EUROPEAN JOURNAL OF COGNITIVE PSYCHOLOGY
2007, 19 (4/5), 528–558

Psychology Press
Taylor & Francis Group

Test format and corrective feedback modify the effect of testing on long-term retention

Sean H. K. Kang, Kathleen B. McDermott, and Henry L. Roediger, III
Washington University in St Louis, St Louis, MO, USA

We investigated the effects of format of an initial test and whether or not students received corrective feedback on that test on a final test of retention 3 days later. In Experiment 1, subjects studied four short journal papers. Immediately after reading each paper, they received either a multiple choice (MC) test, a short answer (SA) test, a list of statements to read, or a filler task. The MC test, SA test, and list of statements tapped identical facts from the studied material. No feedback was provided during the initial tests. On a final test 3 days later (consisting of MC and SA questions), having had an intervening MC test led to better performance than an intervening SA test, but the intervening MC condition did not differ significantly from the read statements condition. To better equate exposure to test-relevant information, corrective feedback during the initial tests was introduced in Experiment 2. With feedback provided, having had an intervening SA test led to the best performance on the final test, suggesting that the more demanding the retrieval processes engendered by the intervening test, the greater the benefit to final retention. The practical application of these findings is that regular SA quizzes with feedback may be more effective in enhancing student learning than repeated presentation of target facts or taking an MC quiz.

A wealth of empirical research has found that retention of studied material can be enhanced by testing. A memory test does not merely measure the amount of learning; it has an impact on the state of that memory itself (Lachman & Laughery, 1968; Tulving, 1967). Subjects who

Correspondence should be addressed to Sean Kang, Department of Psychology, Washington University, St Louis, MO 63130-4899, USA. E-mail: seankang@wustl.edu

This research was conducted as part of the Master's Thesis of the first author, and was supported by a grant from the Institution of Education Sciences (R305H030339) to the second and third authors. The results from Experiment 2 were presented as a poster at the 17th annual convention of the American Psychological Society, Los Angeles, CA, in May 2005. The authors acknowledge Mark McDaniel for his valuable comments on an earlier version of this manuscript. Appreciation is also extended to Seth Goodman for assistance with subject testing and data entry.

http://www.psypress.com/ecp DOI: 10.1080/09541440601056620

receive an intervening test after the initial learning experience typically perform better on a later final test, relative to subjects given only the final test. This phenomenon has come to be referred to as the *testing effect*, and has been demonstrated with diverse study stimuli, including word lists (Darley & Murdock, 1971), paired associates (Runquist, 1986), pictures (Wheeler & Roediger, 1992), general knowledge facts (McDaniel & Fisher, 1991), and prose passages (LaPorte & Voss, 1975; for a review, see Roediger & Karpicke, 2006a). One question that remains to be resolved is how the format of the initial test influences the testing effect. Do multiple choice (MC) or short answer (SA) tests differ in the benefit they produce on a final test? Does the answer depend on the format of the final test? This issue is important for both practical and theoretical reasons.

EDUCATIONAL RELEVANCE

In education, tests have typically been employed solely as assessment tools for evaluating learning and academic progress (Dempster, 1996). High-stakes achievement tests have acquired a less than desirable reputation in recent years, with critics charging that such tests are culturally biased and encourage instructors to devote too much time teaching to the test (Anderson, 1998). One concern is that this may result in a spillover effect, such that educators develop an aversion to all types of testing. The fact overlooked in this debate is that tests can be utilised as instruments to promote learning and retention, as demonstrated by Spitzer (1939), who tested the entire sixth-grade population of 91 elementary schools in Iowa. After students read a paper, he varied the number of tests and the retention intervals between tests, and found that students who were tested soon after reading the paper retained the material better on a test given 63 days later. Despite this and other early studies yielding similar findings (Gates, 1917; Jones, 1923), this benefit of testing has remained largely untapped by educators, with the research findings rarely communicated in teacher education courses or implemented in pedagogical practice (Dempster & Perkins, 1993). A major goal of educators is the long-term retention of knowledge acquired by students (Halpern & Hakel, 2002). The current study is an effort to provide evidence-based recommendations for the use of testing to enhance learning and retention, by using more educationally relevant materials than word lists. To discover the type of testing that may be best for implementation in educational settings, our study compared two testing formats that are commonly used in classrooms—MC and SA—to ascertain which type of format would be more beneficial for later retention. An MC test involves recognition: The subject has to discriminate among the options provided in order to choose

the answer. An SA test, on the other hand, involves response production: The subject must retrieve and generate an answer in response to the question cue.

THEORETICAL ISSUES

In addition to the practical issues surrounding the testing effect, we were also interested in examining the theoretical mechanisms that underlie it. One theoretical account is that the engagement of retrieval processes during an initial test modifies the memory trace of target items (Bjork, 1975), increasing the probability of successful retrieval later. Many studies have shown that receiving a test, relative to having no test, improves later retention (e.g., Darley & Murdock, 1971; Runquist, 1983, 1986, 1987). However, this sort of comparison leaves open the possibility that it is the re-presentation of an item which occurs during the test, rather than the act of retrieval per se, that enhances retention. Indeed, from the results of a multitrial free recall experiment in which the number and sequence of study and test trials were varied, Tulving (1967) concluded that study and test trials seem to facilitate subsequent recall to the same extent. Other researchers have also argued that a presentation at test (i.e., successfully recalling an item) is functionally similar to a study presentation, and that any effect of testing is due to an overlearning of a subset of items (Slamecka & Katsaiti, 1988; Thompson, Wenger, & Bartling, 1978). More recent research, however, has shown that the testing effect cannot be wholly accounted for by the amount of exposure to the tested material, since the testing effect is still obtained when memory for tested items is compared to memory for items that are re-presented but not tested (Carrier & Pashler, 1992; Kuo & Hirshman, 1996; Roediger & Karpicke, 2006b; Wheeler, Ewers, & Buonanno, 2003).

According to the transfer appropriate processing framework (Blaxton, 1989; Morris, Bransford, & Franks, 1977), memory performance depends on the overlap between encoding and retrieval processes. Applied to the current context, this framework provides an alternative explanation for the testing effect: It is the engagement of similar operations on the intervening and final tests that results in better performance for previously tested items, relative to items that were not initially tested or only restudied.

Several studies have examined the issue of how test format affects later memory performance. It has been demonstrated that taking a test of a particular format can still lead to a positive transfer to a later test of a different format: For example, prior recall tests facilitate subsequent recognition (Hanawalt & Tarr, 1961; Lockhart, 1975; Wenger, Thompson, & Bartling, 1980), prior recognition tests facilitate subsequent recall, at

least sometimes (Hogan & Kintsch, 1971, Exp. 1; Runquist, 1983, Exp. 1), and items tested initially in either MC or SA format still show a testing effect when the format is reversed on the final test (Nungester & Duchastel, 1982). However, to confidently adjudicate between the transfer appropriate processing and retrieval effort hypotheses, it is necessary to manipulate the formats of both the intervening and final tests, and only a handful of studies have done so. Support for the transfer appropriate processing view comes from a study by Duchastel and Nungester (1982), who found that taking an intervening MC test produced better performance on a final MC test than taking an intervening SA test, and likewise taking an intervening SA test produced (numerically) better performance on a final SA test than taking an intervening MC test. The authors attributed this benefit to performance for items tested in the same format as before to a "test practice effect".

There has also been some evidence that the more demanding or effortful the retrieval, the greater the enhancement to later memory performance. Glover (1989, Exps 4a, 4b, and 4c) compared the effect of three different types of intervening tests—free recall, cued recall, and recognition—on different types of final tests 4 days later. Regardless of the format of the final test, subjects who received an intervening free recall test performed best, followed in order by those who received an intervening cued recall test, those who received an intervening recognition test, and those who did not receive a prior test. Glover assumed that the amount or completeness of retrieval processing increased from recognition to cued recall to free recall, and concluded that the more complete the retrieval operations during the intervening test, the greater the benefit to final memory performance. Unfortunately, the number of idea units from the studied passage tested on the intervening tests was not equated across the three test types (e.g., on the cued recall test, 12 of the 24 idea units were tested, whereas on the recognition test, six idea units were tested with six distractor sentences), thus the possibility that differential amounts of testing led to the pattern of results cannot be definitively precluded. Also, Glover did not include a condition where subjects were reexposed to the material without taking a test; thus it is possible that it was the representation at test—not necessarily retrieval during the test—that produced the effect.

