Memory & Cognition
2010, 38 (4), 407-418
doi:10.3758/MC.38.4.407

Memorial consequences of multiple-choice testing on immediate and delayed tests

Lisa K. Fazio
Duke University, Durham, North Carolina

Pooja K. Agarwal
Washington University, St. Louis, Missouri

Elizabeth J. Marsh
Duke University, Durham, North Carolina

and

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

Multiple-choice testing has both positive and negative consequences for performance on later tests. Prior testing increases the number of questions answered correctly on a later test but also increases the likelihood that questions will be answered with lures from the previous multiple-choice test (Roediger & Marsh, 2005). Prior research has shown that the positive effects of testing persist over a delay, but no one has examined the durability of the negative effects of testing. To address this, subjects took multiple-choice and cued recall tests (on subsets of questions) both immediately and a week after studying. Although delay reduced both the positive and negative testing effects, both still occurred after 1 week, especially if the multiple-choice test had also been delayed. These results are consistent with the argument that recollection underlies both the positive and negative testing effects.

Multiple-choice exams are commonly used in classrooms, since they are easy to grade and their scoring is perceived as objective. Although much has been written about the assessment function of such tests, less research has focused on the consequences of this form of testing for long-term knowledge. This gap in the literature is troubling, because the available results suggest that tests can change knowledge, in addition to assessing it. The most well-known example is the *testing effect*, the finding that taking an initial test often increases performance on a later test (see Roediger & Karpicke, 2006a, for a review).

Whereas earlier work on testing tended to rely on simple word list stimuli, more recently the emphasis has shifted to studying the effects of testing in educationally relevant situations (Butler, Marsh, Goode, & Roediger, 2006; Marsh, Agarwal, & Roediger, 2009; Marsh, Roediger, Bjork, & Bjork, 2007; Roediger, Agarwal, Kang, & Marsh, 2010; Roediger & Marsh, 2005). In the typical experiment, subjects read nonfiction passages on a variety of topics and then take an initial multiple-choice test. A few minutes later, they take a final cued recall test that includes questions that were tested on the prior multiple-choice test, as well as new questions. Subjects are more likely to answer final cued recall questions correctly if they were tested on the prior multiple-choice test, thus showing the testing effect.

A second effect in this sort of experiment is more problematic: Multiple-choice testing can also have negative effects on students' knowledge. The reason is that multiple-choice tests expose students to incorrect answers (lures), in addition to correct responses. Just as Brown (1988) and Jacoby and Hollingshead (1990) showed that exposure to incorrect spellings of words increased later misspellings, one could predict that reading lures on a multiple-choice test would increase errors on later tests. Supporting this logic, Toppino and his colleagues showed that students rated previously read multiple-choice lures as truer than novel false facts (Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). Similarly, Roediger and Marsh (2005) found that multiple-choice testing increased the intrusion of multiple-choice lures as answers on a final general knowledge test, even though subjects were warned not to guess on that test. Consistent with an interference account, multiple-choice questions that paired the correct answer with a greater number of lures increased this negative effect of testing.

Prior work has established that multiple-choice tests can have both positive and negative consequences. But

L. K. Fazio, lkf@duke.edu

how persistent are these effects? Prior research has established that positive testing effects persist over at least a week's delay. For example, Spitzer (1939) had 3,605 sixth-graders in Iowa read a passage on bamboo. The children were tested on the passage according to different testing schedules. In one group, children were tested on the passage immediately after reading it and again 1 week later. Another group was tested on the passage for the first time 1 week after reading it. When both groups were tested 1 week after reading the passages, performance was much higher in the group that had been tested previously on the material than in the group being tested for the first time. In other words, the benefits of initial testing persisted over a delay of 1 week. Roediger and Karpicke (2006b) observed similar effects in college students. Their students read nonfiction passages; some of these were restudied during the initial session and others were tested. After 2 days or 1 week, recall of the passages was higher if they had been tested initially than if they had been restudied. To be clear, performance was always lower on delayed tests than on immediate tests, but there was less forgetting over time following testing than after an equivalent time spent restudying.

The question we address in the present research is whether negative testing effects persist over a delay, similar to what occurs with positive testing effects. Butler and Roediger (2008) found that negative testing effects can be nullified if feedback is provided after the multiple-choice test. However, this step is often not taken in the classroom, in order to protect items from the test bank. If negative testing effects do *not* persist for long after a multiple-choice test, this fact would remove concerns about the negative effects of testing. On the other hand, if the negative effects do persist over time, the implication for educators would be to include feedback with all tests.

The effects of delay are also of theoretical interest. Typically, manipulations of delay have different effects on memory errors, depending on the mechanism underlying the error. Consider the standard explanation for the effects of delay in the *false fame paradigm*. In a prototypical experiment, subjects study famous and nonfamous names, some of which were presented during an initial study session. Afterward, the subjects judge the fame of each of a series of names, including new famous names, new nonfamous names, and studied nonfamous names. On an immediate test, the subjects are *less* likely to call repeatedly studied nonfamous names "famous," because they are able to recollect the source of the names' familiarity: the earlier study phase. In contrast, if the fame judgments are delayed for a day, the subjects are *more* likely to call repeatedly studied nonfamous names "famous." After a day, the names are still familiar, but the subjects are less able to recollect the source of that familiarity (Jacoby, Kelley, Brown, & Jasechko, 1989).

An increased reliance on familiarity over time (as recollection drops) is used to explain the effects of delay in numerous paradigms.[1] For example, consider the finding that prior shallow processing of campus scenes increases subjects' belief that they have visited locations that they had never actually been to (Brown & Marsh, 2008). Exposure increases a scene's familiarity, but after a delay of 1 or 3 weeks, subjects misattribute that familiarity to prior personal experience with the place. The type of familiarity proposed to underlie these results is similar to the representations that support long-term priming over months and years (e.g., Cave, 1997; Mitchell, 2006). Thus, the level of false memories is likely to be consistent over time (or even increase) if they result from a misattribution of this type of familiarity. Returning to the issue of multiple-choice tests, a previously selected multiple-choice lure may easily come to mind at test, and this retrieval ease may be misinterpreted as confidence in the answer (Kelley & Lindsay, 1993), rather than as its presence on the earlier test. Thus, delaying the final test may have no effect on the negative testing effect or may even increase it. To be clear, we are not suggesting that familiarity does not decrease over time. Rather, as subjects become more reliant on familiarity, they may produce lures that they would have rejected on an immediate test (because they remembered that the answer was presented on the multiple-choice test).

In contrast, some memory errors actually decrease over time. For example, consider what happens when people learn falsehoods from fictional stories. In this paradigm, subjects read short stories that contain statements about the world, some of which are false. Subjects intrude these story errors on later general knowledge tests even when they are warned against guessing. Suggestibility is robust on an immediate test but is reduced on a delayed test (Barber, Rajaram, & Marsh, 2008; Marsh, Meade, & Roediger, 2003). In this case, subjects learn specific falsehoods that need to be recollected, and thus, delay reduces the effect. Returning to the issue of multiple-choice tests, it is possible that the negative testing effect depends on recollection of the multiple-choice lures. If so, delay should reduce the negative testing effect.

