penalized for guessing on multiple-choice tests; rather, the instructor simply calculates the proportion of answers that are correct. A final interest in the present research was trying to assess both student responding and student belief in what answers were actually correct. If students acquire information from a multiple-choice test (whether correct or incorrect), do they believe that the information is correct? When students retrieve information on a test, they can assess various aspects of knowledge to determine its accuracy. In most metamemory situations, such monitoring processes lead to relatively accurate memory reports, because people control whether or not to report the information retrieved. However, educational testing is one area in which forced report dominates. On most classroom tests, the potential for full or partial credit exists, and there is often no penalty for guessing. As a result, students are encouraged to answer every question regardless of the perceived accuracy of the candidate response, making it hard to ascertain whether or not they actually believe in the correctness of any given response.

An interesting way to assess students' belief in the correctness of their knowledge is to manipulate the report option on the final test. Forced report instructions require subjects to respond to every question or to produce a predetermined number of responses, generally resulting in the production of a large amount of incorrect information (see, e.g., Roediger & Payne, 1985). In contrast, free report instructions allow subjects to volunteer or withhold responses, which often leads to enhanced memory accuracy in comparison with forced report (e.g., Koriat & Goldsmith, 1994, 1996). Thus, a manipulation of report option will help to gauge what people really know and what they will report on a test.

### The Present Research

The present research examined the effects of three variables on learning from a multiple-choice test in hopes of finding situations that maximize positive effects of multiple-choice testing while minimizing the negative effects. Students were randomly assigned to one of three initial study conditions: no exposure to the material (no study), a brief reading of the material (study), or a brief reading of the material combined with a rereading of the key sentences (restudy). The no-study and study conditions were similar to those employed in previous research on this topic (e.g., Roediger & Marsh, 2005). The restudy condition was designed to boost performance on the multiple-choice test above that of the study condition. The rereading of the key sentences was intended to be analogous to students reading through their notes or returning to the parts of the passage that they highlighted. Next, all subjects took a multiple-choice test with equal numbers of two-, four-, and six-alternative questions. For each response on the multiple-choice test, they received no feedback, immediate feedback, or delayed feedback. An additional subset of items was never tested to serve as a baseline for comparison with the testing conditions. Finally, after a 1-week delay, subjects returned for a comprehensive cued recall test. This final test used Koriat and Goldsmith's (1996) procedure in which a forced report phase that required guessing was followed by a free report phase in which the responses were judged for correctness.

On the basis of the testing-effect literature reviewed previously, we predicted an overall benefit in performance on the final cued recall test for items tested on the initial multiple-choice test relative to items not initially tested. However, we expected the magnitude of this testing effect to be determined by the amount of prior study and the number of multiple-choice alternatives. More specifically, a greater amount of prior study should lead to better performance on the initial multiple-choice test, resulting in a higher proportion of correct responses and a lower proportion of intrusions (lure responses from the initial multiple-choice test) on the final cued recall test. Similarly, fewer multiple-choice alternatives should lead to better performance on the initial multiple-choice test, resulting in a higher proportion of correct responses and lower proportion of intrusions on the final cued recall test. Thus, performance on the initial multiple-choice test was expected to play a large role in determining performance on the subsequent cued recall test when feedback was not provided. We expected feedback to have positive effects on both correct and incorrect responses (perhaps elimi-

nating the negative effects of multiple-choice testing). We anticipated that feedback would allow students to correct their errors, leading to a reduction of the proportion of intrusions produced on the cued recall test, and that feedback would help to maintain correct responses made on the initial multiple-choice test.

A primary purpose of the present experiment was to explore any novel interactions among the three variables of interest. Although each of the variables included has been investigated in previous research, no study has manipulated all three within a single experiment. As was described previously, the amount of prior study and the number of multiple-choice alternatives variables were expected to have separate and additive effects on cued recall test performance in the absence of feedback. However, it was less clear whether (and how) the pattern of cued recall test performance produced by these two variables would be altered when feedback was provided. There were at least two potential hypotheses about how feedback would interact with the amount of prior study and the number of multiple-choice alternatives variables. One possible outcome was that the provision of feedback would increase the proportion of correct responses and reduce the proportion of intrusions, but would leave the overall pattern of effects observed in the no-feedback condition intact (e.g., a smaller increase in production of lures on the final test as a function of number of alternatives on the prior test). Another possible outcome would be for feedback to completely eliminate the effects of prior study and the number of multiple-choice alternatives on final cued recall, bringing performance in all conditions up to the same level. Similarly, several hypotheses could be generated about the optimal timing of feedback. However, we predicted that delayed feedback would lead to superior performance in comparison with immediate feedback because of the added benefits of allowing the incorrect response to dissipate (see, e.g., Kulhavy & Anderson, 1972) and providing a spaced presentation of the material (see Dempster, 1989) in the case of delayed feedback.

We also manipulated report option on the final cued recall test. We expected that free report would reduce the overall proportion of intrusions in comparison with forced report. However, students' ultimate success at restricting their report to correct responses hinges upon their ability to differentiate between correct and incorrect responses. If students cannot effectively make such a distinction, then free report may result in the reduction of both correct and incorrect responses in comparison with forced report. Thus, manipulating report option permits us to examine students' metamemorial knowledge of their responding.

## METHOD

### Subjects

Seventy-two undergraduate psychology students at Washington University in St. Louis participated for course credit or pay ($20). They were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" (American Psychological Association, 2002).

### Design

The experiment used a 3 (amount of prior study: no study, study, restudy) × 3 (number of multiple-choice alternatives: two, four, six) × 3 (feedback condition: no feedback, immediate feedback, delayed feedback) × 2 (report option: forced report and free report) mixed design. In addition, the experiment included a control condition in which no multiple-choice test was given on some material (no test). This condition could not be crossed with the number of multiple-choice alternatives factor and is therefore not fully incorporated into the main design (see the counterbalancing section). The number of multiple-choice alternatives and the feedback-condition variables were manipulated within subjects and between materials. The amount-of-prior-study variable was manipulated between subjects. Report option was manipulated within subjects during the cued recall test on all items.