More recently, Carpenter and DeLosh (2006, Exp. 1), using word lists and including a restudy control condition, replicated Glover (1989, Exp. 4). However, it is remarkable that on all the types of final tests, the intervening free recall condition, which produced the best performance relative to the other intervening test conditions, did not significantly outperform the restudy condition. This could possibly be due to the rather brief delay (i.e., 5 min) before the administration of the final test, as

retention interval has been shown to be an important moderator of the testing effect (Roediger & Karpicke, 2006b; Wenger et al., 1980; Wheeler et al., 2003).

The present experiments examined the theoretical underpinnings of the testing effect in a manner that would have direct application to pedagogical practice. Like Glover (1989, Exp. 4) and Carpenter and DeLosh (2006, Exp. 1), we factorially manipulated the formats of the intervening and final tests, with a within-subjects design (Glover used a between-subjects design, and Carpenter & DeLosh used a mixed design). Unlike Carpenter and DeLosh, however, we used educationally relevant journal papers as our study material instead of word lists, and our retention interval before the final test was longer (i.e., 3 days instead of 5 min). Also, unlike Glover, we included a condition in which subjects read equivalent target statements, which provided a focused reexposure to the target material, so as to determine to what extent prior testing boosts later retention beyond representation. If the processes engaged during memory retrieval are crucially responsible for the testing effect, then one might expect that the more demanding or effortful the retrieval during a test, the better that material will be remembered later. This retrieval demands hypothesis would predict that an intervening SA test would result in better performance on the final test than an intervening MC test, regardless of whether the final test was MC or SA format. A straightforward prediction from the transfer appropriate processing framework would be that performance on a final memory test benefits most when the test format matches that of the earlier test, in which case an intervening MC test would enhance final MC items more than an intervening SA test, and an intervening SA test would enhance final SA items more than an intervening MC test. Such a prediction, of course, presumes that MC and SA tests engage disparate memorial operations or processes.

Two experiments were conducted, and they were identical except that the second experiment incorporated feedback. In both experiments, subjects studied brief journal papers and then received either an MC test, an SA test, read statements that repeated the relevant information, or did a filler questionnaire. Three days later, subjects returned for a final test. Experiment 1 was conducted without the provision of feedback to subjects, as an analogue of situations when a classroom instructor dispenses with feedback. In Experiment 2, the correct answer was provided after subjects answered each question on the initial test. This was done to examine the role of corrective feedback in test enhanced learning.

EXPERIMENT 1

Method

Subjects

Forty-eight undergraduates from the Washington University Psychology Subject Pool participated in partial fulfilment of course requirements or for $20 cash.

Materials

Study passages. Four papers from the journal *Current Directions in Psychological Science* (American Psychological Society) were selected as study material. Tables and figures, if present, were removed to homogenise the papers. (Information contained in the tables and figures was redundant for these papers, and hence their removal did not compromise the coherence of the papers.) The average length of the papers was about 2500 words.

Tests. From each paper, eight facts or concepts were selected. These facts were tested in multiple choice (MC) and short answer (SA) formats. In the MC format, subjects had to choose a response among four options, whereas in the SA format, they had to fill in the blank or generate a phrase or sentence to answer the question (questions taken from Anastasio, Rose, & Chapman, 1999; Eagly, Kulesa, Chen, & Chaiken, 2001; Garry & Polaschek, 2000; Treiman, 2000; see Appendix). These facts were also rephrased into one-sentence statements for use in the read statements condition, where subjects read the answers to the test questions without actually attempting the questions.

Design

A 4 (intervening task: MC, SA, read statements, or filler/control) $\times$ 2 (final test format: MC or SA) within-subjects design was used. During the first session, subjects studied the four papers. The order of the intervening tasks was kept constant across subjects: They took an MC test immediately after reading the first paper, took an SA test after reading the second paper, read a list of statements (which corresponded to answers to test questions) after reading the third paper, and completed a filler questionnaire after reading the fourth paper. The order of the four papers used was fully counterbalanced across subjects.

During the second session 3 days later, subjects were tested on all four papers. In this final test, questions alternated between MC and SA formats (i.e., each paper was tested with a total of four MC and four SA questions). This final test was administered in two forms, either odd-numbered

questions given in MC format and even-numbered questions given in SA format or vice versa, counterbalanced across subjects. The facts tested in this final test were identical to those tested in the first session (for conditions in which subjects took an intervening test), although the question format differed for half of the questions.

An additional 24 subjects were tested on the final tests only (i.e., without having read the four papers), to gain a baseline measure and ensure that performance in all our other conditions was above that baseline.

Procedure

Subjects were tested in groups of 10 or fewer during two experimental sessions. During the first session, subjects were seated at computer terminals, and were given paper copies of the papers, one at a time, and asked to read them carefully because they would be tested on the material later. At the outset, subjects were told to expect different types of tests after each paper, although they never knew the nature of the test prior to reading any specific paper. They were also told to feel free to underline or mark any part of a paper during reading. They were given 15 min to read each paper, and a timer on the computer screen counted down the minutes. After the 15 min elapsed, subjects were instructed to put away the paper. In the conditions in which subjects received a test, test questions appeared on the computer screen successively, one at a time, and subjects wrote their answers on response sheets provided. The test was self-paced, and subjects were told not to amend previous responses once they had advanced to the subsequent questions. After subjects completed the test, they proceeded to read the next paper. In the read statements condition, subjects were given a list of eight statements to read at their own pace after the paper, although a 3-min delay was inserted into the computer program, such that subjects could not proceed to the next paper before 3 min had elapsed. In the control condition, subjects completed a filler questionnaire after reading the paper, after which they were dismissed and reminded to return for the second session. The first session lasted about 1 hour and 20 min.

Three days later (a window period of 70–74 hours after the start of the first session was allowed), subjects returned for the second session. They were tested on all the four papers they had previously read, in both MC and SA formats. As before, test questions appeared one at a time on the computer screen, and subjects wrote their answers on response sheets provided. The second session took about 15 min. At the end of the experiment, subjects were debriefed and thanked for their participation.

Results

Scoring

For MC questions, responses were counted as either correct (1 point) or incorrect (0 points). For SA questions, responses were judged as either correct (1 point), partially correct (½ point), or incorrect (0 points). Scoring was done by a single rater. As a reliability check, 10% of the SA responses were submitted to a second rater for scoring. The interrater agreement and reliability were both .94.

Initial test performance

Although the focus is on final test performance, intervening test performance was also examined. For each participant, we computed the proportion of items correctly answered on the intervening MC ($M = 0.86$, $SD = 0.14$) and SA ($M = 0.54$, $SD = 0.23$) tests. The much higher performance on the MC relative to the SA test has implications for their effects on the later tests, as discussed below.

Final test performance

The proportion of questions answered correctly by each participant from each intervening task condition was computed separately for the two final test formats (MC and SA), and the means can be seen in Figure 1. Due to scaling differences between the MC and SA tests, we analysed the final MC and SA performance separately using one-way repeated measures ANOVAs, with intervening task as a within-subjects factor. Baseline performance of subjects who took the tests without having read the papers was .30 and .02 for the MC and SA tests, respectively, far below performance in all four conditions. The α-level for all analyses was set at .05.

Multiple choice. The type of intervening task did affect final MC performance, $F(3, 141) = 5.54$, $MSE = 0.62$, $\eta^2 = .11$. Post hoc comparisons using paired samples *t*-tests revealed that the intervening MC test condition had greater final performance than the intervening SA test and the control conditions, $t(47) = 2.14$, $d = 0.36$, and $t(47) = 3.08$, $d = 0.61$, respectively. Similarly, the read statements condition had greater final performance than the intervening SA test and the control conditions, $t(47) = 2.05$, $d = 0.40$, and $t(47) = 3.64$, $d = 0.66$, respectively. No other pairwise comparison was significantly different.