The prior literature allows for both possibilities. On the one hand, the effects of testing have been linked to enhanced recollection (Chan & McDermott, 2007; Karpicke, McCabe, & Roediger, 2006). Prior testing increases the number of *remember* responses on a later recognition test, and process dissociation measures show that the effects of testing are primarily recollection driven, rather than familiarity driven. From these studies, we would predict that both positive and negative testing effects would depend on recollection and, thus, should be similarly affected by delay. On the other hand, Brainerd and Reyna (1996) have shown that delay increases the likelihood that children will select a lure from a prior recognition test on a second test, suggesting a role for familiarity in this memory error. From this study, we would predict that familiarity underlies negative testing effects and, thus, that the level of lure intrusions on a final test should remain constant or even increase over time.

In the experiment presented here, we asked a number of questions about how delay affects the memorial consequences of testing. All the subjects visited the laboratory twice, with 1 week separating the two sessions. Of interest was the subjects' ability to answer questions about facts from 36 nonfiction passages on initial and delayed tests. The different delays were all manipulated within subjects.

All of the subjects took all of the tests, and across subjects, the assignment of passages to testing schedules was counterbalanced, as shown in Table 1.

During the first session, the subjects read one half of the nonfiction passages; reading status was manipulated to ensure a wide range of performance. The goal was for some questions to be difficult because the passages had not been read (and thus, potentially more likely to yield negative testing effects) and for some to be easier following passage reading (and thus, more likely to be remembered correctly after a delay of 1 week). After the reading phase, all the subjects took an initial multiple-choice test on two thirds of the passages (see Table 1). Each multiple-choice question paired the correct answer with one, three, or five lures; in other words, the subjects answered two-, four-, and six-alternative forced choice questions. On immediate tests, testing with additional multiple-choice lures increases the negative testing effect (Roediger & Marsh, 2005); of interest here was whether that effect would persist over a delay.

After completion of the initial multiple-choice test, all the subjects completed an initial cued recall test. Critically, this test included questions on half the facts tested on the initial multiple-choice test (see Table 1). One week later, the subjects returned and took a second multiple-choice test (on the remaining one third of the passages that had not yet been tested on a multiple-choice test) and a final cued recall test on all items. The subjects were instructed to answer all cued recall questions, just as students attempt to answer all exam questions, even if unsure. Because forced responding increases guessing, the subjects also rated their confidence in each answer so that we could ascertain whether guessing was responsible for any negative testing effects that might be observed.

The design yielded three testing schedules, all of which have real-world parallels in educational situations. Schedule A (immediate multiple-choice and cued recall tests) mimics students' self-quizzing immediately before an exam. Schedule B (immediate multiple-choice and delayed cued recall tests) is similar to cases in which a teacher gives a quiz 1 week before a larger, more comprehensive test. Finally, in Schedule C (delayed multiple-choice and cued recall tests), students have read the material earlier and are then quizzing themselves just before the exam. It should be noted that Schedules A and C model different situations; although both involve multiple-choice testing immediately before a cued recall test, in Schedule A this testing occurs immediately after reading the passages, whereas in Schedule C the testing is delayed a week after passage reading. In some ways, Schedule C is the most likely scenario in the real world; students learn information but then delay self-testing and other study behaviors until immediately before the exam.

This design allowed us to answer three important questions about the persistence of the positive and negative consequences of multiple-choice testing. First, what is the effect of delaying the cued recall test until a week after the initial multiple-choice test? To answer this, we compared performance on the initial cued recall test in Schedule A with performance on the final cued recall test in Schedule B. The second question involved any effects of delaying the multiple-choice test by 1 week. To answer this question, we compared performance on the final cued recall test in Schedule B (following immediate multiple-choice testing) with that observed in Schedule C (following delayed multiple-choice testing). Performance should be higher on the immediate multiple-choice test, perhaps magnifying the benefits and minimizing the costs of testing. In contrast, more errors might be selected on a delayed multiple-choice test, possibly increasing the costs of testing. The final question involved whether the effects of testing persist from the first cued recall test to the final cued recall test. Would the costs and benefits of testing observed on an initial exam persist a week later? Again, the focus was on performance on the final cued recall test, but the key comparison was between the initial and final cued recall tests in Schedule A.

METHOD

Subjects

Seventy-two Washington University undergraduates participated in the experiment, either for partial fulfillment of a course requirement or for monetary compensation.

Design

The experiment had a 2 (passage status: read or not read) × 4 (number of alternatives on the multiple-choice test: zero [not tested], two, four, or six) × 3 (testing schedule: A, B, or C, as shown in Table 1) design. All the factors were manipulated within subjects and were counterbalanced across subjects.

Materials

We used the same nonfiction passages as did Roediger and Marsh (2005); these were selected from reading comprehension sections of

Table 1
Within-Subjects Design of the Experiment

	Session 1			Session 2	
Schedule	Tested in MC 1?	Tested in CR 1?	Delay (1 Week)	Tested in MC 2?	Tested in Final CR?
A (12 passages: 6 read, 6 not read)	yes	yes		no	yes
B (12 passages: 6 read, 6 not read)	yes	no		no	yes
C (12 passages: 6 read, 6 not read)	no	no		yes	yes

Note—One half of the passages were read, and one half were not; both types of passages were rotated through the three testing schedules (A, B, C) shown. Assignment of passages to reading condition (read vs. not read), passages to testing schedule (A, B, C), and facts to multiple-choice format (not tested vs. two, four, or six alternatives) was counterbalanced across subjects. MC, multiple-choice; CR, cued recall.

TOEFEL, SAT, and GRE practice test books. The passages spanned a variety of topics, including famous people (e.g., Louis Armstrong), science (e.g., the sun), history (e.g., the founding of New York City), places (e.g., Mt. Rainier), and animals (e.g., sea otters). Roediger and Marsh created four questions for each passage, each of which was tested in all four formats necessary for the design (two-, four-, and six-alternative multiple choice, plus cued recall). The multiple-choice questions were created by generating five plausible lures for each question, and the six options (the lures plus the correct answer) were randomly ordered. Two lures were randomly removed to create each four-alternative question; two more were randomly removed from each to create the two-alternative questions. Across subjects, the four questions corresponding to each of the passages were rotated through the four multiple-choice conditions (zero [not tested], two, four, or six alternatives).

The 36 passages were divided into two sets to allow counterbalancing of reading status; each reading set was further subdivided into three groupings to allow counterbalancing of testing schedules. Thus, there were six groups of 6 passages; texts on similar subjects (e.g., the ozone layer and the sun) were placed in different groups. Half the subjects read the passages in Set 1; the other half read the passages in Set 2. Therefore, each subject read only half the passages but was tested on all 36. Across subjects, both read and nonread passages were rotated through the three testing schedules (A, B and C, as depicted in Table 1). All the items were included on the final cued recall test; we manipulated which passages were tested (and in what format) prior to that final test. For one set of passages, Schedule A, the subjects took the multiple-choice test and a cued recall test in Session 1 (as well as the final cued recall test on all items in Session 2). For a second set of passages, Schedule B, the subjects took the multiple-choice test in Session 1 but did not take a cued recall test until Session 2. For the third set of passages, Schedule C, the items were not tested in Session 1. Rather, the multiple-choice test was administered in Session 2, prior to the final cued recall test.