### Materials and Counterbalancing

Stimuli consisted of a set of 12 prose passages covering a variety of historical topics (e.g., the Khmer Rouge). The passages were developed using information obtained from two online encyclopedias (www.encyclopedia.com and www.en.wikipedia.org). Each passage contained approximately 400 words arranged into four paragraphs. Four facts were identified in each passage, with each fact corresponding to one of the four paragraphs. The "key sentences" that subjects in the restudy condition were given to reread consisted of the sentences from the passage that contained these facts. A question was designed to test each fact from the passage using a fill-in-the-blank format. The correct response to each question (henceforth referred to as the *target*) consisted of a short phrase between one and three words in length. For example, *Many of the leaders of the Khmer Rouge were educated in* _____ (target: *France*). For the purposes of the multiple-choice test, five plausible lures were developed for each question for the six-alternative condition (the correct answer plus five lures). Two lures were randomly removed to create the four-alternative condition, and four lures were randomly removed to create the two-alternative condition. No lure or target appeared as a potential answer to another question.

The experimental materials were counterbalanced in several ways. First, across subjects, each passage was used in each condition an equal number of times. In order to accomplish this, the materials were divided into four sets of three passages. The four sets were then rotated through the three feedback conditions (no feedback, immediate feedback, delayed feedback) and the no-test control condition to create four versions of the multiple-choice test. Then, within each version, the three passages in each test condition were rotated through the number of multiple-choice alternative conditions (two, four, six) to create a total of 12 versions of the multiple-choice test. Second, for each of the 12 versions of the multiple-choice test, the target appeared equally in each possible position in comparison with the lures across the items within each multiple-choice alternatives condition. For example, in the four-alternative condition, the target appeared three times in the first, second, third, and fourth positions.

### Procedure

The entire experiment was conducted on PCs using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002) and involved two sessions that were spaced 1 week apart.

**Session 1**. Subjects began with a study phase in which treatment was given according to a randomly assigned condition: no study, study, or restudy. Those in the no-study condition did not read the passages and skipped ahead to the filler task (playing a Pac-Man video game for 5 min) that separated the study phase from the multiple-choice test. Each subject in the study condition read through all 12 passages in a different order, which was randomly determined by the computer at the start of the session. The passages were presented one half at a time (approximately 200 words), with each half appearing on the screen for 30 sec. Subjects in the restudy condition read through the full set of passages in the same manner

as did those in the study condition. However, after reading the passages, they engaged in the filler task for 2 min and then reread the key sentences that contained information on the tests from each passage. The key sentences were grouped by passage, and each group of sentences was displayed for 30 sec. Again, the computer randomly determined a different presentation order for each subject.

Prior to taking the multiple-choice test, all subjects engaged in a filler task (Pac-Man) for 5 min. After the filler task, they received instructions about the multiple-choice test (those in the no-study condition were told that it was a general knowledge test). The test was self-paced and consisted of 42 questions. The first 6 questions were always filler items in order to ensure that subjects did not have information from the last passage in working memory. Subjects in the study and restudy conditions were told that these items were practice questions. The remaining 36 questions corresponded to the passages in the three feedback conditions (no feedback, immediate feedback, delayed feedback). With the exception of the filler questions, which were always presented first, the computer randomized the presentation of the questions so that each subject received a different order. Each question was presented at the top of the screen with the alternatives listed below. The position of the target in comparison with the lures was counterbalanced as described previously. Subjects were instructed to press the button corresponding to the position of the correct answer (e.g., press 1 for the alternative in position 1). The position number preceded each alternative to facilitate responding. Feedback was presented for 10 sec either immediately after the response (immediate feedback) or at the end of the test (delayed feedback). In order to equate the amount of time spent on each question, a message ("Please wait for the next question to load") was displayed for 10 sec after items in the no-feedback condition. Feedback consisted of an indication of the accuracy of the response (correct–incorrect), a re-presentation of the question, the response selected, and the correct response. After students finished the multiple-choice test, they were reminded of the second session and dismissed.

**Session 2**. One week after the first session, subjects returned to take a final, comprehensive cued recall test that incorporated a procedure adopted from Koriat and Goldsmith (1994, 1996) in which a forced report phase is followed by a free report phase. The questions on the final cued recall test were exactly the same as those on the initial multiple-choice test (with the addition of the subset of untested items). Similar to the multiple-choice test, each fill-in-the-blank question on the cued recall test was presented at the top of the screen (e.g., *Many of the leaders of the Khmer Rouge were educated in* _____). However, instead of choosing from a list of alternatives, subjects had to produce the response from memory and type it in by using the keyboard. The test was self-paced and consisted of two phases. In the forced report phase, subjects were given instructions to provide a response to every question, even if they had to guess. After each response, they were asked to rate their confidence in the response on a scale of 0 to 100. After answering all questions, students proceeded to the free report stage, in which they were given the opportunity to go back through their responses from the forced report phase and to decide whether to keep or omit each response. They were shown the question and their response, but not their confidence estimate. The stated goal was to keep as many correct and omit as many incorrect answers as possible. After the second session was complete, subjects 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. In the analysis of repeated measures, a Greenhouse–Geisser correction was used for violations of the sphericity assumption (Geisser & Greenhouse, 1958).

**Table 1**
**Proportion Correct on the Initial Multiple-Choice Test As a Function of the Number of Multiple-Choice Alternatives and the Amount-of-Prior-Study Condition**

| Study Condition | Number of Alternatives | | | |
|---|---|---|---|---|
| | Two | Four | Six | *M* |
| No study | .56 | .34 | .28 | .39 |
| Study | .69 | .56 | .47 | .57 |
| Restudy | .84 | .72 | .70 | .75 |
| *M* | .69 | .54 | .48 | .57 |

### Initial Multiple-Choice Test

Table 1 displays the proportion of correct responses on the initial multiple-choice test as a function of the number of multiple-choice alternatives and the amount of prior study (the data are collapsed across feedback condition because the manipulation had not yet been introduced). The data were analyzed by way of a 3 (amount of prior study) × 3 (number of multiple-choice alternatives) repeated measures ANOVA. This analysis revealed two significant main effects, both of which were expected. First, there was a main effect of prior study [$F(2,69) = 66.25$, $MS_e = 0.035$, $\eta_p^2 = .66$] in which the restudy condition produced a higher proportion of correct responses than did the study condition [$t(46) = 13.68$, $SED = 0.026$, $d = 1.29$, $p_{rep} = 1.00$ ($p_{rep}$ is an estimate of the probability of replicating the direction of an effect; see Killeen, 2005)], which in turn was higher than the no-study condition [$t(46) = 5.99$, $SED = 0.031$, $d = 1.20$, $p_{rep} = 1.00$]. Second, there was a main effect of the number of multiple-choice alternatives in which the proportion of correct responses decreased as the number of alternatives increased. Both the linear [$F(1,69) = 73.76$, $MS_e = 0.022$, $\eta_p^2 = .52$] and quadratic [$F(1,69) = 7.05$, $MS_e = 0.018$, $\eta_p^2 = .09$] effects were significant. No other effects approached significance. Our interest centered on how multiple-choice performance would affect the cued recall test that was given a week later.