Short answer. The type of intervening task did affect final SA performance, $F(3, 141) = 10.85$, $MSE = 4.06$, $\eta^2 = .19$. Post hoc comparisons

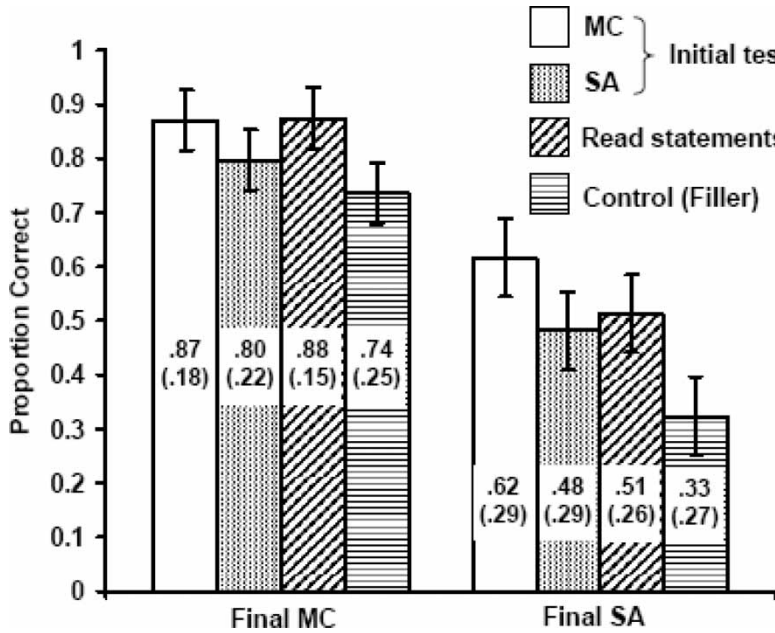

Figure 1. Mean final test performance as a function of intervening task in Experiment 1. Error bars are 95% confidence intervals. *M*s and *SD*s for each condition are listed in the respective bars.

using paired samples *t*-tests indicated that the intervening MC test condition had greater final performance than the intervening SA test and the control conditions, $t(47) = 2.38$, $d = 0.45$, and $t(47) = 6.01$, $d = 1.03$, respectively, and marginally higher final performance than the read statements condition, $t(47) = 1.91$, $p = .06$, $d = 0.37$. The intervening SA test and read statements conditions both had greater final performance than the control condition, $t(47) = 2.91$, $d = 0.55$, and $t(47) = 4.34$, $d = 0.71$, respectively, but were not significantly different from each other.

Lures on MC test. Prior research has shown that MC tests may cause interference when subjects select lure items and hence acquire false knowledge (Roediger & Marsh, 2005). A supplementary analysis was done to ascertain the proportion of instances in which wrong responses on the intervening MC test led to the same incorrect response being endorsed or produced on the final test. If a subject got an item wrong on the intervening MC test, the probability of endorsing or producing the same incorrect answer was .65 and .11 on the final MC and SA tests, respectively.

Effectiveness score

Due to the different levels of performance on the initial MC and SA tests, we analysed test performance using a metric introduced by Lockhart (1975) and extended by Bjork, Hofacker, and Burns (1981) to assess the degree to which subsequent retrieval is enhanced by prior retrieval, while avoiding item selection artifacts that can compromise the interpretation of raw conditional probabilities. The effectiveness score, *a*, is based on a simple finite state model that classifies target items into one of four states: whether or not an item is retrievable at the initial test (*C* or *N*, respectively) and whether or not it is retrievable at the final test (*C* or *N*, respectively). An item in state *CN* is one that would be retrieved on the initial test (if there was an initial test), but would not be retrieved on the final test. Only items in this state can potentially be facilitated by an act of initial retrieval. The *a* score is an estimate of the probability that items in state *CN* make a transition into state *CC* as a result of an initial retrieval being permitted to occur, and the formula is

$$a = \frac{P_T - P_{NT}}{P_{CN} + P_T - P_{NT}},$$

where P_T = final performance for the conditions in which an intervening test was given, P_{NT} = final performance for the no intervening test control condition, P_{CN} = the expected proportion of items in state *CN* (i.e., the conditional probability of getting items wrong on the final test, given correct responding on the intervening test). A more in-depth explanation of how the equation was derived can be found in McDaniel, Kowitz, and Dunay (1989).

For the purpose of this analysis, we rescored the responses to SA questions such that responses that were previously scored either as partially or fully correct were now considered correct or retrievable (i.e., *C*). The *a* scores could not be calculated in some cases due either to the denominator of the equation being zero, or the conditional probability P_{CN} being undefined (see Table 1 for the means; the number of cases each mean is based on is given in parentheses beside each mean). There was no difference in *a* scores for either the MC, $F(1, 24) = 0.345$, or SA, $F(1, 34) = 0.01$, final test as a function of intervening test format. This analysis suggests that successful retrieval on an intervening MC test was as effective as successful retrieval on an intervening SA test in enhancing subsequent retrieval, for both final test formats. It should be noted that this conclusion is, at best, tentative due to the rather large number of missing values.

TABLE 1
Mean *a* scores as a function of format of intervening and final test and feedback

Test format		a *score*		
Intervening test	*Final test*	*Experiment 1–no feedback (no fb)*	*Experiment 2–feedback (fb)*	*Difference (fb–no fb)*
MC	MC	.86 ($n = 35$)	.64 ($n = 36$)	−.22
MC	SA	.38 ($n = 45$)	.32 ($n = 47$)	−.06
SA	MC	.79 ($n = 30$)	.91 ($n = 36$)	.12
SA	SA	.41 ($n = 38$)	.67 ($n = 40$)	.26

The *a* scores could not be calculated in some cases due either to the denominator of the equation being zero, or the conditional probability P_{CN} being undefined. The number of cases that each mean *a* score is based on is given in parentheses. Total $n = 48$.

Discussion

The results in Figure 1 showed that taking an intervening MC test boosted final test performance more than taking an intervening SA test, regardless of whether the final test was MC or SA format. Compared to the control condition, having an intervening MC test resulted in a robust testing effect. Having an intervening SA test led to significantly better performance than the control condition only when the final test was SA format. Different types of tests provide differentially effective cues for accessibility of memories (Tulving & Pearlstone, 1966). Having an intervening SA test may have increased the accessibility to target items, leading to better performance on a subsequent SA test (relative to the control condition), but this effect may not have been apparent on a subsequent MC test because accessibility is already high on a recognition test where copy cues are provided (Darley & Murdock, 1971; Hogan & Kintsch, 1971). The intervening MC test condition was not significantly different from the read statements condition in the final test performance, thus suggesting that the enhancement in retention due to prior MC testing may be equivalent to a focused restudying of the target facts.

One issue that clouds interpretation of the results in Experiment 1 is the differing levels of performance on the initial tests. MC performance (86%) was much higher than SA performance (54%), so it may be no surprise that the greater testing effect shown in Figure 1 for MC tests could have been due to this factor. In addition, in the condition in which subjects restudied the critical facts, they were of course exposed to 100% of them, which may be why even MC testing did not show an advantage relative to reading. Of course, the control of having subjects read the critical statements that would be later tested is quite conservative (and unrealistic) in that students would not normally be able to selectively study only facts

on the upcoming test. Reading the entire paper might be a more externally valid rereading control.

Because of these difficulties in comparing MC and SA tests with their varying levels of performance, we used the effectiveness measure (*a* scores). In this analysis, the results showed that SA and MC tests were equally effective in producing gains on the final test when the no test control was used as a baseline. The conclusion differs from that which arises from using the raw scores (Figure 1), but as noted above the analysis should be considered tentative due to large numbers of missing data. Experiment 2 was conducted using the same conditions and design as in Experiment 1, except that feedback was provided after subjects answered the initial test questions so as to equate for exposure to the material on the initial test. Providing feedback in this way should permit a more accurate assessment of how MC and SA testing affects final criterial performance without differences in test performance playing so great a role.

EXPERIMENT 2

The idea that feedback plays an important role in learning is not new (Thorndike, 1913). Kulhavy (1977) proposed that the crucial instructional significance of feedback is to correct erroneous responses during tests. In Experiment 2, subjects were presented with the correct answer after they responded to each question on the intervening tests. The purpose was to ensure that the intervening test conditions would not be penalised by lower exposure to accurate information (relative to the read statements condition), since performance on the initial tests was not at ceiling. Given that supplying the correct answer after subjects respond on an initial test, compared to conditions providing no feedback or feedback that merely states whether or not a response was correct, has been found to greatly augment final retention of verbal material (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005), we expected the inclusion of feedback on the intervening tests to allow the effect of testing to manifest itself more fully.