The first multiple-choice test contained 96 questions: 24 fillers and 72 critical questions (half corresponding to read passages). The fillers were questions from the Nelson and Narens (1980) norms and were used to provide separation between questions from the same passage (fillers were used for this purpose on the other multiple-choice and cued recall tests, too). There were 12 different versions of this test, so that, across subjects, all the items appeared in all four multiple-choice formats (not tested or two, four, or six alternatives) and all the passages were sometimes tested in this immediate multiple-choice condition.

The first cued recall test contained 72 questions: 24 fillers and 48 critical items (half from read passages and half from nonread passages). Each question was followed by a space for writing the answer and a box for recording confidence. Confidence was rated on a 4-point scale ranging from *not sure* to *very sure*. There were three versions of this test, so that, across subjects, all the passages were sometimes tested on this test.

The second multiple-choice test contained 48 questions: 12 fillers and 36 critical questions (18 from studied passages). As with the first multiple-choice test, 12 versions were needed for counterbalancing purposes.

All the subjects took the same final cued recall test. This test contained 144 critical questions and 72 fillers, for a total of 216 questions. Each question was followed by a space for writing the answer and a box for recording confidence. Confidence was rated on the same 4-point scale as that used on the first cued recall test. All the tests were in paper-and-pencil format.

Procedure

The experiment consisted of two sessions, separated by 1 week. In the first session, the subjects read 18 of the 36 passages. The amount of time devoted to each passage was determined in pretesting; the amount of time allotted to each passage varied because the passages differed in length. On average, the subjects were given up to 90 sec. to read each passage. The goal was for all the subjects to finish reading each passage once. The subjects were given a sheet on which they indicated when they had completed reading the passage; the experimenter monitored the subjects for completion and moved the subjects to the next passage when all of them had finished reading.

Immediately after the passages had been read, the first multiple-choice test was administered. The experimenter read the instructions aloud to the subjects, telling them that they were going to take a multiple-choice test, with no mention of the prior reading phase. They were told,

> You must answer each and every question. You will not know the answers to all of the questions. That's okay. If you have to, just guess. Sometimes a question will have two possible answers, sometimes four, and sometimes six. For each question, read the question carefully, read all the possible answers, and then circle the best answer. Again, you should answer all of the questions even if you have to guess. We would like you to answer the questions in the order in which they appear. Do not go back and change your answers. Rather, read each question and its answers once, and simply select the best possible answer and move on to the next question.

The subjects were told that they would receive up to 14 min for completion of the test and that they would be given verbal warnings about how time was passing. Pretesting determined that this amount of time would be more than enough for the subjects to finish the test. Those who finished early were instructed to turn over their tests and wait quietly for the next set of instructions. All the subjects then worked on a spatial filler task for 5 min.

After the filler task, the subjects had up to 12 min to complete the first cued recall test. The experimenter read the following instructions aloud to subjects:

> You will now take a second general knowledge test. This time, the questions are open-ended. So, you will read each question and write down your answer. Again, we would like you to answer all of the questions even though some of them are very difficult. Please write an answer for each and every one even if you have to guess. Again, answer the questions in the order in which they appear, and do not go back and change your answers. For each answer, please rate how sure you are that you are correct, using the following scale: 1 = *very sure*, 2 = *sure*, 3 = *somewhat sure*, and 4 = *not sure*. Please write the appropriate number in the box labeled confidence rating, next to the blank on which you'll write your answer.

The subjects were informed that the test had 72 questions and that they would be given 12 min to complete the test; pretesting had established that this was more than enough time for subjects to complete the test. The subjects followed the instructions, answering an average of 98% of the cued recall questions.

One week later, the subjects returned to the lab for Session 2. The session began with the second multiple-choice test, which was prefaced with the same instructions as the first test. The subjects were given up to 7 min for completing this test (again, this time was determined through pretesting). Following the multiple-choice test, all the subjects worked on a spatial filler task for 5 min. After the filler task, all the subjects took the final cued recall test and rated their confidence in each answer, using the same 4-point scale as that used on the earlier cued recall test. The subjects were given up to 35 min to complete the final test, with the same instructions as those used on the first cued recall test. No reference was made to the reading phase or to the earlier tests. No subjects had difficulty in completing any of the tests in the time allotted. As with the first cued recall test, the subjects followed the instructions, answering an average of 98% of the cued recall questions.

Table 2
Proportions Correct on the Multiple-Choice (MC) Tests, As a Function of Timing of the MC Test, Number of MC Alternatives, and Reading Status of the Passages

MC Test Timing	Reading Status	Number of MC Alternatives: Two	Four	Six	*M*
Immediate	Read	.86	.78	.72	.79
	Not read	.68	.48	.41	.52
Delayed	Read	.76	.60	.52	.63
	Not read	.68	.49	.38	.52
M		.75	.59	.51	

RESULTS

All the results were significant at the .05 level of confidence, unless otherwise noted.

Performance on the Multiple-Choice Tests

The data from the multiple-choice tests are shown in Table 2. The subjects correctly answered more multiple-choice questions when they had read the passages containing the tested facts ($M = .71$) than when they had not read the relevant passages ($M = .52$) [$F(1{,}71) = 242.38$, $MS_e = .03$, $\eta_p^2 = .77$]. In addition, as the number of multiple-choice alternatives increased, the subjects were less likely to answer the multiple-choice question correctly [$F(2{,}142) = 186.65$, $MS_e = .02$, $\eta_p^2 = .72$]. This effect was larger when the subjects had *not* read the passages containing the facts. When the subjects had not read the passages, performance decreased from .68 when they chose between two alternatives to .48 with four alternatives to only .40 with six alternatives. In other words, when the passages had not been read, performance dropped 28% when the alternatives were increased from two to six, as compared with the smaller drop of 19% when the subjects *had* read the passages. This interaction between reading status and number of alternatives was significant [$F(2{,}142) = 6.02$, $MS_e = .03$, $\eta_p^2 = .08$]. To be clear, the number of alternatives affected performance for both read [$F(2{,}142) = 65.01$, $MS_e = .01$, $\eta_p^2 = .48$] and nonread [$F(2{,}142) = 103.90$, $MS_e = .01$, $\eta_p^2 = .59$] passages, but this effect was larger when the subjects had not read the passages.

As was expected, the students' ability to correctly answer multiple-choice questions depended on the timing of the test. The subjects answered more questions correctly on the immediate multiple-choice test ($M = .66$) than on the delayed test ($M = .57$) [$F(1{,}71) = 49.08$, $MS_e = .03$, $\eta_p^2 = .41$]. Delay interacted with only one variable, reading status [$F(1{,}71) = 36.91$, $MS_e = .04$, $\eta_p^2 = .34$], which confirms the obvious point that when the subjects had not read the passages, the delay between reading and testing did not affect performance ($M = .52$ for both tests). The advantage gained from reading the passages was reduced after a week's delay (*M*s = .79 and .63 on the immediate and delayed tests, respectively), the usual finding of forgetting over time.

Performance on the Cued Recall Tests

The design allowed us to answer a number of questions about how multiple-choice testing affects later cued recall performance. Rather than analyzing all the conditions together, we made the comparisons necessary to answer questions of interest. We begin with an analysis of performance on the initial cued recall test in Schedule A (following immediate multiple-choice testing). This condition served as the control for most of the questions of interest and also extended the design in Roediger and Marsh (2005) from a cued recall test with a warning against guessing to a cued recall test with forced responding and confidence ratings.