### Final Cued Recall Test: Forced Report

The data from the forced report phase of the cued recall test were analyzed first. There were three potential outcomes for each item in this phase: a correct response (correct), the production of a lure from the prior multiple-choice test (intrusion), or an incorrect response that had not appeared previously as a lure (incorrect other).[1]

**Correct responses**. The upper portion of Table 2 shows the proportion of correct responses as a function of the amount of prior study and feedback condition, with the no-test condition included for comparison purposes. The data are collapsed across the number of multiple-choice alternatives because this variable did not interact with any other variable of interest, as reported below (see the Appendix for the full data). The data were analyzed with a 3 (amount of prior study) × 3 (number of multiple-choice alternatives) × 3 (feedback condition) repeated measures ANOVA. First, a main effect of feedback condition was observed [$F(2,138) = 66.79$, $MS_e = 0.051$, $\eta_p^2 = .49$] in which delayed feedback led to a higher proportion of cor-

**Table 2**
**Proportion Correct on the Final Cued Recall Test As a Function of Amount-Of-Prior-Study and Feedback Conditions (Including the No-Test Condition)**

| Study Condition | Feedback Condition: No Test | No Feedback | Immediate Feedback | Delayed Feedback | *M* |
|---|---|---|---|---|---|
| | | Forced Report | | | |
| No study | .10 | .18 | .42 | .57 | .32 |
| Study | .11 | .33 | .43 | .54 | .35 |
| Restudy | .22 | .41 | .50 | .57 | .43 |
| *M* | .14 | .31 | .45 | .56 | .37 |
| | | Free Report | | | |
| No study | .04 | .12 | .39 | .52 | .27 |
| Study | .06 | .28 | .38 | .50 | .31 |
| Restudy | .16 | .35 | .46 | .52 | .37 |
| *M* | .09 | .25 | .41 | .51 | .32 |

Note—Data have been collapsed across the number of initial multiple-choice alternatives.

rect responses than did immediate feedback [$t(71) = 4.79$, $SEM = 0.022$, $d = .56$, $p_{rep} = 1.00$] and immediate feedback was higher than no feedback [$t(71) = 6.40$, $SEM = 0.022$, $d = .69$, $p_{rep} = 1.00$]. Second, there was a marginally significant main effect of prior study [$F(2,69) = 2.92$, $MS_e = 0.197$, $p = .06$]. Pairwise comparisons showed only one significant difference: The restudy condition led to a higher proportion of correct responses than did the no-study condition [$t(46) = 3.02$, $SEM = 0.035$, $d = .75$, $p_{rep} = .98$]. Finally, there was also a significant interaction between the feedback condition and prior study [$F(4,138) = 5.77$, $MS_e = 0.051$, $\eta_p^2 = .14$]. The proportion of correct responses increased substantially as the amount of prior study increased (i.e., going from no study to study to restudy) in the no-feedback condition, but not in the delayed-feedback condition. In fact, in the delayed-feedback condition, it did not matter whether students had previously studied the material at all; performance was roughly the same in all three of the study conditions. Finally, the proportion of correct responses differed only slightly as a function of the number of multiple-choice alternatives (see the Appendix). There was a numerical trend in the no-feedback condition in which the proportion of correct responses decreased as the number of multiple-choice alternatives increased. However, the linear trend did not reach significance when tested with an ANOVA conducted on the data from no-feedback condition alone [$F(1,69) = 2.60$, $MS_e = 0.048$, $p = .11$]. Nevertheless, the numerical trend was in the right direction, and there may have been insufficient power to detect the effect (observed power = .36). Other research making this comparison has found effects of roughly the same size as that seen in the present study (e.g., Roediger & Marsh, 2005). No other effects approached significance. Although the main analyses did not include the no-test control condition, an additional *t* test revealed that the no-feedback condition (which received a test) produced a significantly greater proportion of correct responses than did the no-test condition [$t(71) = 9.11$, $SEM = 0.018$, $d = .97$, $p_{rep} = 1.00$], showing the basic testing effect.

**Intrusions**. The upper portion of Table 3 displays the proportion of intrusions made on the cued recall test as a function of the amount-of-prior-study and feedback conditions, with the no-test condition included for comparison purposes (the data are again collapsed across the number of multiple-choice alternatives because this variable did not interact with any other variable of interest, as reported below; see the Appendix). The data were again analyzed by a 3 (amount of prior study) × 3 (number of multiple-choice alternatives) × 3 (feedback condition) repeated measures ANOVA. Several significant effects emerged. First, there was a main effect of feedback condition [$F(2,138) = 22.62$, $MS_e = 0.031$, $\eta_p^2 = .25$] in which no feedback produced a significantly higher proportion of intrusions than did the immediate feedback [$t(71) = 5.90$, $SEM = 0.018$, $d = .65$, $p_{rep} = .99$] and delayed feedback [$t(71) = 4.65$, $SEM = 0.019$, $d = .81$, $p_{rep} = 1.00$]. Second, a main effect of prior study was observed [$F(2,69) = 3.07$, $MS_e = 0.034$, $\eta_p^2 = .08$] in which the no-study condition led to the production of more lures than did the restudy condition [$t(46) = 2.45$, $SEM = 0.016$, $d = .35$, $p_{rep} = .95$]. Third, there was a linear trend in the number of multiple-choice alternatives [$F(1,69) = 5.30$, $MS_e = 0.039$, $\eta_p^2 = .07$] in which a greater number of alternatives led to a higher proportion of intrusions, as shown in the Appendix. Finally, there was an interaction between prior study and feedback condition [$F(4,138) = 5.88$, $MS_e = 0.031$, $\eta_p^2 = .15$]. Greater amount of prior study decreased the proportion of intrusions in the no-feedback condition (as in Roediger & Marsh, 2005), but the amount of prior study was neutralized by feedback, which reduced the number of intrusions. No other effects reached significance. As before, the main analysis did not include the no-test condition, but an additional *t* test confirmed that the no-feedback condition produced a higher proportion of intrusions than did the no test control condition in which no lures had been shown [$t(71) = 4.49$, $SEM = 0.019$, $d = .65$, $p_{rep} = .99$]. Feedback on the multiple-choice test reduced lure intrusions to this baseline level.