Method

Subjects

Fifty-five undergraduates from the Washington University Psychology Subject Pool participated in partial fulfilment of course requirements or for $20 cash. Seven subjects either failed to return for the second session or did not fully adhere to instructions, so their data were excluded from the analysis (leaving data from 48 subjects).

Materials and design

These were the same as in Experiment 1, except that subjects received feedback after each response on the intervening tests in the first session.

Procedure

This was identical to Experiment 1, with the exception that after completing each question on the MC and SA tests in the first session subjects would press a key and the correct answer would appear on the computer screen. They were instructed to press the key only after they had finished responding to each question, and not to change their responses after feedback was provided. Subjects viewed the feedback and proceeded to the following questions at their own pace.

Results and discussion

Initial test performance

Although the focus is on final test performance, the performance on the intervening tests was also examined. For each participant, we computed the proportion of items correctly answered on the intervening MC ($M = 0.85$, $SD = 0.15$) and SA ($M = 0.56$, $SD = 0.19$) tests. Again, MC performance was much higher, but because subjects received feedback immediately, we can assume that initial exposure to correct answers was equated.

Final test performance

The proportion of questions answered correctly by each participant from each intervening task condition was computed separately for the two final test formats (MC and SA), and the means can be seen in Figure 2. Again, the final MC and SA performance was analysed separately using one-way repeated measures ANOVAs, with intervening task as a within-subjects factor.

Multiple choice. There was a significant main effect of intervening task, $F(3, 141) = 14.19$, $MSE = 0.60$, $\eta^2 = .23$. Post hoc comparisons using paired samples t-tests indicated that the intervening SA test condition yielded higher final performance than the intervening MC test, read statements, and control conditions, $t(47) = 2.55$, $d = 0.41$; $t(47) = 3.23$, $d = 0.62$; $t(47) = 6.24$, $d = 1.18$, respectively. The intervening MC test and read statements conditions both had greater final performance than the control condition, $t(47) = 4.22$, $d = 0.78$, and $t(47) = 2.80$, $d = 0.61$, respectively, but were not significantly different from each other.

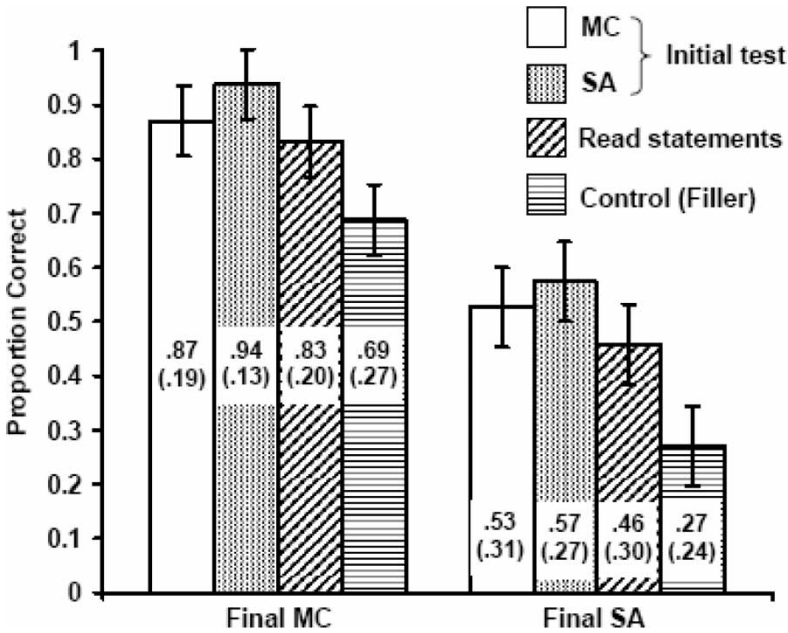

Figure 2. Mean final test performance as a function of intervening task in Experiment 2. Error bars are 95% confidence intervals. *M*s and *SD*s for each condition are listed in the respective bars.

Short answer. There was a significant main effect of intervening task, $F(3, 141) = 12.92$, $MSE = 4.19$, $\eta^2 = .22$. Post hoc comparisons using paired samples *t*-tests revealed that the intervening SA test condition had higher final performance than the read statements and control conditions, $t(47) = 2.18$, $d = 0.40$, and $t(47) = 6.70$, $d = 1.18$, respectively, but performance was not significantly different from the intervening MC test condition. The intervening MC test and read statements conditions both had higher final performance than the control condition, $t(47) = 4.94$, $d = 0.93$, and $t(47) = 3.66$, $d = 0.69$, respectively, but they were not significantly different from each other.

In summary, the results showed that having an intervening SA test enhanced final test performance the most (relative to the read statements or no test control conditions). More importantly, this increase in final retention due to taking an SA test was superior to being presented with statements that reflected answers to test questions (i.e., the read statements condition). Although an intervening MC test did enhance final retention more than the control condition, this effect of testing was not significantly greater than the read statements condition.

The provision of feedback during the intervening tests led to a rather different pattern of results from Experiment 1. The superiority of the

intervening SA test condition on both final test formats (MC and SA), relative to the read statements condition, supports the idea that the more retrieval effort expended during a test, the greater the benefit to retention. Even when retrieval is unsuccessful or erroneous on the intervening test, if corrective feedback is supplied, the benefit to final retention from taking a SA test exceeds that of being presented with the test answers for additional study.

Lures on MC test. Again, a supplementary analysis was done to ascertain the likelihood of a wrong response on the intervening MC test being endorsed or produced on the final test. If a subject got an item wrong on the intervening MC test, the probability of endorsing or producing the same incorrect answer was .30 and .03 on the final MC and SA tests, respectively. Compared to the results from Experiment 1, it is clear that the provision of corrective feedback diminished the negative effect of MC lures (Butler & Roediger, 2006).

Effect of feedback

Final test performance. To examine directly whether the role of test format in enhancing final retention was moderated by the provision of corrective feedback during the intervening tests, we compared the data from Experiments 1 and 2, and analysed final test performance using a mixed ANOVA with intervening and final test formats as within-subjects factors, and feedback as a between-subjects factor. Crucially, intervening test format interacted with feedback, $F(1, 94) = 11.44$, $MSE = 0.05$, partial $\eta^2 = .11$. Post hoc comparisons using independent samples *t*-tests revealed that whereas the provision of feedback during the intervening SA tests led to greater final test performance, $t(190) = 2.75$, $d = 0.40$, feedback during the intervening MC tests did not make a difference to final performance, $t(190) = -1.06$, $p = .29$.

Effectiveness score. An alternative way to look at the role of feedback is to examine its impact on *a*. Again, we compared the data from Experiments 1 and 2, and for each subject, the *a* score was calculated for each intervening (MC and SA) and final (MC and SA) test condition. Since Experiment 2 incorporated corrective feedback during the intervening tests, the *a* scores derived from those data would gauge not only the impact of initial retrieval, but also of feedback as well. As in Experiment 1, the *a* scores could not be calculated in some cases (see Table 1 for the means, with the number of cases contributing to each mean shown in parentheses).

The *a* scores were analysed using a mixed ANOVA with intervening and final test formats as within-subjects factors, and feedback as a between-

subjects factor. An intervening SA test yielded higher *a* scores than an intervening MC test, $F(1, 45) = 7.87$, $MSE = 0.23$, partial $\eta^2 = .15$. But this main effect was qualified by an interaction between intervening test format and feedback, $F(1, 45) = 8.14$, $MSE = 0.23$, partial $\eta^2 = .15$. As can be seen in the last column of Table 1 showing the differences in *a* scores as a function of feedback, while the provision of feedback did not significantly affect *a* scores in the intervening MC condition, $t(161) = -1.10$, $p = .27$, feedback marginally boosted *a* scores in the intervening SA condition, $t(142) = 1.67$, $p < .10$.

Once again, it should be noted that the results of the *a* score analysis are provisional, given the large number of missing values. Nevertheless, the combined analysis of data from Experiments 1 and 2, both in terms of final test performance and *a* scores, suggests that SA tests paired with corrective feedback are especially efficacious for enhancing final retention.