Immediate Cued Recall: An Extension of Positive and Negative Consequences of Testing

As in Roediger and Marsh (2005), the number of prior multiple-choice alternatives had two separate, opposite effects on an immediate cued recall test. These data are shown in the top panels of Tables 3 (proportion of cued recall questions answered correctly) and 4 (proportion of cued recall questions answered with multiple-choice lures).

First, testing benefited later memory: The subjects correctly answered a greater proportion of cued recall ques-

Table 3
Proportions of Cued Recall (CR) Questions Answered Correctly, As a Function of Passage Reading, Number of Alternatives on the Prior Multiple-Choice (MC) Test, and Timing of Tests

		Number of Prior MC Alternatives				
Timing	Reading Status	Zero (Not Tested)	Two	Four	Six	*M*
Immediate cued recall, MC in Session 1	Read	.48	.76	.71	.65	.65
	Not read	.21	.54	.39	.39	.38
	M	.35	.65	.55	.52	
Delayed cued recall, MC in Session 1	Read	.30	.45	.47	.49	.43
	Not read	.23	.33	.29	.28	.28
	M	.27	.39	.38	.38	
Delayed cued recall, MC in Session 2	Read	.31	.63	.52	.45	.48
	Not read	.23	.51	.39	.33	.37
	M	.27	.57	.46	.39	
Delayed cued recall, CR and MC in Session 1	Read	.37	.59	.58	.57	.53
	Not read	.24	.38	.33	.35	.33
	M	.31	.49	.46	.46	

Table 4
Proportions of Cued Recall (CR) Questions Answered With Multiple-Choice (MC) Lures, As a Function of Passage Reading, Number of Alternatives on the Prior MC Test, and Timing of Tests

		Number of Prior MC Alternatives				
Timing	Reading Status	Zero (Not Tested)	Two	Four	Six	*M*
Immediate cued recall, MC in Session 1	Read	.13	.09	.14	.21	.15
	Not read	.19	.23	.40	.40	.30
	M	.16	.16	.27	.30	
Delayed cued recall, MC in Session 1	Read	.17	.14	.15	.14	.15
	Not read	.18	.21	.26	.30	.24
	M	.18	.18	.21	.22	
Delayed cued recall, MC in Session 2	Read	.19	.18	.30	.38	.26
	Not read	.19	.25	.36	.44	.31
	M	.19	.21	.33	.41	
Delayed cued recall, CR and MC in Session 1	Read	.17	.12	.15	.19	.16
	Not read	.18	.21	.29	.32	.25
	M	.18	.16	.22	.26	

tions if they had been tested previously on the multiple-choice test ($M = .57$) than if they had not ($M = .35$) [$F(1,71) = 161.24$, $MS_e = .02$, $\eta^2_p = .69$]. As on the initial multiple-choice test, the subjects answered more cued recall questions correctly if they had read the relevant passages [$F(1,71) = 190.13$, $MS_e = .03$, $\eta^2_p = .73$]. Reading status did not interact with testing ($F < 1$); the benefits of testing were equally strong for questions corresponding to read and not-read passages.

However, not all forms of prior testing were equal. That is, prior testing with two alternatives led to 65% correct on the cued recall test; this dropped to 55% following testing with four alternatives and 52% with six alternatives. This effect of number of prior multiple-choice alternatives was significant even when never-tested items were removed from the analysis [$F(2,142) = 16.41$, $MS_e = .04$, $\eta^2_p = .19$], and there was no interaction between passage reading and number of prior alternatives [$F(2,142) = 2.29$, $MS_e = .04$, $p > .10$].

A second negative consequence of testing was the intrusion of multiple-choice lures as answers on the immediate cued recall test; the relevant data are shown in the top panel of Table 4. That is, we scored whether each answer was one of the five possible multiple-choice lures for that item. The subjects were more likely to produce multiple-choice lures when they had not read the relevant passages ($M = .30$) than after reading the passages ($M = .15$) [$F(1,71) = 135.10$, $MS_e = .03$, $\eta^2_p = .66$]. Most important, the number of prior multiple-choice alternatives (zero [not tested], two, four, or six) affected the level of lure intrusions on the cued recall test [$F(3,213) = 23.29$, $MS_e = .03$, $\eta^2_p = .25$]. Multiple-choice lure intrusions increased linearly with number of prior alternatives for both read [$F(1,71) = 10.56$, $MS_e = .03$, $\eta^2_p = .13$] and nonread [$F(1,71) = 60.49$, $MS_e = .04$, $\eta^2_p = .46$] passages, but the pattern was stronger for nonread passages. In other words, the interaction between number of prior alternatives and reading status was significant [$F(3,213) = 8.86$, $MS_e = .03$, $\eta^2_p = .11$]. For nonread passages, multiple-choice lure intrusions increased from .19 without testing to .40 after testing with six alternatives, an increase of 21% [$t(71) = 6.79$, $SEM = .03$]. In contrast, after the relevant passages had been read, lure intrusions increased from .13 with zero alternatives to .21 with six prior alternatives—a smaller but still significant increase of 8% [$t(71) = 2.83$, $SEM = .03$].

As was described earlier, the subjects rated their confidence in their cued recall answers. These confidence ratings were used to assess the role of guessing in the negative testing effect. Critically, a similar pattern occurred when the lowest confidence (*not sure*) responses were removed from the analyses. The subjects produced more multiple-choice lure intrusions when the passages were nonread, as compared with read [$F(1,71) = 19.33$, $MS_e = .02$, $\eta^2_p = .21$], and the number of prior multiple-choice alternatives affected production of lure answers [$F(3,213) = 11.65$, $MS_e = .02$, $\eta^2_p = .14$]. As in the analysis with all the answers, multiple-choice lure intrusions increased linearly with number of previously read alternatives (zero [not tested], two, four, or six) for questions that referred to both read [$F(1,71) = 4.98$, $MS_e = .02$, $\eta^2_p = .07$] and nonread [$F(1,71) = 28.45$, $MS_e = .02$, $\eta^2_p = .29$] passages. Again, the increase in lure production was larger for nonread passages, leading to an interaction between the number of multiple-choice alternatives and prior reading [$F(3,213) = 6.75$, $MS_e = .02$, $\eta^2_p = .09$].

Finally, we examined the persistence of errors made on the multiple-choice test. That is, given that a lure was selected on the multiple-choice test, how likely was it that a lure was produced on the cued recall test? This analysis includes all the lures produced on the final test (as opposed to requiring it to be the same lure as that selected on the multiple-choice test), because prior work has shown that almost all lures produced on the final test match earlier selections (e.g., Marsh et al., 2009; Roediger & Marsh, 2005). Following the selection of a multiple-choice lure, 65% of the corresponding cued recall questions were answered with multiple-choice lures. In the later sections, we will use this number as a base rate to examine the effects of delay on the persistence of errors.

In short, as in Roediger and Marsh (2005), multiple-choice testing led to benefits on a cued recall test a few minutes later (a positive testing effect), and these benefits were reduced if the prior multiple-choice test had paired the correct answer with additional alternatives. The subjects were more likely to answer cued recall questions with multiple-choice lures following testing with additional multiple-choice alternatives, especially for read passages. In addition, the negative testing effect was not due to guessing on the cued recall test: The effect persisted even after guesses were removed from the analyses.

Did Delaying the Cued Recall Test Change the Impact of the Initial Multiple-Choice Test?