**Table 3**
**Proportion Intrusions on the Final Cued Recall Test As a Function of Amount-of-Prior-Study and Feedback Conditions (Including the No-Test Condition)**

| Study Condition | Feedback Condition: No Test | No Feedback | Immediate Feedback | Delayed Feedback | *M* |
|---|---|---|---|---|---|
| | | Forced Report | | | |
| No study | .17 | .32 | .14 | .13 | .19 |
| Study | .16 | .25 | .18 | .13 | .18 |
| Restudy | .15 | .16 | .14 | .15 | .15 |
| *M* | .16 | .24 | .15 | .14 | .17 |
| | | Free Report | | | |
| No study | .05 | .19 | .07 | .07 | .10 |
| Study | .06 | .15 | .08 | .07 | .09 |
| Restudy | .05 | .08 | .09 | .07 | .07 |
| *M* | .05 | .14 | .08 | .07 | .09 |

Note—Data have been collapsed across the number of initial multiple-choice alternatives.

### Final Cued Recall Test: Free Report[2]

In the free report phase, subjects had the option of keeping or omitting each response they made during the forced report phase. The subsequent analysis focuses on the responses that subjects chose to keep in order to examine the extent to which students believed that the information learned on the multiple-choice test was true. The free report phase follows the forced report phase and could be affected by the earlier phase, so the data and analysis presented below should be interpreted with this influence in mind.

**Correct responses**. The lower portion of Table 2 displays the proportion of correct responses kept during the free report phase as a function of prior study and feedback condition (including the no-test condition for comparison).[3] A 3 (amount of prior study) × 3 (feedback condition) repeated measures ANOVA was conducted, and it revealed the same effects as did the forced report analysis: a main effect of feedback condition [$F(2,138) = 77.634$, $MS_e = 0.017$, $\eta_p^2 = .53$], a marginally significant main effect of prior study [$F(2,69) = 2.762$, $MS_e = 0.065$, $p = .07$], and an interaction between the prior study and feedback conditions [$F(4,138) = 6.053$, $MS_e = 0.0167$, $\eta_p^2 = .15$]. This interaction appears to be driven by the no-feedback condition in the same manner as that in the forced report data. The only difference between the forced and free results is that the proportion correct in each cell has decreased by between 3% and 6% in free report due to response withholding. When report option (forced, free) was entered into the analysis as a within-subjects variable, a main effect of report option did emerge [$F(1,69) = 82.67$, $MS_e = 0.004$, $\eta_p^2 = .55$], indicating that free report instructions led to a reduction in the proportion of correct responses. Report option did not interact with any of the other factors.

**Intrusions**. The lower portion of Table 3 displays the proportion of intrusions kept during the free report phase as a function of the prior-study and feedback conditions (including the no-test condition). A 3 (amount of prior study) × 3 (feedback condition) repeated measures ANOVA revealed a main effect of feedback condition [$F(2,138) = 16.65$, $MS_e = 0.007$, $\eta_p^2 = .19$] and an interaction between prior study and feedback condition [$F(4,138) = 4.352$, $MS_e = 0.007$, $\eta_p^2 = .11$]. This interaction probably represents the differential effect of the amount of prior study on the no-feedback condition. A greater amount of prior study decreased the proportion of intrusions in the no-feedback condition, but had no effect on any of the other feedback conditions. Overall, students managed to reduce the proportion of intrusions relative to the forced report phase, and the magnitude of this reduction differed slightly across the feedback conditions. When report option (forced, free) was entered into the analysis, a main effect of report option emerged [$F(1,69) = 229.90$, $MS_e = 0.005$, $\eta_p^2 = .77$], indicating that subjects were able to reduce the proportion of intrusions in free report as opposed to forced report. Report option did not interact with any of the other factors.

### Final Cued Recall: Confidence

Performance on the final cued recall test was compared with the confidence estimate given by subjects during the forced report phase. The subjects were asked to rate their confidence in each response on a scale of 0–100, in which "0" indicated no confidence (i.e., a pure guess) and "100" indicated that the response was definitely correct. Of interest was whether any of the manipulated variables would influence the subjects' ability to assess the accuracy of their knowledge. Feedback on the multiple-choice test increased the subjects' ability to assess the accuracy of their responses on the cued recall test, as indicated by the absolute correspondence between proportion correct and confidence estimate. The overall mean proportions correct for the delayed-feedback ($M = .56$) and immediate-feedback ($M = .49$) conditions were almost identical to the mean confidence estimate (means of 55 and 46, respectively). In contrast, the overall mean proportion correct in the no-test ($M = .14$) and no-feedback ($M = .31$) conditions was lower than the mean confidence estimate (means of 23 and 39, respectively), indicating overconfidence. The relationship between performance and confidence did not differ as a function of any of the other variables.

Mean confidence estimates were also computed for the intrusions produced on the forced cued recall test. Numerically, there was little difference between the confidence estimates in the no-feedback ($M = 42$), immediate-feedback ($M = 38$), and delayed-feedback ($M = 38$) conditions. However, subjects assigned greater confidence to intrusions produced in these three conditions (overall $M = 39$) than to those in the no-test condition ($M = 26$) [$t(71) = 4.11$, $SEM = 2.98$, $d = .56$, $p_{rep} = .99$]. Thus, prior exposure to lure items on the multiple-choice test seemed to increase confidence in the intrusion responses.

### Conditional Analyses[4]

Conditional analyses were conducted to investigate the relationship between performance on the initial multiple-choice test and on the final cued recall test. Of interest was how response outcome on the multiple-choice test (correct–incorrect) influenced the production of correct responses on the final cued recall test as a function of the amount-of-prior-study and testing conditions. One question was whether the overall pattern of results obtained in the main analyses (e.g., the superiority of delayed feedback) would hold for both items that were initially correct and those that were incorrect. For the purposes of these conditional analyses, the data were again collapsed across the number of multiple-choice alternatives.