The conditional probabilities of correctly recalling an item on the final test given the accuracy of response at the intervening test also point to the same conclusion (see Table 2, which displays the conditional probabilities of being correct on the final test, broken down by final test format, intervening test format, and accuracy at the intervening test). The effect of feedback was more pronounced on items that were unretrievable at the intervening test, so given the lower performance on the intervening SA test (relative to MC test), more items in that condition would benefit from feedback. For items that were answered wrongly on the intervening MC test, with the provision of feedback the proportion answered correctly on the final MC test increased from .27 (Experiment 1) to .52 (Experiment 2), and for the final SA test the increase was from .11 (Experiment 1) to .19 (Experiment 2). For items that were answered wrongly on the intervening SA test, with the provision of

TABLE 2
Conditional probabilities of getting item correct on the final test as a function of test format and feedback

Final test	*Intervening test format*	*Accuracy at intervening test*	*Experiment 1 (no fb)*	*Experiment 2 (fb)*	*Difference (fb−no fb)*
MC correct	MC	Wrong	.27	.52	.25
		Correct	.96	.93	−.03
	SA	Wrong	.59	.84	.25
		Partially correct	.90	.94	.04
		Correct	.94	1.00	.06
SA correct	MC	Wrong	.11	.19	.08
		Correct	.62	.51	−.11
	SA	Wrong	.06	.27	.21
		Partially correct	.17	.26	.09
		Correct	.76	.76	.00

feedback the proportion answered correctly on the final MC test increased from .59 (Experiment 1) to .84 (Experiment 2), and for the final SA test the increase was from .06 (Experiment 1) to .27 (Experiment 2). In sum, the conditional probability of getting an item correct on the final test given that the response was wrong or omitted at the intervening test increased with the provision of feedback, and this increase for the intervening SA condition was numerically greater than the intervening MC condition.

GENERAL DISCUSSION

The results of the experiments reported here provide further evidence that testing can improve retention of studied material. More specifically, this study showed that test format and corrective feedback do modulate the testing effect. Taking an SA test was found to boost final test performance more than additional focused exposure to test-relevant information, if corrective feedback was provided to ameliorate poor initial test performance. Although taking an MC test improved final retention relative to doing an unrelated filler task after study, performance was not significantly different from receiving additional exposure to test-relevant information. Based on the current results, taking a MC test may seem functionally equivalent to a reexposure to target information. However, such a conclusion may be premature because in the final SA tests, a tendency existed in both experiments for the intervening MC condition to outperform the read statements condition (and this effect was marginally significant in Experiment 1). So it could be that the benefit of taking an intervening MC test only begins to show when retention is assessed in a manner that is sensitive to the accessibility of target items in memory (Hogan & Kintsch, 1971). Also, it is worth noting again that focused reading of target facts is not the norm in the classroom; it is probably more common for students to reread the entire passage or set of materials. Further research is needed to ascertain if and how the effect of taking an MC test differs from simply reading the material again, but both produced gains relative to the control condition that received no additional exposure to the material.

The pattern of results obtained was not predicted by the transfer appropriate processing framework. A match in the formats of the intervening and final tests did not result in the best final performance. Perhaps a way to reconcile our findings with such a framework is to think of MC and SA tests as not so much engaging fundamentally different processes; rather, MC and SA tests possibly differ in terms of the degree or amount of particular memory processes that are required. In terms of Jacoby's (1991) dual-process model, it could be that SA tests rely more on the intentional, recollective component than MC tests, and that this deeper engagement of recollection

during an intervening SA test results in better retention. MC tests, on the other hand, may rely relatively more on familiarity, so an intervening MC test may have less positive transfer to a final SA test (as predicted by the transfer appropriate processing framework). In addition, performance on a final MC test is compromised by the increased familiarity of the incorrect lures seen on the previous MC test (Roediger & Marsh, 2005). Therefore, expecting an intervening MC test to transfer most to a final MC test and an intervening SA test to transfer most to a final SA test may be too simplistic an application of the transfer appropriate processing framework.

Our findings contradict those of Duchastel and Nungester (1982), who found that taking an initial MC test benefited a final MC test more than taking an initial SA test, and taking an initial SA test benefited a final SA test more than taking an initial MC test (although this latter comparison did not reach statistical significance). Possible factors that could have contributed to the discrepancy in findings include differences in the amount of information required for each test question (i.e., their test questions required a brief—often one-word—answer, whereas our test questions often required more elaborate sentence-length answers), differences in retention interval (i.e., their final test was administered after a 2-week delay, whereas our final test was 3 days after the study phase), and the provision of feedback (i.e., they did not provide feedback during the initial test, whereas we did so in Experiment 2). The level of performance on the initial tests could also have played a role; unfortunately, Duchastel and Nungester did not report their initial test results. Also, Duchastel and Nungester conducted their study on students in actual, intact classes, and a questionnaire they administered at the end of the study revealed that 29% of the students discussed the studied passage with their friends during the 2-week retention interval. This could have introduced additional confounds beyond the effect of the initial test.

Other sources of evidence that congruence between the intervening and final tests plays a role in the testing effect come from studies that have shown that when the test questions in the intervening and final tests are phrased similarly, final performance is better than when the questions are paraphrased (Anderson & Biddle, 1975; McDaniel & Fisher, 1991). Also, McDaniel et al. (1989) varied the type of the cues (phonemic or semantic) in a cued-recall test, and found that final recall performance was best when the cues on the final test were of the same type as the cues on the initial test. Future research will be needed to determine the circumstances in which the overlap in the processes engaged by the initial and final tests contributes to the testing effect.

The effect of retrieval

The superiority of taking an intervening SA test with feedback over an MC test on final retention, regardless of the final test format, strongly implicates the role of retrieval effort in improving retention. The findings suggest that the greater the depth or difficulty of the retrieval attempt, the greater the benefit to retention, in a way replicating Glover (1989, Exp. 4) and Carpenter and DeLosh (2006, Exp. 1). Glover and Carpenter and DeLosh did not provide corrective feedback during the initial test or retrieval, but yet found that the most demanding intervening retrieval condition produced the best final performance, whereas in our study the SA test was most effective only when feedback was provided (Experiment 2). The likely explanation is that initial retrieval levels were lower in our study than in theirs, and hence it was necessary for corrective feedback to restore the effectiveness of testing, which is reduced when performance at initial retrieval is low (Wenger et al., 1980).

A recent study by McDaniel, Anderson, Derbish, and Morrisette (2007, this issue), conducted on students enrolled in an online college course, found that on criterial MC exams, those who took weekly SA quizzes outperformed those who took weekly MC quizzes or read target facts (feedback was provided during the weekly quizzes). Their findings dovetail nicely with the results of Experiment 2, and provide additional support for the idea that recall tests are more beneficial for subsequent retention than recognition tests or additional study. Bjork (1975) suggested two possible ways by which retrieval attempts result in the testing effect: The more effortful or difficult the retrieval, (a) the more the memory trace or representation is strengthened, becoming more durable and less vulnerable to interference, and (b) the more the retrieval routes are elaborated and multiplied, increasing accessibility for subsequent retrieval. The findings of the present study are consistent with such an account.

The effect of SA testing bears noticeable similarity to the generation effect. It has been amply demonstrated that when target items are self-generated by subjects in response to cues provided by the experimenter, those items are better retained than items merely presented to be read (Slamecka & Graf, 1978). Importantly, Slamecka and Fevreiski (1983) found that the generation effect is obtained even when subjects fail to correctly generate an item, if the correct response is presented after the failed generation attempt. This, of course, does not mean that the causal mechanisms behind the generation effect and the SA testing advantage obtained in our study are identical. Investigations into the generation effect have shown that factors other than retrieval can contribute to the effect (e.g., allocation of attentional resources at encoding; Schmidt, 1990).

The effect of feedback

The advantage of an intervening SA test only became evident in our study when corrective feedback was provided at test, thus revealing feedback as an important moderator of the effects of testing, especially when performance levels are relatively low during the intervening test. These findings are consistent with those of Pashler et al. (2005), who found that in paired associate learning, supplying the correct answer after an incorrect response during an initial cued recall test produced a great boost to final retention 1 week later. Our analysis of the combined data from Experiments 1 and 2 suggests that the effect of providing feedback on final performance depends on the format of the intervening test. Whereas feedback during an intervening SA test significantly benefited final performance, feedback during an intervening MC test seemed not to make a difference.