To isolate the effects of delaying the cued recall test, we compared performance on passage facts tested on the initial cued recall test in Schedule A with performance on the final cued recall test in Schedule B. In this comparison, the multiple-choice test always immediately followed the reading period, and the cued recall test occurred either immediately or 1 week after the multiple-choice test. The immediate condition is the one reported in the last section (the top panel in Tables 3 and 4), and the delayed condition is reported in the second panel of Tables 3 and 4.

We begin with an analysis of correct answers on the cued recall test, as shown in Table 3. Not surprisingly, delaying the cued recall test led to lower performance than was observed on the immediate cued recall test [$F(1,71) = 109.86$, $MS_e = .02$, $\eta^2_p = .61$]. Delaying the test also reduced the effects of having read the passages [$F(1,71) = 31.71$, $MS_e = .02$, $\eta^2_p = .31$].

Of particular interest was whether delaying the cued recall test would change the effects of prior testing. Interestingly, delaying the final test led to a reduction in the positive testing effect [$F(1,71) = 23.68$, $MS_e = .02$, $\eta^2_p = .25$]. As has been reported already, testing increased the proportion of final questions answered correctly on the immediate test to .57, relative to .35 in the nontested condition. When the final test was delayed, prior testing only increased the proportion of questions answered correctly from .27 (nontested) to .38 (tested). This 11% difference was significant [$t(71) = 7.06$, $SEM = .02$], but it was smaller than the increase from testing observed on the initial test [$M = .22$; $t(71) = 12.70$, $SEM = .02$]. There was also a marginally significant three-way interaction between passage reading, prior testing, and delay [$F(1,71) = 3.67$, $MS_e = .02$, $p = .06$, $\eta^2_p = .05$]. On the immediate test, the testing effect was similar for read and nonread passages (a benefit of 23% for previously tested items). However, delay reduced the testing effect more for nonread passages than for read passages. After a delay, the difference between tested and nontested items was 17% for read passages but only 7% for nonread passages. Having read the passages helped protect the benefits of testing over the delay.

Next, we examined whether delaying the final cued recall test would have consequences for the negative testing effect. Two analyses are relevant to this question. First is whether the positive testing effect was smaller following testing with additional lures. The second analysis involves the proportion of cued recall questions answered with multiple-choice lures.

If one looks only at performance on the delayed cued recall test (the second panel in Table 3), the effect of number of prior multiple-choice alternatives on correct recall disappeared ($F < 1$). The proportion of correct cued recall answers remained constant at .38 following testing with two, four, or six alternatives. There was a hint that the number of prior multiple-choice alternatives had different effects on correct answers for read passages (actually increasing performance following testing with more alternatives) than for nonread passages (where performance decreased following testing with more alternatives), but the interaction failed to reach significance [$F(2,142) = 2.52$, $MS_e = .03$, $p = .08$, $\eta^2_p = .03$]. Overall, performance on the delayed test differed from that observed on the immediate test (where cued recall performance decreased when the number of prior alternatives increased from two to six, for both read and nonread passages). The different patterns on the immediate and delayed tests led to an interaction between the number of prior alternatives and delay [$F(2,142) = 7.72$, $MS_e = .04$, $\eta^2_p = .10$]. The three-way interaction between delay, reading status, and number of prior multiple-choice alternatives was nonsignificant [$F(2,142) = 1.45$, $MS_e = .03$, $p = .24$].

Second, did the subjects still answer the cued recall questions with multiple-choice lures if the final test was delayed for 1 week? The answer is yes; an analysis of the delayed cued recall test revealed that multiple-choice lure intrusions increased linearly with number of previously read alternatives [$F(1,71) = 6.71$, $MS_e = .03$, $\eta^2_p = .09$]. This increase, however, was smaller when the cued recall test was delayed by 1 week, as reflected by an interaction between delay and number of prior alternatives [$F(3,213) = 5.22$, $MS_e = .03$, $\eta^2_p = .07$]. Lure production increased from .16 with zero alternatives (not-tested items) to .30 with six alternatives on the immediate test, a difference of 14% [$t(71) = 6.64$, $SEM = .02$]. On the delayed test, lure production increased from .18 with zero alternatives to .22 with six alternatives, a difference of only 4%, but this difference was still significant [$t(71) = 2.17$, $SEM = .02$]. Lure production remained stable over the delay for questions referring to read passages but decreased over time for questions referring to nonread passages. This led to an interaction between reading status and delay [$F(1,71) = 14.08$, $MS_e = .03$, $\eta^2_p = .17$]. The three-way interaction between delay, reading status, and number of multiple-choice alternatives was not significant [$F(3,213) = 1.85$, $MS_e = .03$, $p = .14$].

Similar negative testing effects were observed after the lowest confidence responses were removed from the analyses. Paralleling the main analyses, there was an interaction between delay and number of prior multiple-choice alternatives after guesses were removed [$F(3,213) = 3.89$, $MS_e = .02$, $\eta^2_p = .05$]. Multiple-choice lure intrusions increased from .08 with zero prior alternatives (not tested) to .15 following six alternatives on the immediate test, an increase of 7% [$t(71) = 4.15$, $SEM = .02$]. The increase in multiple-choice lure intrusions was smaller but still significant on the delayed test. Lure intrusions increased from .06 for not-tested items to .08 for questions previously tested with six alternatives [$t(71) = 2.06$, $SEM = .01$]. Again, delay reduced lure intrusions for nonread passages, whereas the overall level of lure intrusions did not change over time for read passages, resulting in an interaction between delay and reading status [$F(1,71) = 5.36$, $MS_e = .02$, $\eta_p^2 = .07$].

Finally, we examined whether delay affected the persistence of errors made on the multiple-choice test. Of interest was whether a cued recall question would be answered with one of the multiple-choice lures, given that an error was made on the parallel multiple-choice question. Critically, delay reduced the likelihood that a multiple-choice error led to a lure intrusion on the final test. Sixty-five percent of the initial multiple-choice errors led to lure intrusions on the immediate cued recall test, whereas only 36% of the multiple-choice errors led to lure intrusions on cued recall test after 1 week [$t(71) = 10.70$, $SEM = .03$].

In summary, delaying the cued recall test reduced both the positive and negative effects of testing. Prior testing increased later production of correct answers on both the immediate and delayed tests, but the increase was smaller when the tests were separated by 1 week. Delay reduced both negative consequences of testing. First, after a delay, the number of prior multiple-choice alternatives no longer affected correct answers on the cued recall test. The positive testing effect was similar following testing with two, four, or six prior alternatives. Second, delaying the cued recall test also reduced the intrusion of multiple-choice lures, although this negative testing effect was not eliminated.

Did Delaying the Initial Multiple-Choice Test Change Its Impact on the Final Cued Recall Test?

To isolate the effects of the timing of the initial multiple-choice test, this analysis was limited to performance on the final cued recall test. We compared performance on the final test as a function of whether passages were assigned to the immediate multiple-choice testing condition (Schedule B in Table 1) or the delayed multiple-choice testing condition (Schedule C in Table 1). Thus, the delay between study and the final cued recall test was constant in the two groups; only the placement of the multiple-choice test varied.