Table 4 displays the proportion of correct responses on the cued recall test for items that were correctly and incorrectly answered on the initial multiple-choice test as a function of prior-study and feedback conditions. For items that were answered correctly on the multiple-choice test, delayed feedback led to a higher proportion of correct responses on the cued recall test than did immediate feedback, which in turn produced a higher proportion than did no feedback. For the most part, increases in the amount of prior study had little effect on the maintenance of correct responses from multiple choice to final cued recall. The only exception was the no-study–no-feedback condition, which produced a much lower proportion of correct responses than did the other prior-study conditions that did not receive feedback. For items that were initially

**Table 4**
**Proportion of Correct Responses on the Final Cued Recall Test for Items That Were Correctly and Incorrectly Answered on the Initial Multiple-Choice (MC) Test As a Function of Amount-of-Prior-Study Condition and Feedback Condition**

| Study Condition | Feedback Condition: No Feedback | Immediate Feedback | Delayed Feedback | *M* |
|---|---|---|---|---|
| | | MC–Correct | | |
| No Study | .35 | .53 | .61 | .50 |
| Study | .56 | .52 | .64 | .57 |
| Restudy | .52 | .57 | .62 | .57 |
| Mean | .50 | .55 | .63 | .56 |
| | | MC–Incorrect | | |
| No Study | .08 | .35 | .53 | .31 |
| Study | .07 | .27 | .41 | .24 |
| Restudy | .12 | .30 | .35 | .25 |
| Mean | .08 | .31 | .46 | .28 |

Note—All data are for performance under forced report instructions.

answered incorrectly, a similar overall pattern emerged among the testing conditions: Delayed feedback produced the highest proportion of correct responses, followed by immediate feedback and no feedback. The amount of prior study did not have clear overall effects on performance. However, the magnitude of the differences between the testing conditions appears to be attenuated as the amount of prior study increases.

Thus, the superiority of delayed feedback in comparison with immediate feedback that emerged in the main analyses held for both items that were initially correct and those that were incorrect on the multiple-choice test. The benefit of providing feedback (either delayed or immediate) as opposed to not providing feedback also held for both sets of items. The impact of increasing the amount of prior study was less clear, but this is probably because prior study increases correct performance and decreases errors on the initial test. Of course, conditional analyses are always subject to item-selection artifacts, and the results presented previously should be interpreted with this caution in mind. Still, delayed feedback on the multiple-choice test provided consistently better performance on the final cued recall test whether or not the multiple-choice item was correctly answered.

## DISCUSSION

The present experiment investigated the predictions that feedback on a multiple-choice test would enhance the testing effect for items answered correctly and reduce or eliminate negative effects on items answered incorrectly, as assessed on a cued recall test given a week later. Our findings confirmed these hypotheses. We replicated several previous findings within a single experiment and found novel interactions between two of the three variables being investigated. First, a testing effect emerged on the final cued recall test: Students performed better on items that were tested on the prior multiple-choice test than on items not initially tested, regardless of the amount of prior study or feedback. Second, when items were tested and feedback was not given, the pattern of performance on the cued recall test was largely determined by performance on the initial multiple-choice test. Greater amounts of prior study led to a higher proportion of correct responses in cued recall, but the number of multiple-choice alternatives did not have a significant effect on correct responses. Less prior study and a greater number of initial multiple-choice alternatives resulted in a higher proportion of intrusions, as in prior research (Butler et al., 2006; Roediger & Marsh, 2005). Third, the initial predictions regarding the effect of feedback on performance were substantiated: Feedback on the multiple-choice test increased the proportion of correct responses on the final cued recall test, whereas the proportion of intrusions was sharply reduced. Delayed feedback led to a higher proportion of correct responses than did immediate feedback, but both feedback schedules were equally effective at reducing the amount of misinformation acquired. Finally, when given the option of free report, subjects succeeded in reducing the proportion of intrusions reported; however, they also eliminated many correct responses. These results are discussed further in the subsequent sections, focusing first on the learning and retention of correct responses and then on the acquisition of misinformation. After placing the results in the context of other studies, we will conclude by discussing the implications of this research for educational practice.

### The Learning and Retention of Correct Responses

We first consider performance in the no-feedback condition, which was conceptually the most similar to previous studies (Butler et al., 2006; Roediger & Marsh, 2005). Just as in the initial multiple-choice test, the proportion of correct responses on the cued recall test was influenced by both the amount of prior study and the number of multiple-choice alternatives. A greater amount of prior study led to a higher proportion of correct responses, replicating Roediger and Marsh (2005), who included a study–no study manipulation. The restudy condition in our experiment extends their finding by showing that increasing the amount of prior study (by selective restudying of facts) can enhance recall. Increasing the number of multiple-choice alternatives led to a lower proportion of correct responses on the cued recall test, but not significantly so. Previous studies have found significant effects in which a greater number of alternatives led to a lower proportion of correct responses on a subsequent recall test (e.g., Butler et al., 2006; Roediger & Marsh, 2005; but see Whitten & Leonard, 1980). However, these effects tend to be relatively small in size, presumably because as more lures are included, the plausibility of each additional lure decreases. Nevertheless, the numerical trend was in the predicted direction in the present study, suggesting that greater power may have been needed to detect this effect.

When compared with the no-test condition, the no-feedback condition also shows the mnemonic benefit of taking a prior multiple-choice test. Retrieval of the correct response on the multiple-choice test helped students to learn and retain that response, which is no surprise because the testing effect is generally quite robust (see Roediger

& Karpicke, 2006a). However, a more stringent way of assessing the benefits of testing is to take into account any negative effect that occurs as a result of multiple-choice testing. When both correct responses and intrusions are considered, there is usually still a net benefit of testing (see, e.g., Roediger & Marsh, 2005). Thus, it is surprising to find that in the no-study–no-feedback condition of the present experiment, the proportion of intrusions produced ($M = .32$) was substantially greater than the proportion of correct responses ($M = .18$). Although the no-study–no-test condition produced fewer correct responses ($M = .10$), it also produced fewer (spontaneous) intrusions ($M = .17$). In other words, if students had not studied the material, they would have been worse off if they were tested than if not tested. Although a net benefit of prior testing was obtained in the study and restudy conditions, this negative effect of testing in this one condition is particularly important, because it indicates that there is a point at which multiple-choice testing ceases to be beneficial to students.