One possible reason for the interaction between provision of feedback and intervening test format is that intervening SA test performance was lower (relative to intervening MC performance), hence there were more opportunities for feedback to correct errors and improve final performance. Another possibility is that the retrieval processes engendered by the type of intervening test affected the subsequent processing of feedback, such that taking a SA test, with relatively greater retrieval demands, led to more thorough encoding of feedback than taking an MC test. Although this is admittedly speculative based on the current data, there is evidence that feedback processing can be influenced by metacognitive knowledge (e.g., subjective certainty of whether or not a response is correct) and the accuracy of the response (Kulhavy & Stock, 1989). For instance, Bahrick and Hall (2005) proposed that retrieval failures are informative to subjects, and can spur them to modify their encoding strategies during subsequent presentations of the target material. In addition, Butterfield and Metcalfe (2006) found that subjects focus more attention on corrective feedback when they make an erroneous response with high confidence. Also, Auble and Franks (1978) showed that the more subjects puzzled over seemingly incomprehensible sentences (e.g., "The party stalled because the wire straightened") before they were given a key word that rendered the sentences comprehensible (*cockscrew*), the better their retention of the key word on a later test. This benefit of "effort after meaning" resembles the advantage of SA testing with feedback over just reading the test answers (i.e., the read statements condition) on final performance. To better elucidate the role of feedback, future research should directly examine how the format of a memory test can affect metacognitive judgements and processing of corrective feedback. Nevertheless, regardless of the specific mechanisms involved, the present findings have obvious practical implications.

Practical implications for pedagogy

In recent years, there has been increasing appreciation for the need to base educational practice on scientific evidence (US Department of Education, 2003). Unfortunately, relevant empirical findings from basic psychological research are often not disseminated widely in teacher education programmes (Newcombe, 2002). Although psychological science has had some impact on the teaching of reading (Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001), pertinent research from other domains of cognitive psychology (e.g., human learning and memory) has not had much influence on actual pedagogical practice (Matlin, 2002). Aside from the typical barriers when trying to communicate across disciplines, part of the problem might be researchers failing to see the potential application of their findings, and educators feeling sceptical about whether laboratory findings generalise to the classroom.

A major impetus for this study was the desire to examine a question about testing that would have straightforward implications for teaching practice, using materials and tests that closely approximate college-level course materials. So as not to be bogged down by the current controversy over high-stakes achievement testing, a change in mindset away from assessing achievement would be useful: Administering tests can be a valuable pedagogical tool to enhance learning (Demptser, 1992; Roediger & Karpicke, 2006b). Our findings clearly support this view by showing that of the two test formats often used by teachers—MC and SA—an SA test is more beneficial for long-term retention than restudying. Also, to obtain the benefit of an SA test, it is necessary to provide corrective feedback, especially when performance on the test is not high. This is a point worth noting, especially since educators at the higher levels (e.g., in college in North America) often do not provide corrective feedback after tests, or at least make it inconvenient for students to view the feedback (e.g., require students to make an appointment to view their test forms), so as to save class time or maintain security of questions in a test bank. The timing of the feedback may also be a factor (Butler & Roediger, 2006; Kulik & Kulik, 1988). Whereas feedback in our study (Experiment 2) was provided after each item, in classrooms it is probably more common for feedback to be delayed for at least a few days until the test scripts have been scored. Future research is required to ascertain whether the current findings generalise to a situation where feedback is delayed. Although MC tests are easier to score and hence more convenient to administer, they tend not to be as effective in improving retention as SA tests. Moreover, a disadvantage of MC or recognition tests is that the presence of (incorrect) lures has the potential to create false knowledge; students may subsequently accept these lures as correct (Butler, Marsh, Goode, & Roediger, 2006; Mandler & Rabinowitz,

1981; Roediger & Marsh, 2005), although feedback does ameliorate the negative effect of MC testing (Butler & Roediger, 2006). Carroll, Campbell-Ratcliffe, Murnane, and Perfect (2007, this issue) demonstrated that retrieval practice during an initial cued recall test could, in some situations, impair memory for untested material on a final test (but see Chan, McDermott, & Roediger, 2006). It should be noted that for conditions in which the effect was present, it was short-lived. Items that underwent retrieval practice, on the other hand, displayed relatively larger and more enduring facilitation on the final test. In addition to the memorial advantages, regular testing has other benefits. It encourages preparation for class, reduces test anxiety, and focuses attention on important course content (Snooks, 2004).

In conclusion, educators have at their disposal a readily available instrument for enhancing learning and retention—SA tests. Instead of giving students handouts summarising key points and facts, a better alternative would be the regular administration of SA quizzes, followed by instructor feedback.

REFERENCES

Anastasio, P. A., Rose, K. C., & Chapman, J. (1999). Can the media create public opinion? A social-identity approach. *Current Directions in Psychological Science*, *8*, 152–155.

Anderson, R. C., & Biddle, W. B. (1975). On asking people questions about what they are reading. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 9 (pp. 89–132)). New York: Academic Press.

Anderson, R. S. (1998). Why talk about different ways to grade? The shift from traditional assessment to alternative assessment. *New Directions for Teaching and Learning*, *74*, 5–16.

Auble, P. M., & Franks, J. J. (1978). The effects of effort toward comprehension on recall. *Memory and Cognition*, *6*, 20–25.

Bahrick, H. A., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*, 566–577.

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

Bjork, R. A., Hofacker, C., & Burns, M. J. (1981, November). *An "effectiveness-ratio" measure of tests as learning events.* Paper presented at the 22nd annual meeting of the Psychonomic Society, Philadelphia, PA.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 3–9.

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

Butler, A. C., & Roediger, H. L. (2006, May). *Feedback neutralizes the detrimental effects of multiple-choice testing*. Poster presentation at the 18th annual convention of the Association for Psychological Science, New York.

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

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

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

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology*, *19*, 580–606.

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

Darley, C. F., & Murdock, B. B., Jr. (1971). Effects of prior free recall testing on final recall and recognition. *Journal of Experimental Psychology*, *91*, 66–73.

Dempster, F. N. (1992). Using tests to promote learning: A neglected classroom resource. *Journal of Research and Development in Education*, *25*, 213–217.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego, CA: Academic Press.

Dempster, F. N., & Perkins, P. G. (1993). Revitalizing classroom assessment: Using tests to promote learning. *Journal of Instructional Psychology*, *20*, 197–203.

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

Eagly, A. H., Kulesa, P., Chen, S., & Chaiken, S. (2001). Do attitudes affect memory? Tests of the congeniality hypothesis. *Current Directions in Psychological Science*, *10*, 5–9.

Garry, M., & Polaschek, D. L. L. (2000). Imagination and memory. *Current Directions in Psychological Science*, *9*, 6–10.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, No. 40, 1–104.

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

Halpern, D. F., & Hakel, M. D. (2002). Learning that lasts a lifetime: Teaching for long-term retention and transfer. *New Directions for Teaching and Learning*, *89*, 3–7.

Hanawalt, N. G., & Tarr, A. G. (1961). The effect of recall upon recognition. *Journal of Experimental Psychology*, *62*, 361–367.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 562–567.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*, 513–541.

Jones, H. E. (1923). Experimental studies of college teaching: The effect of examination on permanence of learning. *Archives of Psychology*, *No. 68*, 5–70.

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

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.-L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*, 79–97.

Kuo, T. M., & Hirshman, E. (1996). Investigations of the testing effect. *American Journal of Psychology*, *109*, 451–464.

Lachman, R., & Laughery, K. R. (1968). Is a test trial a training trial in free recall learning? *Journal of Experimental Psychology*, *76*, 40–50.

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, *67*, 259–266.

Lockhart, R. S. (1975). The facilitation of recognition by recall. *Journal of Verbal Learning and Verbal Behavior*, *14*, 253–258.

Mandler, G., & Rabinowitz, J. C. (1981). Appearance and reality: Does a recognition test really improve subsequent recall and recognition? *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 79–90.

Matlin, M. W. (2002). Cognitive psychology and college-level pedagogy: Two siblings that rarely communicate. *New Directions for Teaching and Learning*, *89*, 87–103.

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). Tests and test feedback as learning sources. *Contemporary Educational Psychology*, *16*, 192–201.

McDaniel, M. A., Kowitz, M. D., & Dunay, P. K. (1989). Altering memory through recall: The effects of cue-guided retrieval processing. *Memory and Cognition*, *17*, 423–434.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Newcombe, N. S. (2002). Biology is to medicine as psychology is to education: True or false? *New Directions for Teaching and Learning*, *89*, 9–18.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*, 18–22.

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

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, *2*, 31–74.

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

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

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory and Cognition*, *11*, 641–650.