A comparison of the second and third panels of Table 3 reveals that the positive testing effect was larger when the multiple-choice test occurred in the second session, immediately before the cued recall test, rather than a week earlier [$F(1,71) = 10.90$, $MS_e = .02$, $\eta_p^2 = .13$]. When both tests occurred in the second session (panel 3), cued recall performance was much better for previously tested items ($M = .47$) than for previously untested items ($M = .27$) [$t(71) = 11.51$, $SEM = .02$]. When the multiple-choice test had occurred a week earlier (panel 2), subjects still correctly answered more cued recall questions from passages that had been tested previously ($M = .38$) than from nontested passages ($M = .27$) [$t(71) = 7.06$, $SEM = .02$]. However, this testing effect was reduced relative to the testing effect observed when both the multiple-choice and the cued recall test were delayed.

The timing of the multiple-choice test also affected whether or not all the forms of testing were equivalent. When both the multiple-choice and cued recall tests occurred during the second session, performance decreased from .57 to .46 to .39 as the number of prior alternatives increased from two to four to six [$F(2,142) = 30.17$, $MS_e = .04$, $\eta_p^2 = .30$]. As was reported in the previous section, when the multiple-choice and cued recall tests occurred in different sessions, there was no effect of number of prior multiple-choice alternatives (two vs. four vs. six) on cued recall performance. These two different patterns led to an interaction between timing of the multiple-choice test and number of prior multiple-choice alternatives [$F(2,142) = 15.70$, $MS_e = .04$, $\eta_p^2 = .18$]. The three-way interaction between timing of the multiple-choice test, number of prior multiple-choice alternatives, and reading status was not significant [$F(2,142) = 1.51$, $MS_e = .03$, $p = .22$].

The timing of the multiple-choice test significantly affected the production of multiple-choice lure intrusions on the final cued recall test. These data appear in the second and third panels of Table 4. The effect of testing with additional multiple-choice alternatives was larger when the two tests occurred in the same session, as reflected in an interaction between delay and number of prior alternatives [$F(3,213) = 13.77$, $MS_e = .03$, $\eta_p^2 = .16$]. When facts were tested twice in the second session, multiple-choice lure answers increased from .19 for not-tested items to .41 for items tested with six alternatives [$t(71) = 10.16$, $SEM = .02$]. In contrast, when the multiple-choice test had occurred a week earlier, multiple-choice lure answers on the final test showed a smaller (but still significant) increase to .22 after testing with six alternatives (as compared with a baseline of .18) [$t(71) = 2.17$, $SEM = .02$]. Delaying the multiple-choice test until the second session also reduced the benefits of having read the passages. When the multiple-choice test occurred just before the cued recall test, lure production was high, and reading provided less protection against the negative testing effect [$F(1,71) = 4.85$, $MS_e = .03$, $\eta_p^2 = .06$].

The timing of the multiple-choice test still affected the negative testing effect after guesses were removed from the analysis. Multiple-choice lure answers increased 10% with increasing alternatives when both tests occurred in the second session, as compared with 2% when the multiple-choice test occurred a week earlier [$F(3,213) = 4.90$, $MS_e = .02$, $\eta_p^2 = .07$]. After guesses were removed, the interaction between delay and reading status was no longer significant ($F < 1$). Questions referring to both read and nonread passages produced equal lure production at both delays.

Finally, we examined the proportion of multiple-choice errors that persisted onto the final cued recall test. That is, given a multiple-lure selection, how likely were subjects to produce a multiple-choice lure on the corresponding final cued recall question? More errors persisted when the two tests were held in the same session ($M = .48$) than when the tests occurred a week apart ($M = .36$) [$t(71) = 4.73$, $SEM = .02$].

In summary, delaying the multiple-choice test increased both its positive and negative effects on the final cued recall test. Prior testing increased correct answers in both conditions, but especially when the multiple-choice test was close in time to the final test. The delayed multiple-choice test also led to greater intrusions of multiple-choice lures on the final test, as reflected in the higher persistence rate.

Did Testing Effects Observed on the Immediate Cued Recall Test Persist Until the Delayed Cued Recall Test?

To examine whether testing effects observed on an immediate cued recall test persisted over a 1-week delay, we compared performance on the initial cued recall test (following multiple-choice testing) with performance with the same items on the final cued recall test. Referring to Table 1, we compared performance on the initial and final cued recall tests for Schedule A.

The positive testing effect observed on the initial cued recall test (as shown in the top panel of Table 3) was retained on the delayed cued recall test (as shown in the bottom panel of Table 3). On the final test, the subjects correctly answered 47% of the items that had been tested on both the multiple-choice and cued recall tests in the first session. This was significantly above the baseline of 31% for items that had been tested on the initial cued recall test but had *not* been tested on the initial multiple-choice test [$t(71) = 8.86$, $SEM = .02$]. However, this testing effect was significantly smaller than the one observed on the immediate cued recall test, where performance increased from .35 to .57, leading to an interaction between testing and timing of the cued recall test [$F(1,71) = 18.90$, $MS_e = .01$, $\eta_p^2 = .21$]. There was also a three-way interaction between passage reading, prior testing, and delay [$F(1,71) = 10.83$, $MS_e = .01$, $\eta_p^2 = .13$]. Delay affected the testing effect only for nonread passages. For read passages, testing boosted performance by 23% on the immediate test and 21% on the delayed test. In contrast, for nonread passages, testing boosted performance by 23% on the immediate test, but this dropped to 11% on the final test.

On the final test, all forms of prior multiple-choice testing led to similar levels of correct responding. Although the number of prior multiple-choice alternatives had an effect on correct answers on the immediate cued recall test, this effect did not appear when the same questions were asked again on the second cued recall test [leading to an interaction between delay and number of prior alternatives; $F(2,142) = 14.04$, $MS_e = .01$, $\eta_p^2 = .17$]. On the first test, as described earlier, correct answers declined when the subjects had been tested with more multiple-choice alternatives, from .65 to .52. On the second test, however, performance dropped from .49 to .46, and a linear trend analysis on these delayed data was not significant [$F(1,71) = 1.39$, $MS_e = .04$, $p = .24$].

However, as is shown in Table 4, the pattern of lure intrusions seen on the first cued recall test also appeared on the final test, albeit to a lesser extent. An examination of the final test revealed that multiple-choice lure intrusions increased linearly with number of prior multiple-choice alternatives [$F(1,71) = 19.55$, $MS_e = .03$, $\eta_p^2 = .22$]. However, this pattern was not as strong as that observed on the first cued recall test, leading to an interaction between delay and number of prior alternatives [$F(3,213) = 6.54$, $MS_e = .01$, $\eta_p^2 = .08$]. On the initial cued recall test, lure intrusions increased from .16 with zero alternatives (not tested) to .30 after prior multiple-choice testing with six alternatives [$t(71) = 6.64$, $SEM = .02$]. This difference was significant but smaller on the second cued recall test. Lure intrusions increased from .18 following zero alternatives to .26 following six alternatives [$t(71) = 3.85$, $SEM = .02$]. Lure production dropped more over the delay for nonread passages, as compared with read passages. This led to an interaction between delay and reading status [$F(1,71) = 31.14$, $MS_e = .01$, $\eta_p^2 = .31$].

Removing guesses from the analyses did not change the conclusions about the persistence over 1 week of the negative testing effects observed on the initial cued recall test. When guesses were excluded, intrusions increased from .08 to .15 on the first cued recall test and from .07 to .11 on the second cued recall test, meaning that number of prior alternatives and delay interacted [$F(3,213) = 3.67$, $MS_e = .01$, $\eta_p^2 = .05$]. Again, questions referring to nonread passages showed larger decreases in lure production over the delay, but the interaction between delay and reading status was now only marginally significant [$F(1,71) = 3.69$, $MS_e = .01$, $p = .06$, $\eta_p^2 = .05$].