The testing effect observed in comparing the no-test and no-feedback conditions was enhanced by the provision of feedback. Both immediate and delayed feedback led to large gains in the proportion of correct responses on the cued recall test in comparison with the no-feedback condition. The added benefit of feedback is likely due to the correction of errors (Butterfield & Metcalfe, 2001; Pashler et al., 2005) and the maintenance of correct responses that otherwise might have been forgotten or switched to an attractive alternative (Butler et al., in press). Greater amounts of prior study led to a small increase in the proportion of correct responses in the immediate-feedback condition, but had no effect on the delayed-feedback condition (although the interaction was not statistically significant; $p = .11$). Assuming the differential effect exists, a possible explanation derives from the *interference-perseveration theory* (Kulhavy, 1977; Kulhavy & Anderson, 1972; Kulhavy & Stock, 1989). According to this theory, immediate feedback produces response competition when the correct response is presented immediately after an incorrect response is made. With a greater amount of prior study, fewer incorrect responses are made; therefore, the potential for response competition to occur should decrease. In contrast, a delay in the presentation of feedback may allow incorrect responses to dissipate, making the correct response easier to learn and negating the impact of any differences in the number of incorrect responses made.

In accordance with this theory, the timing of feedback also had a large influence on the learning and retention of correct responses. Delayed feedback led to a higher proportion of correct responses (overall $M = .56$) than did immediate feedback (overall $M = .45$). The superiority of delayed feedback in the present results can best be explained by invoking two different—but compatible—theories. First, as was just noted, the interference-perseveration theory offers an explanation for why delayed feedback might also benefit initially incorrect responses (Kulhavy, 1977; Kulhavy & Anderson, 1972; Kulhavy & Stock, 1989). As explained above, this theory revolves around the idea that immediate feedback produces competition between the incorrect response and the presented correct response. Accordingly, delayed feedback is more effective because it allows incorrect responses to dissipate, making the correct response easier to learn. A second theory posits that delayed feedback leads to better subsequent recall because it provides an additional spaced presentation of the material. The superiority of spaced (or distributed) study in comparison with massed study for enhancing the retention of verbal material has been well established (see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a review). If a correct response to a test question is considered a study trial, then immediate feedback represents a massed study trial, and delayed feedback represents a spaced study trial. Thus, delayed feedback should be superior to immediate feedback for initially correct responses, but equally effective for initially incorrect responses for which both feedback timings would represent a spaced study trial (see, e.g., McConnell, Hunt, & Smith, 2006). However, delayed feedback led to a higher proportion of correct responses in the present study regardless of whether the initial response was correct or incorrect. Importantly, these two theories are not mutually exclusive, because the spaced presentation and interference-perseveration theories focus on the effect of feedback timing after correct and incorrect responses, respectively. Therefore, a combination of the two theories provides a comprehensive account of the present results.

As a final consideration, feedback also helped students to better gauge the accuracy of their responses on the final cued recall test. Students' ability to differentiate between correct and incorrect responses was explored through the absolute correspondence between the confidence estimates and the proportion of correct responses. This analysis revealed that subjects were almost perfectly calibrated for items in the delayed-feedback condition and only slightly overconfident in the immediate-feedback condition. However, students were highly overconfident in the no-test and no-feedback conditions. Thus, testing with feedback also helps students to judge better what they know and what they do not know.

### The Acquisition of Misinformation

Focusing first on the no-feedback condition, the proportion of intrusions produced on the forced report phase of the cued recall test was heavily influenced by both the prior study and number of alternatives variables. Decreasing the amount of prior study and/or increasing the number of alternatives led to a higher proportion of intrusions, replicating previous research (Butler et al., 2006; Roediger & Marsh, 2005). The overall pattern of intrusions closely resembled the incorrect response data from the multiple-choice test, suggesting that performance on the multiple-choice test mediated the influence of these two variables on the acquisition of misinformation. This conclusion is bolstered by the fact that 75% of the intrusions produced in the no-feedback condition were lures that had been (incorrectly) selected on the initial multiple-choice test (the other 25% were initially correct responses that were subsequently switched to lures on the cued recall test; see Butler et al., in press). Interestingly, the amount of misinformation acquired varied widely among the different conditions: Over

a third of the responses in the no-study–six-alternative condition were intrusions ($M = .38$), whereas only a rather small proportion in the restudy–two-alternative condition were intrustions ($M = .14$). Note that the latter proportion was no greater than the overall mean in the no-test ($M = .16$) and feedback conditions (immediate feedback, $M = .15$; delayed feedback, $M = .14$). Theoretically, intrusions on the cued recall test likely result from lure responses blocking previously learned correct responses (or causing them to be unlearned), similar to the retroactive interference created in the misinformation paradigm (Loftus, Miller, & Burns, 1978) or the in classic A–B, A–D interference paradigm (McGeoch, 1932). Alternatively, the correct response may never have been learned, and people may just have guessed and learned the wrong response as a result. However, research suggests that errors resulting from faulty reasoning are much more likely to persist than are guesses (Huelser & Marsh, 2006).

When feedback was provided after the multiple-choice test, a very different pattern of results emerged. First, the overall amount of misinformation acquired was sharply reduced. Second, the effects observed in the no-feedback condition were neutralized: Neither the amount of prior study nor the number of multiple-choice alternatives had an influence on the proportion of intrusions produced. Armed with knowledge about whether their response was correct or incorrect (as well as the correct answer), subjects were able to correct many of their errors and refrain from producing the lures on the cued recall test. This result fits nicely with previous investigations that have demonstrated the error-correcting function of feedback (e.g., Butterfield & Metcalfe, 2001; Pashler et al., 2005). Furthermore, the timing of the feedback did not seem to matter: Immediate and delayed feedback were equally effective at reducing the amount of misinformation acquired. It might be argued that even with feedback, multiple-choice tests were harmful because a small but sizeable number of intrusions were produced (overall $M = .15$) even with feedback on the multiple-choice test. However, we believe that this implication is erroneous; if the proportion of intrusions spontaneously produced in the no-test condition ($M = .16$) is used as a baseline, then it is clear that taking a multiple-choice test with feedback is no more harmful than not taking a multiple-choice test at all.