Runquist, W. N. (1986). The effect of testing on the forgetting of related and unrelated associates. *Canadian Journal of Psychology*, *40*, 65–76.

Runquist, W. N. (1987). Retrieval specificity and the attenuation of forgetting by testing. *Canadian Journal of Psychology*, *41*, 84–90.

Schmidt, S. R. (1990). A test of resource-allocation explanations of the generation effect. *Bulletin of the Psychonomic Society*, *28*, 93–96.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior*, *22*, 153–163.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604.

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 716–727.

Snooks, M. K. (2004). Using practice tests on a regular basis to improve student learning. *New Directions for Teaching and Learning*, *100*, 109–113.

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

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 210–221.

Thorndike, E. L. (1913). *Educational psychology: Vol. 1. The original nature of man*. New York: Columbia University.

Treiman, R. (2000). The foundations of literacy. *Current Directions in Psychological Science*, *9*, 89–92.

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

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior*, *5*, 381–391.

US Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Washington, DC: Author.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 135–144.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.

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.

APPENDIX: Test questions used in Experiments 1 and 2

From Garry and Polaschek (2000)

MC	*SA*	*Read Statements*
1. Loftus (1993) was the first systematic study to show what? a. Detailed false memories for a whole event could be implanted b. Emotional events tend to be particularly salient and memorable c. Counterfactual thoughts can affect people's judgment of outcomes d. People tend to misremember childhood events	l. Loftus (1993) was the first systematic study to show what?	l. Loftus (1993) was the first systematic study to show that detailed false memories for a whole event could be implanted.
2. According to Sarbin (1998), what strategy do people rely on when they try to remember an event that they do not remember? a. Fabrication of the details b. Exhaustive search of their memory store c. Imagination of the event d. Look for retrieval cues in the environment	2. According to Sarbin (1998), what strategy do people rely on when they try to remember an event that they do not remember?	2. According to Sarbin (1998), people rely on imagination as a strategy when trying to remember an event they do not remember.

Appendix (*Continued*)

MC	*SA*	*Read Statements*
3. Subjects become more confident they have experienced a counterfactual event after they imagine the event. This is called a. a false memory b. imagination inflation c. illusory vividness d. confirmatory bias	3. Subjects become more confident they have experienced a counterfactual event after they imagine the event. This is called _____.	3. Subjects become more confident they have experienced a counterfactual event after they imagine the event. This is called imagination inflation.
4. What is source confusion? a. Confusing details from an imagined event with details from an experienced event. b. Forgetting what the source of a memory was. c. When the vividness of a memory is no longer a good indicator of its veracity. d. Misattributing content of a memory to the wrong source.	4.What is source confusion?	4. Source confusion occurs when one misattributes the content of a memory to the wrong source.
5. Two mechanisms have been proposed to account for the boost in confidence of having experienced an imagined counterfactual event. One is source confusion, the other is a. strength of memory trace. b. recollection. c. vividness of memory. d. familiarity	5. Two mechanisms have been proposed to account for the boost in confidence of having experienced an imagined counterfactual event. One is source confusion, the other is_____.	5. Two mechanisms have been proposed to account for the boost in confidence of having experienced an imagined counterfactual event. One is source confusion, the other is familiarity.
6. According to Heaps and Nash (1999), which of the following factors predicts people's tendency to become more confident that they have actually experienced an event after imagining the event? a. Their susceptibility to influence of an authoritative person b. The vividness of their mental imagery c. Their predisposition to hypnotic suggestion d. Their arousal to emotional stimuli	6. State a factor that predicts people's tendency to become more confident that they have actually experienced an event after imagining the event (according to Heaps & Nash, 1999).	6. According to Heaps and Nash (1999), a person's predisposition to hypnotic suggestion predicts one's tendency to become more confident that one has actually experienced an event after imagining the event.
7. One might be tempted to regard the confidence boosting effect of imagining an event as merely the statistical phenomenon of a. regression towards the mean b. restriction of range c. homogeneity of regression d. a spurious correlation	7. One might be tempted to regard the confidence boosting effect of imagining an event as merely the statistical phenomenon of _____.	7. One might be tempted to regard the confidence boosting effect of imagining an event as merely the statistical phenomenon of regression towards the mean.
8. Why do the findings of memory-related effects of repeatedly imagination have clinical implications? a. Because patients might be imagining their disorder/illness b. Because various psychotherapy techniques involve imagining situations and actions c. Because the therapist may find it difficult to distinguish reality from imagination d. Because repeated imagination of events can lead to hallucinations	8. Why do the findings of memory-related effects of repeatedly imagination have clinical implications?	8. The findings of memory-related effects of imagination have clinical implications because various psychotherapy techniques involve imagining situations and actions.

From Anastasio et al. (1999)

MC	*SA*	*Read Statements*
1. What is one of the most blatant examples of how the media can induce public opinion? a. Biased news coverage. b. "Live" telecast of events. c. Advertisements. d. Selective censorship of news stories.	1. What is one of the most blatant examples of how the media can induce public opinion?	1. One of the most blatant examples of how the media can induce public opinion is via advertisements.
2. What difference did Archer et al. (1983) find in the way men and women are typically portrayed in news photographs? a. Men are often pictured in job-related roles, whereas women feature more prominently in home-related roles. b. Photographs of men tend to be more close-up compared to that of women. c. The facial expressions of men in photographs tend to be more solemn than that of women. d. Men tend to be photographed alone, whereas photos of women tend to feature them in a group.	2. What difference did Archer et al. (1983) find in the way men and women are typically portrayed in news photographs?	2. Archer et al. (1983) found that in news photographs, men tended to be portrayed more close-up than women.
3. Persons depicted in photographs high in "face-ism" tend to be rated as more _____. a. friendly b. confident c. trustworthy d. intelligent	3. Persons depicted in photographs high in "face-ism" tend to be rated as more _____.	3. Persons depicted in photographs high in "face-ism" tend to be rated as more intelligent.
4. According to Mullen et al. (1986), how was newscaster Peter Jennings different when discussing Ronald Reagan's 1984 campaign compared to when he discussed the campaign of Reagan's political opponent? a. Peter Jennings smiled more when discussing Reagan. b. Peter Jennings was more critical of Reagan. c. Peter Jennings used more hand gestures when discussing Reagan. d. Peter Jennings looked directly at the camera more often when discussing Reagan.	4. According to Mullen et al. (1986), how was newscaster Peter Jennings different when discussing Ronald Reagan's 1984 campaign compared to when he discussed the campaign of Reagan's political opponent?	4. According to Mullen et al. (1986), newscaster Peter Jennings smiled more when discussing Ronald Reagan's 1984 campaign compared to when he discussed the campaign of Reagan's political opponent.
5. According to Gilens (1996), how can the media's portrayal of America's poor affect public perception of poverty? a. The media's overrepresentation of African Americans in poverty can create the perception of more blacks in poverty than there actually are. b. The media's portrayal of poor people as lacking in motivation can lead to less public support for social welfare and public assistance. c. The media's portrayal of people in poverty as being lazy can increase negative attitudes toward poor people. d. The media's underrepresentation of certain groups in their portrayal of poverty can lead to those groups being neglected in social welfare policies.	5. According to Gilens (1996), how can the media's portrayal of America's poor affect public perception of poverty?	5. According to Gilens (1996), the media's overrepresentation of African Americans in poverty can create the perception of more blacks in poverty than there actually are.

Appendix (*Continued*)

MC	*SA*	*Read Statements*
6. What is the *hostile media bias*? a. The subtle effects of media portrayal on people's perceptions and opinion. b. The media's influence on hostile and aggressive behaviour. c. The media's reinforcement of negative stereotypes of out-group. d. People on both sides of a controversy perceiving the media as hostile to their group.	6. What is the *hostile madia bias*?	6. *Hostile media bias* refers to the phenomenon where people on both sides of a controversy perceive the media as hostile to their group.
7. Advertising that uses an attractive person to promote a product is relying on the _____ route of persuasion. a. central b. secondary c. peripheral d. fundamental	7. Advertising that uses an attractive person to promote a product is relying on the _____ route of persuasion.	7. Advertising that uses an attractive person to promote a product is relying on the peripheral route of persuasion.
8. A study conducted by the authors (which involved subjects judging guilt/innocence of a fraternity member on charges of vandalism) found that the subject's tendency to side with one's in-group disappeared when_____. a. the subject was exposed to the opinion of an authority figure b. the subject was exposed to evenly mixed opinions of in-group and out-group members c. opinions of others were homogeneous and perfectly correlated with group membership d. the subject was given time to consider all the evidence	8. A study conducted by the authors (which involved subjects judging guilt/innocence of a fraternity member on charges of vandalism) found that the subject's tendency to side with one's in-group disappeared when _____.	8. A study conducted by the authors (which involved subjects judging guilt/innocence of a fraternity member on charges of vandalism) found that the subject's tendency to side with one's in-group disappeared when the subject was exposed to evenly mixed opinions of in-group and out-group members.