Finally, we examined whether errors on the initial multiple-choice test were associated with errors on the cued recall tests. Errors on the multiple-choice test were more likely to lead to errors on the immediate cued recall test ($M = .65$) than on the delayed cued recall test ($M = .48$) [$t(71) = 9.05$, $SEM = .02$]. In other words, some of the errors that were repeated on the first cued recall test were forgotten by the final test.

In short, when the same questions were asked on immediate and delayed cued recall tests, similar effects of prior multiple-choice testing were observed on the two tests, although the effects were reduced on the delayed test.

DISCUSSION

The first contribution of this experiment was to extend Roediger and Marsh's (2005) finding of positive and negative testing effects to a test with forced responding. Whereas Roediger and Marsh instructed subjects not to guess on the final cued recall test and to answer only the questions to which they knew the answer, we instructed subjects to answer every question, even if they had to guess. This instruction is much more similar to what occurs in educational situations. Given that most instructors do not penalize students for guessing, there is a strong incentive for students to answer every question, even if they have to guess. We thought the results might change with the new instructions, with the possibility that allowing guesses would increase the negative effects of testing.

On the whole, our results were similar to those found by Roediger and Marsh (2005). On an immediate cued recall test, there was a positive testing effect: The subjects were more likely to answer cued recall questions correctly if they had occurred on the multiple-choice test. This positive effect of testing decreased following exposure to additional lures on the prior multiple-choice test. Having read additional lures also increased the likelihood that cued recall questions would be answered with multiple-choice lures. All of these results nicely parallel those of Roediger and Marsh. One difference involves the overall level of lure intrusions, which was much higher in the present experiment (*M* = .22) than in Roediger and Marsh (*M* = .09). However, because this increase was also observed in the baseline (not-tested) condition, it does not change the conclusions. The only substantive difference between the two experiments involved the effects of having read the passages. In the present study, the negative testing effect was reduced following passage reading. This pattern is similar to that in Roediger and Marsh numerically, although the interaction between reading status and number of prior alternatives did not reach significance in their study. In general, passage reading protects against the negative effects of multiple-choice testing. When students are well prepared for the multiple-choice test, the negative effects of testing are reduced. Interestingly, the present experiment shows that if passage reading and the multiple-choice test are separated by 1 week (Schedule C), reading no longer protects against lure intrusions.

The second, larger, contribution of this experiment was to examine the effects of delay on positive and negative testing effects. We asked three main questions. First, does taking a multiple-choice test still yield positive and negative testing effects if the final cued recall test is delayed 1 week? Second, does delaying the multiple-choice test (to a week after reading) change its impact on the final cued recall test? Third, do the testing effects observed on an initial cued recall test appear on a final cued recall test a week later? We discuss the answers to these questions below, before turning to a more general discussion of the experiment.

To guide this discussion, Figures 1 and 2 show a summary of the effects of delay on the positive and negative effects of prior testing. For the purposes of the figures, we collapsed across read and nonread passages. These simplified figures highlight the most important findings to be discussed below. To preview, all the positive and negative testing effects were significant, but the size of the effects differed dramatically across conditions.

Delayed Effects of Immediate Multiple-Choice Testing

To determine whether a multiple-choice test still affected later responding after a delay, we compared performance on the initial cued recall test in Schedule A with performance on the final cued recall test in Schedule B (see Table 1). That is, we examined performance on the cued recall test as a function of whether it was taken immediately or 1 week after the multiple-choice test. To summarize, across our different dependent measures (some of which excluded guesses), positive testing effects were reduced but were still present when the cued recall test occurred 1 week after the initial session. On both immediate and delayed cued recall tests, the subjects correctly answered more questions if they had been previously tested on the multiple-choice test. This positive testing effect was larger, however, when the cued recall test immediately followed the multiple-choice test. In addition, increasing numbers of multiple-choice alternatives decreased performance on the immediate cued recall test but had no effect after 1 week. Likewise, the negative effects of testing were reduced over the delay but still occurred. Taking a multiple-choice test increased production of multiple-choice lures on the final cued recall test, especially after more multiple-choice alternatives had been

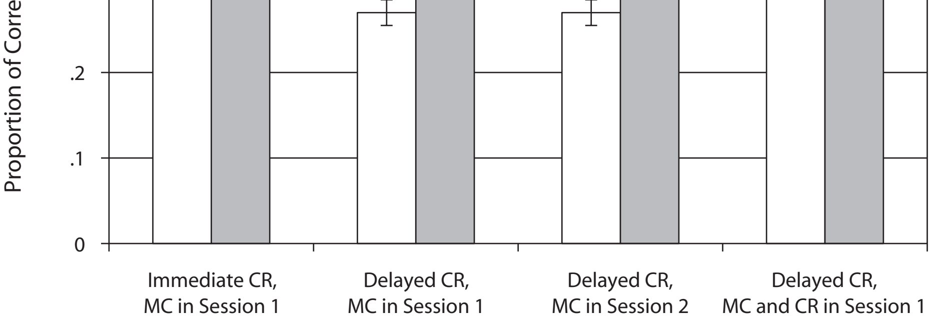

Figure 1. Positive effects of prior multiple-choice (MC) testing on later cued recall (CR) performance, as a function of test timing.

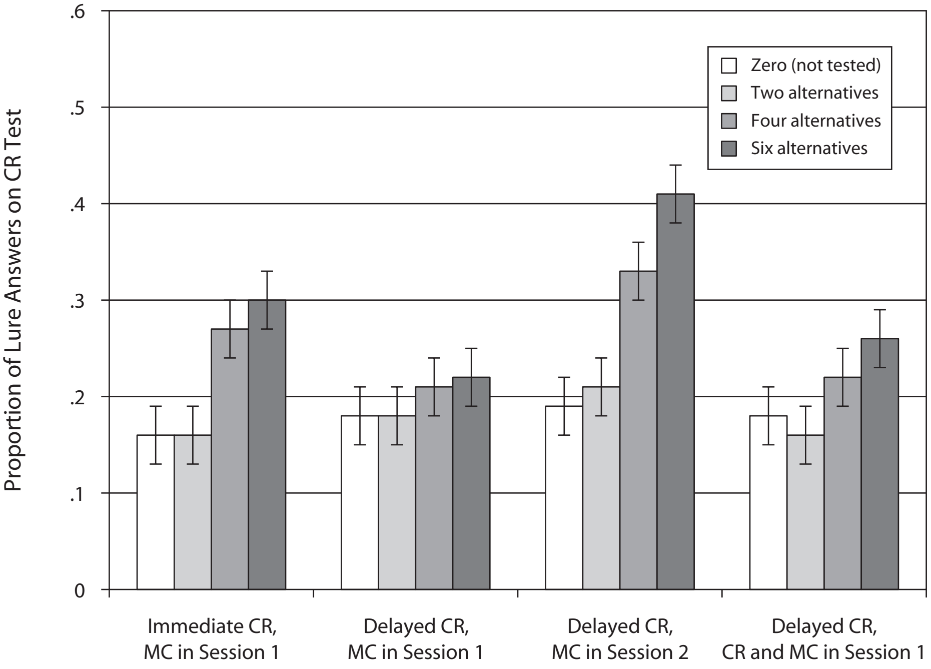

Figure 2. Negative effects of prior multiple-choice (MC) testing on later cued recall (CR) performance, as a function of test timing.

read. Although these effects persisted over the delay, they were reduced.