Another goal of the experiment was to investigate the extent to which students believe that the misinformation acquired from the multiple-choice test is true. Students' confidence estimates for intrusions indicated roughly the same level of confidence regardless of whether or not they had received feedback. This result indicates that feedback works in an all-or-none manner: If students do not successfully correct the error, then feedback does not diminish the potency of the misinformation. The only obvious difference in confidence estimates for intrusions was between the no-test condition ($M = 26$) and the other three testing conditions (overall $M = 39$), indicating that prior exposure to lures led to greater confidence in intrusions in comparison with intrusions that were spontaneously produced (presumably due to the familiarity of the lures). When students were allowed to revisit their forced report responses and omit any responses they believed to be incorrect, they succeeded in reducing the proportion of intrusions, more than halving the number of lure items reported. However, this reduction in the proportion of intrusions was roughly equivalent across all the conditions, and the same overall pattern of effects that was obtained under forced report remained. That is, the no-feedback condition still produced the highest proportion of intrusions in comparison with the other conditions, and this proportion increased as the amount of prior study decreased. Remarkably, even under free report, subjects in the no-study–no-feedback condition decided to keep a large proportion of intrusions ($M = .19$). Overall, these results indicate that students strongly believed in the veracity of misinformation acquired during the multiple-choice test.

**Implications for Educational Practice**

The present experiment demonstrates that students acquire both correct and incorrect information from multiple-choice tests. Taking a multiple-choice test leads to a substantial benefit in retention of correct responses, but the exposure to misinformation in the form of multiple-choice lure items can lead to the intrusion of these lures on a subsequent test, especially when the lure is (incorrectly) endorsed on the initial multiple-choice test. The magnitude of these positive and negative effects is greater with little prior study and with increasing numbers of lures on the multiple-choice test. If the material has not been sufficiently studied prior to taking a multiple-choice test, or a test is made more difficult by increasing the number of alternatives, students acquire a greater amount of misinformation. Although these two factors have been emphasized in the present study, any factor that negatively affects performance on a multiple-choice test (e.g., test anxiety, time restrictions, increasing the attractiveness of lures, etc.) will probably have the same effect.

A pragmatic solution to the possible negative effects of multiple-choice tests is to ensure that students always receive feedback after testing. Feedback enhances the positive effects of taking a test and helps students correct their errors, thereby reducing the acquisition of misinformation. The latter outcome is especially important when the same questions and alternatives from a first test are reused on a later test, because the production of misinformation often increases the chance that it will be produced again on a later test (Roediger, Jacoby, & McDermott, 1996; Roediger, Wheeler, & Rajaram, 1993). One positive aspect of the present results is that feedback need not be given immediately; a delay in the presentation of feedback seems to be beneficial to learning. Of course, in our conditions, what we are calling delayed feedback is what many instructors who cannot use computerized testing would see as immediate feedback; students in our delayed feedback condition were actually given feedback soon after taking the test (but not immediately after answering each item). Further research will be needed to determine if feedback may cease to be effective if it is delayed for too long. For example, in many classroom settings, feedback on a test is provided a week or two after the test is given, in order to permit time for grading the test. Would feedback under

these conditions lose its effectiveness or be even more effective?

One final consideration involves the type of to-be-learned information used in the present study. Although our test questions focused on relatively basic factual information, the results of the present study probably extend to more complex conceptual information as well. The critical mechanism for promoting the retention of information is the successful retrieval of that information. Thus, if a test leads students to successfully retrieve conceptual information, then the retention of that conceptual information will be enhanced. Indeed, there is some evidence to suggest that testing on conceptual information leads to even bigger testing effects (Wildman & McDaniel, 2007). Marsh, Roediger, Bjork, and Bjork (2007) reported that both positive and negative effects of multiple-choice testing were apparent in complex materials when no feedback was given. We expect that effects of feedback would diminish errors in these complex materials—as with our current factual materials—but testing this conjecture must await future research.

In summary, the present research highlights the importance for educators to provide (briefly) delayed feedback following multiple-choice tests in order to maintain correct answers and to correct erroneous answers. Even students who have not studied the material thoroughly will benefit, at least for the information that is tested. Our results provide further evidence for the importance of judicious testing in order to enhance educational performance.

## AUTHOR NOTE

We thank Patrick Flanagan for his help in collecting data. This research was supported with a Collaborative Activity Award from the James S. McDonnell Foundation and Grant R305H030339 from the Institute of Education Sciences. Correspondence may be addressed to A. C. Butler, Department of Psychology, Campus Box 1125, Washington University, One Brookings Drive, St. Louis, MO 63139-4899 (e-mail: butler@wustl.edu).

## REFERENCES

AGARWAL, P. K., KARPICKE, J. D., KANG, S. H. K., ROEDIGER, H. L., III & MCDERMOTT, K. B. (in press). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*.

AMERICAN PSYCHOLOGICAL ASSOCIATION (2002). *Ethical priniciples of psychologists and code of conduct*. Retrieved June 1, 2003, from www.apa.org/ethics/code2002.html.

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.

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.

BANGERT-DROWNS, R. L., KULIK, J. A., & KULIK, C. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, **85**, 89-99.

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

BRACKBILL, Y., BRAVOS, A., & STARR, R. H. (1962). Delay-improved retention of a difficult task. *Journal of Comparative & Physiological Psychology*, **55**, 947-952.

BROWN, A. S. (1988). Experiencing misspellings and spelling performance: Why wrong isn't right. *Journal of Educational Psychology*, **80**, 488-494.

BROWN, A. S., SCHILLING, H. E. H., & HOCKENSMITH, M. L. (1999). The negative suggestion effect: Pondering incorrect alternative may be hazardous to your knowledge. *Journal of Educational Psychology*, **91**, 756-764.

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

BUTLER, A. C., KARPICKE, J. D., & ROEDIGER, H. L., III (in press). Correcting a metacognitive error: Feedback enhances retention of low confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, & Cognition*.

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.

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

CARRIER, M., & PASHLER, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, **20**, 633-642.

CEPEDA, N. J., PASHLER, H., VUL, E., WIXTED, J. T., & ROHRER, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, **132**, 354-380.

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

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

FOOS, P. W., & FISHER, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology*, **80**, 179-183.

FREDERIKSEN, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, **39**, 193-202.

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.

GILMAN, D. A. (1969). Comparison of several feedback methods for correcting errors by computer-assisted instruction. *Journal of Educational Psychology*, **60**, 503-508.

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

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

JACOBY, L. L., & HOLLINGSHEAD, A. (1990). Reading student essays may be hazardous to your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology*, **44**, 345-358.

JONES, H. E. (1923–1924). The effects of examination on the performance of learning. *Archives of Psychology*, **10**, 1-70.

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.

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

LOFTUS, E. F., MILLER, D. G., & BURNS, H. J. (1978). Semantic integration of verbal material into a visual memory. *Journal of Experimental Psychology: Human Learning & Memory*, **4**, 19-31.