From Treiman (2000)

MC	*SA*	*Read Statements*
1. What is the alphabetic principle? a. appreciating that many languages, in the written form, use a set of symbols or letters b. appreciating how the letters in printed words relate to how the spoken words sound c. appreciating that there are some rules in how letters can be combined in the spelling of words d. appreciating how the spelling of words can be inconsistent	1. What is the alphabetic principle?	1. The alphabetic principle refers to the appreciation of how the letters in printed words relate to how the spoken words sound.
2. What is a phoneme? a. basic sound unit of a language b. the sound structure of a language c. a syllable d. a cluster of consonants	2. What is a phoneme?	2. A phoneme is the basic sound unit of a language.
3. A syllable can be subdivided into a. consonant clusters b. vowel clusters c. letter segments d. onset and rime	3. A syllable can be subdivided into 2 parts: _____ & _____	3. A syllable can be subdivided into 2 parts: onset and rime.

Appendix (*Continued*)

MC	*SA*	*Read Statements*
4. Studies have shown that training in _____ can improve reading and spelling ability in children. a. the names of letters b. analyzing linguistic structure c. phonological awareness d. how to spell their names	4. Studies have shown that training in _____ can improve reading and spelling ability in children.	4. Studies have shown that training in phonological awareness can improve reading and spelling ability in children.
5. What has been the implicit assumption about how children learn letter names and letter sounds? a. they learn them via imitating adult speech b. they learn them unconsciously when listening to adults speak c. they learn them via experimentation with different sounds d. they learn them via rote memorization	5. What has been the implicit assumption about how children learn letter names and letter sounds?	5. The implicit assumption has been that children learn letter names and letter sounds via rote memorization.
6. Recent studies by Treiman et al. have found that an important determinant of knowledge of letter-sounds is a. whether the letter's sound occurs in the name of the letter b. whether the letter is voiced or unvoiced c. the place of articulation of the sound d. the spelling of the child's name	6. Recent studies by Treiman et al. have found that an important determinant of knowledge of letter-sounds is _____.	6. Recent studies by Treiman et al. have found that an important determinant of knowledge of letter-sounds is whether the letter's sound occurs in the name of the letter.
7. Young Joe is more likely to know the _____ of the letter 'j' than Alice or Tom. a. place of articulation b. phoneme c. name d. sound	7. Young Joe is more likely to know the _____ of the letter 'j' than Alice or Tom.	7. Young Joe is more likely to know the name of the letter 'j' than Alice or Tom.
8. There is a widespread view that young children are purely _____ readers, memorizing associations between whole printed words and their spoken form	8. There is a widespread view that young children are purely _____ readers, memorizing associations between whole printed words and their spoken form.	8. There is a widespread view that young children are purely logographic readers, memorizing associations between whole printed words and their spoken form.

From Eagly et al. (2001)

MC	*SA*	*Read Statements*
1. What is the congeniality hypothesis? a. People are motivated to avoid information that challenges their attitudes. b. People's memories are biased in favor of information that agrees with their attitudes. c. People selectively pay attention only to attitudinally agreeable information. d. People tend to more elaborately process information that is inconsistent with their attitudes.	1. What is the congeniality hypothesis?	1. The congeniality hypothesis proposes that people's memories are biased in favor of information that agrees with their attidues.

Appendix (*Continued*)

MC	*SA*	*Read Statements*
2. In the author's meta-analysis of research on memory for attitude-relevant information, what was the trend in results for later compared to early findings? a. Early experiments showed that congenial information was less memorable than uncongenial information, whereas later research tended to yield the reverse pattern or null difference. b. The results tended to be inconsistent regardless of whether the studies were early or more recent. c. More recent studies tended to yield a larger effect of congeniality on memory than early studies. d. Early experiments showed that congenial information was more memorable than uncongenial information, whereas later research tended to yield the reverse pattern or null difference.	2. In the auhors' meta-analysis of research on memory for attitude-relevant information, what was the trend in results for later compared to early findings?	2. In the auhors' meta-analysis of research on memory for attitude-relevant information, they found early experiments tended to show that congenial information was more memorable than uncongenial information, whereas later research tended to yield the reverse pattern or null difference.
3. What is the likely cause of the different trend in findings for earlier vs. later studies? a. Improvements in the procedures used to assess memory. b. Participants in later studies tended to have less polarized attitudes. c. Participants in earlier studies tended to have weaker attitudes. d. Later studies examined more variables than earlier studies.	3. What is the likely cause of the different trend in findings for earlier vs. later studies?	3. Improvements in the procedures used to assess memory is the likely cause of the different trend in findings for earlier vs. later studies.
4. What is the design/procedure for a typical experiment looking at the congeniality effect? a. Participants attitudes toward an issue are measured before and after presentation of information relevant to the issue. b. Participants are presented with information that disagrees with their attitudes, and their subsequent memory for that information assessed. c. Participants with opposing attitudes toward an issue are presented with information on one or both sides of the issue, and their subsequent memory for that information assessed. d. Participants are presented with information that agrees with their attitudes, and their subsequent memory for that information assessed.	4. What is the design/procedure for a typical experiment looking at the congeniality effect?	4. An experiment looking at the congeniality effect typically has the following design/procedure: Participants with opposing attitudes toward an issue are presented with information on one or both sides of the issue, and their subsequent memory for that information assessed.
5. What did the authors propose could account for the weakness of the congeniality effect shown in experiments that were methodologically more rigorous? a. Attitudes have little impact on memory. b. People avoid information that challenges their attitudes. c. People may mount an active defense and hence thoroughly process counterattitudinal information. d. Participants had insufficiently strong attitudes.	5. What did the authors propose could account for the weakness of the congeniality effect shown in experiments that were methodologically more rigorous?	5. The authors proposed that people may mount an active defense and hence thoroughly process counterattitudinal information, thus accounting for the weakness of the congeniality effect found in experiments that were methodologically more rigorous.

Appendix (*Continued*)

MC	*SA*	*Read Statements*
6. In their recent experiment (Eagly et al., 2000), the authors found what difference between congenial and uncongenial information? a. Congenial information was recalled better than uncongenial information. b. Participants had more prior knowledge of congenial information than uncongenial information. c. Uncongenial information was better recalled soon after the message was presented, whereas congenial information was better recalled after a delay. d. Uncongenial information elicited more thought and attention than congenial information.	6. In their recent experiment (Eagly et al., 2000), the authors found what difference between congenial and uncongenial information?	6. In their recent experiment (Eagly et al., 2000), the authors found that uncongenial information elicited more thought and attention than congenial information.
7. The authors propose that to persuade people to accept a position that is highly divergent from their own attitudes, it might be best to a. use an incremental approach whereby each exposure to uncongenial information produces only a small amount of change. b. expose them to large amounts of uncongenial information at one go. c. employ an authority figure to promote the counterattitudinal position. d. encourage them to think global thoughts concerning the issue rather than differentiated thoughts.	7. The authors propose that to persuade people to accept a position that is highly divergent from their own attitudes, it might be best to _____.	7. The authors propose that to persuade people to accept a position that is highly divergent from their own attitudes, it might be best to use an incremental approach whereby each exposure to uncongenial information produces only a small amount of change.
8. According to dual-process theories of social judgement, a recipient who is lacking in motivation and capacity will likely adopt what type of approach when faced with uncongenial information? a. Yield and capitulate to the counterattitudinal viewpoint. b. Adopt an active resistance and confront the uncongenial information. c. Adopt a passive, avoidant approach and process the information less. d. React emotionally and dismiss the information outright.	8. According to dual-process theories of social judgement, a recipient who is lacking in motivation and capacity will likely adopt what type of approach when faced with uncongenial information?	8. According to dual-process theories if social judgment, a recipient who is lacking in motivation and capacity will likely adopt a passive, avoidant approach and process the information less.