Effects of Delaying the Multiple-Choice Test

To determine the effects of delaying the multiple-choice test for 1 week, we compared performance on the final cued recall test in Schedules B and C (see Table 1). That is, we compared performance on the final cued recall test as a function of whether the subjects had taken an immediate multiple-choice test (after the reading phase, a week before the final cued recall test) or a delayed multiple-choice test (immediately before the final test). To summarize, both positive and negative testing effects were larger when the multiple-choice test was delayed and occurred immediately before the final test.

Persistence of Testing Effects

As is shown in Table 1, testing Schedule A provided an opportunity to see whether testing effects observed on an initial cued recall test would persist until the final cued test a week later. Again, the short answer is *yes*. Although the effects were reduced over the delay, in general, the positive and negative testing effects observed on first cued recall test were also observed on the second cued recall test.

In summary, three general points emerged from the experiment. First, both the positive and negative effects of prior testing were strongest when the multiple-choice test and the cued recall test occurred in the same session. It was less important whether both of them had occurred in the first session or in the second session. Rather, separation in time between the tests reduced the effects of testing. Second, the negative testing effect decreased over the delay but was never eliminated. Third, it should be noted that the positive testing effect was very robust. Both immediately and after the delay, the positive testing effect was always greater than the negative testing effect. That is, the increase in correct answers following testing was always larger than the increase in multiple-choice lure answers. The net result of prior multiple-choice testing was always positive.

We first comment on the theoretical implications of our results and then turn to practical recommendations. In particular, our findings are consistent with prior work that suggests that recollective processing underlies the benefits of testing (Chan & McDermott, 2007; Karpicke et al., 2006). Chan and McDermott had subjects study two lists of words: In the tested condition, the subjects were given a free recall test after each list, whereas in the not-tested condition, the subjects solved math problems. At the end of the experiment, both groups completed a final recognition test on the words from both lists. The subjects in the tested condition were better able to remember on which list the words appeared and gave more *remember* responses on the recognition test than did the subjects in the not-tested condition. These results suggest that testing increases later recollection processes, rather than increasing familiarity.

Our work extends this recollection account beyond the positive effects of testing to the negative testing effect. We manipulated a variable thought to have a large impact on recollection; *delay*, and it had similar effects on positive and negative testing effects. The fact that the negative testing effect decreased over the delay suggests that recollecting the multiple-choice lures is a prerequisite for the negative testing effect.[2] This is in contrast to other false memory paradigms such as false fame, where memory errors are caused by a reliance on familiarity in the absence of recollection (Jacoby, Woloshyn, & Kelley, 1989).

On the basis of our results, what advice can be offered to educators? Because teachers should want to retain the positive (but not the negative) effects of testing, the fact that the temporal spacing of tests has an impact on positive testing effects leads to the recommendation that frequent quizzes should be given to enhance students' knowledge. At a delay of 1 week, the positive testing effect still outweighs the negative, but it is an open question as to whether the positive testing effect will still prevail at longer delays. Finally, given that the negative effects of testing persist over time, educators should be aware of the costs of multiple-choice tests and try to reduce these hazards. One easy intervention is providing feedback after a test, which increases later correct responding and decreases later production of multiple-choice lures (Butler & Roediger, 2008). In short, we believe that frequent tests given with feedback will increase students' knowledge, while avoiding the negative effects of testing.

AUTHOR NOTE

This research was supported by a collaborative activity award from the James S. McDonnell Foundation. We thank Holli Sink and Aaron Johnson for help with manuscript preparation. The opinions expressed are those of the authors and do not represent the views of the James S. McDonnell Foundation. Correspondence concerning this article should be addressed to L. K. Fazio, Department of Psychology and Neuroscience, Duke University, 417 Chapel Drive, Durham, NC 27708 (e-mail: lkf@duke.edu).

Note—Accepted by the previous editorial team,
when Brian H. Ross was Editor.

REFERENCES

Barber, S. J., Rajaram, S., & Marsh, E. J. (2008). Fact learning: How information accuracy, delay, and repeated testing change retention and retrieval experience. *Memory*, **16**, 934-946.

Brainerd, C. J., & Reyna, V. F. (1996). Mere memory testing creates false memories in children. *Developmental Psychology*, **32**, 467-478.

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

Brown, A. S., & Marsh, E. J. (2008). Evoking false beliefs about autobiographical experience. *Psychonomic Bulletin & Review*, **15**, 186-190.

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 (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, **36**, 604-616.

Cave, C. B. (1997). Very long-lasting priming in picture naming. *Psychological Science*, **8**, 322-325.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **33**, 431-437.

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.

Jacoby, L. L., Kelley, C. [M.], Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality & Social Psychology*, **56**, 326-338.

Jacoby, L. L., Woloshyn, V., & Kelley, C. [M.] (1989). Becoming famous without being recognized: Unconscious influences of memory produced by dividing attention. *Journal of Experimental Psychology: General*, **118**, 115-125.

Karpicke, J. D., McCabe, D. P., & Roediger, H. L., III (2006, November). *Testing enhances recollection: Process dissociations and metamemory judgments*. Poster presented at the Annual Meeting of the Psychonomic Society, Houston, TX.

Kelley, C. M., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory & Language*, **32**, 1-24.

Marsh, E. J., Agarwal, P. K., & Roediger, H. L., III (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, **15**, 1-11.

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.

Mitchell, D. B. (2006). Nonconscious priming after 17 years: Invulnerable implicit memory? *Psychological Science*, **17**, 925-929.

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning & Verbal Behavior*, **19**, 338-368.

Roediger, H. L., III, Agarwal, P. K., Kang, S. H. K., & Marsh, E. J. (2010). Benefits of testing memory: Best practices and boundary conditions. In G. M. Davies & D. B. Wright (Eds.), *Current issues in applied memory research* (pp. 13-49). Hove, U.K.: Psychology Press.

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.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, **30**, 641-656.

Toppino, T. C., & Brochin, H. A. (1989). Learning from tests: The case of true–false examinations. *Journal of Educational Research*, **83**, 119-124.

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

Yonelinas, A. P., & Levy, B. J. (2002). Dissociating familiarity from recollection in human recognition memory: Different rates of forgetting over short retention intervals. *Psychonomic Bulletin & Review*, **9**, 575-582.

NOTES

1. This argument for long-term persistence of familiarity does not contradict recent findings about familiarity's dropping quickly in the very short term (e.g., Yonelinas & Levy, 2002). It may very well be that familiarity drops off more quickly than recollection initially but that familiarity is more stable than recollection over longer delays.

2. Because both familiarity and recollection likely drop over a delay, it is impossible to definitively say that subjects' reliance on the prior multiple-choice lures is due to recollection. However, our interpretation (that familiarity is more stable over time than recollection is, meaning that delay primarily affects recollection) is consistent with how recollection and familiarity are conceptualized in other paradigms, such as false fame.

(Manuscript received February 8, 2008;
revision accepted for publication March 25, 2010.)