MARSH, E. J., MEADE, M. L., & ROEDIGER, H. L., III (2003). Learning facts from fiction. *Journal of Memory & Language*, **49**, 519-536.

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

MCCONNELL, M. D., HUNT, R. R., & SMITH, R. E. (2006, May). *Differential effects of immediate and delayed feedback on test performance*. Paper presented at the annual meeting of the Midwestern Psychological Society, Chicago, IL.

MCDANIEL, M. A., ANDERSON, J. L., DERBISH, M. H., & MORRISETTE, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, **19**, 494-513.

MCDANIEL, M. A., & FISHER, R. P. (1991). Test and test feedback as learning sources. *Contemporary Educational Psychology*, **16**, 192-201.

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

MCGEOCH, J. A. (1932). Forgetting and the law of disuse. *Psychological Review*, **39**, 352-370.

MCKEACHIE, W. J. (1999). *Teaching tips: Strategies, research, and theory for college and university teachers* (10th ed.). Boston: Houghton Mifflin.

PAIGE, D. D. (1966). Learning while testing. *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, & Cognition*, **31**, 3-8.

ROEDIGER, H. L., III, JACOBY, J. D., & MCDERMOTT, K. B. (1996). Misinformation effects in recall: Creating false memories through repeated retrieval. *Journal of Memory & Language*, **35**, 300-318.

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, & 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., III, 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.

SCHNEIDER, W., ESCHMAN, A., & ZUCCOLOTTO, A. (2002). E-Prime reference guide. Pittsburgh: Psychology Software Tools, Inc.

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

SCHROTH, M. L. (1977). Effects of informative feedback and active training upon concept attainment. *Psychological Reports*, **40**, 647-653.

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

STURGES, P. T. (1969). Verbal retention as a function of the informativeness and delay of informative feedback. *Journal of Educational Psychology*, **60**, 11-14.

SULLIVAN, H. J., SCHUTZ, R. E., & BAKER, R. L. (1971). Effects of reinforcement contingencies. *American Educational Research Journal*, **8**, 135-141.

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

TAIT, K., HARTLEY, J. R., & ANDERSON, R. C. (1973). Feedback procedures in computer-assisted arithmetic instruction. *British Journal of Educational Psychology*, **43**, 161-171.

TOPPINO, T. C., & LUIPERSBECK, S. M. (1993). Generality of the negative suggestion effect in objective tests. *Journal of Educational Psychology*, **86**, 357-362.

TULVING, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning & Verbal Behavior*, **6**, 175-184.

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.

WHITTEN, W. B., & LEONARD, J. M. (1980). Learning from tests: Facilitation of delayed recall by initial recognition alternatives. *Journal of Experimental Psychology: Human Learning & Memory*, **6**, 127-134.

WILDMAN, K., & MCDANIEL, M. A. (2007, August). *Test-enhanced learning for facts and concepts*. Poster session presented at the annual meeting of the American Psychological Association, San Francisco.

## NOTES

1. Performance in a given response category is dependent on performance in the other two response categories (e.g., the proportion of incorrect other responses is necessarily determined by the proportion correct and intrusions). The analysis of incorrect other responses is not reported because the measure was not of primary interest and would be redundant with other measures, but the data can be found in the Appendix. We include the Appendix to provide a complete summary of the forced recall results.

2. As in the forced report phase, there were three possible outcomes for each response: correct, intrusion, or incorrect other. However, unlike the forced report phase, these various response outcomes are potentially independent of each other because a response could also be omitted. The analyses of incorrect other responses are not presented because subjects kept very few of these responses (overall $M = .10$) and there were no significant differences between any of the conditions. The proportions reported for the free report performance were computed by dividing the number of items kept by the total number of items in that condition of the experiment (i.e., not the total number of items that were kept).

3. The number of multiple-choice alternatives factor was initially included in all of the free report analyses. However, this factor did not produce any main effects and did not interact with any of the other factors. Thus, all the free report data were collapsed across the number of multiple-choice alternatives for the sake of brevity.

4. The conditional analyses were conducted on the aggregated data (i.e., across subjects) rather than the alternative method of computing conditionalized means for each individual subject. This method was used in order to avoid the problem of how to replace or estimate a mean for individual subjects 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"). Because we used this form of analysis, inferential statistics could not be computed on the data. They should be considered descriptive.

(*Continued on next page*)

## APPENDIX

**Performance on the Cued Recall Test Under Forced Report Instructions As a Function of Amount-of-Prior-Study Condition, the Number of Initial Multiple-Choice Alternatives, and Feedback Condition (Excluding the No-Test Condition)**

| | | Feedback Condition | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | No Feedback | | | Immediate Feedback | | | Delayed Feedback | | |
| Study Condition | DV | Two | Four | Six | Two | Four | Six | Two | Four | Six |
| No study | Correct | .21 | .17 | .16 | .36 | .43 | .48 | .52 | .57 | .60 |
| | Intrusions | .23 | .34 | .38 | .16 | .14 | .14 | .10 | .14 | .14 |
| | Incorrect other | .56 | .49 | .46 | .48 | .43 | .38 | .38 | .29 | .26 |
| Study | Correct | .40 | .29 | .31 | .41 | .43 | .45 | .47 | .56 | .58 |
| | Intrusions | .20 | .28 | .27 | .16 | .19 | .20 | .13 | .10 | .16 |
| | Incorrect other | .40 | .43 | .42 | .43 | .39 | .35 | .40 | .34 | .26 |
| Restudy | Correct | .43 | .42 | .38 | .51 | .50 | .49 | .57 | .56 | .56 |
| | Intrusions | .14 | .15 | .21 | .13 | .17 | .14 | .15 | .16 | .16 |
| | Incorrect other | .43 | .43 | .41 | .36 | .33 | .37 | .28 | .28 | .28 |
| Mean | Correct | .35 | .29 | .28 | .43 | .45 | .47 | .52 | .56 | .58 |
| | Intrusions | .19 | .26 | .29 | .15 | .17 | .16 | .13 | .13 | .15 |
| | Incorrect other | .46 | .45 | .43 | .42 | .38 | .37 | .35 | .30 | .27 |

Note—The first dependent variable (DV) is the proportion of targets correctly recalled, the second is the proportion of intrusions, and the third is the proportion of incorrect other responses.

(Manuscript received June 28, 2007;
revision accepted for publication October 29, 2007.)