1 Benefits of testing memory

Best practices and boundary conditions

Henry L. Roediger, III,
Pooja K. Agarwal, Sean H. K. Kang
and Elizabeth J. Marsh

The idea of a memory test or of a test of academic achievement is often circumscribed. Tests within the classroom are recognized as important for the assignment of grades, and tests given for academic assessment or achievement have increasingly come to determine the course of children's lives: score well on such tests and you advance, are placed in more challenging classes, and attend better schools. Against this widely acknowledged backdrop of the importance of testing in educational life (not just in the US, but all over the world), it would be difficult to justify the claim that testing is not used enough in educational practice. In fact, such a claim may seem to be ludicrous on the face of it. However, this is just the claim we will make in this chapter: Education in schools would greatly benefit from additional testing, and the need for increased testing probably increases with advancement in the educational system. In addition, students should use self-testing as a study strategy in preparing for their classes.

Now, having begun with an inflammatory claim – we need more testing in education – let us explain what we mean and back up our claims. First, we are not recommending increased use of standardized tests in education, which is usually what people think of when they hear the words "testing in education." Rather, we have in mind the types of assessments (tests, essays, exercises) given in the classroom or assigned for homework. The reason we advocate testing is that it requires students to retrieve information effortfully from memory, and such effortful retrieval turns out to be a wonderfully powerful mnemonic device in many circumstances.

Tests have both indirect and direct effects on learning (Roediger & Karpicke, 2006b). The indirect effect is that, if tests are given more frequently, students study more. Consider a college class in which there is only a midterm and a final exam compared to a similar class in which weekly quizzes are given every Friday, in addition to the midterm and the final. A large research program is not required to determine that students study more in the class with weekly quizzes than in the class without them. Yet tests also have a direct effect on learning; many studies have shown that students' retrieval of information on tests greatly improves their later retention of the tested

material, either compared to a no-intervention control or even compared to a control condition in which students study the material for an equivalent amount of time to that given to students taking the test. That is, taking a test on material often yields greater gains than restudying material, as we document below. These findings have important educational implications, ones that teachers and professors have not exploited.

In this chapter, we first report selectively on findings from our lab on the critical importance of testing (or retrieval) for future remembering. Retrieval is a powerful mnemonic enhancer. However, testing does not lead to improvements under all possible conditions, so the remainder of our chapter will discuss qualifications and boundary conditions of test-enhanced learning, as we call our program (McDaniel, Roediger, & McDermott, 2007b). We consider issues of test format in one section, such as whether multiple-choice or short answer tests produce greater enhancements in performance. Another critical issue, considered in the next section, is the role of feedback: When is it helpful, or is it always helpful? We then discuss how to schedule tests and whether tests should occur frequently with short spacings between them – should we strike memory again when the iron is hot, as it were? Or should tests be spaced out in time, and, if so, how? In the next section, we ask if true/false and multiple-choice tests can ever have negative influences on learning. These tests provide students with erroneous information, either in the form of false statements (in true/false tests) or plausible alternatives that are nearly, but not quite, correct (in multiple-choice tests). Might students pick up misinformation from these kinds of tests, just as they do in other situations (e.g., Loftus, Miller, & Burns, 1978)? We then turn to the issue of metacognition, and examine students' beliefs and practices about testing and how they think it compares to other study strategies. Finally, we discuss how the findings on testing reviewed in this chapter might be applied in the classroom, as recent studies show that test-enhanced learning works in actual classrooms from middle school to college. We end with a few reflections on the role of testing in enhancing educational attainment.

TEST-ENHANCED LEARNING

Psychologists have studied the effects of testing on later memory, off and on, for 100 years (Abbott, 1909). In this section we report two experiments from our own lab to illustrate the power of testing to readers who may not be familiar with this literature and to blunt one main criticism of some testing research (see Roediger & Karpicke [2006b] for a thorough review).

Consider first a study by Wheeler and Roediger (1992). As part of a larger experiment, students in one condition studied 60 pictures while listening to a story. The subjects were told that they should remember the pictures, because they would be tested on the names of the pictures (which were given in the story). The test was free recall, meaning that students were given a blank

sheet of paper and asked to recall as many of the items as possible. After hearing the story, one group of students was permitted to leave the lab and asked to return 1 week later for the test. A second group was given a single test that lasted about 7 minutes. A third group was given three successive tests. That is, a minute after their first test, they were given a new blank sheet of paper and asked to recall as many of the 60 pictures as they could for a second time. After they were finished, the procedure was repeated a third time. The group that was given a single recall test produced 32 items on the test; the group that took three tests recalled 32, 35 and 36, respectively. The improvement in recall across repeated tests (even though each later test is further delayed from original study) is called hypermnesia. However, the real interest for present purposes is how students performed on a final test a week later. All subjects had studied the same list of pictures while listening to a story, so the only difference was whether they had taken 0, 1 or 3 tests just after the study phase of the experiment.

The results from the 1-week delayed test are shown in Figure 1.1. Subjects who did not take a test during the first session recalled 17 items, those who had taken one test recalled 23 items, whereas those who had taken three tests recalled 32 items. The number of tests given just after learning greatly

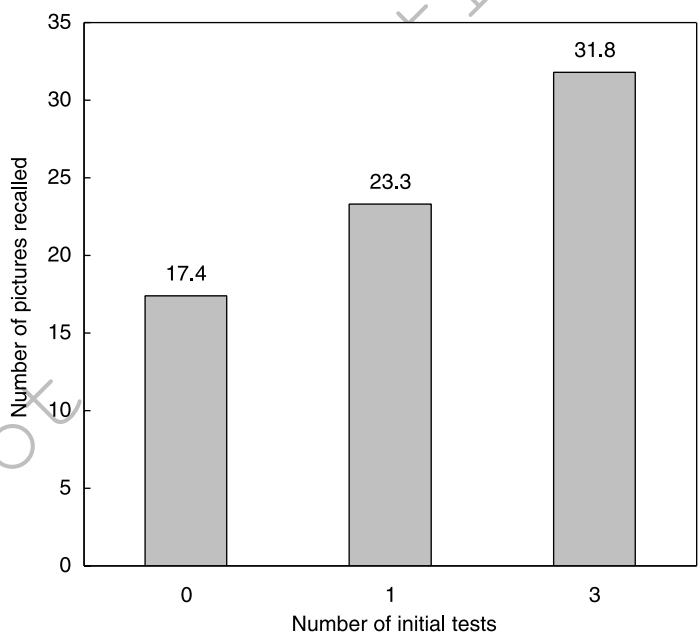

Figure 1.1 Number of pictures recalled on a 1-week delayed test, adapted from Table 1.1 in Wheeler and Roediger (1992). The number of tests given just after learning greatly affected performance a week later; three prior tests raised recall over 80% relative to the no-test condition, and the act of taking three tests virtually eliminated the forgetting process.

affected performance a week later; three prior tests raised recall over 80% relative to the no-test condition (i.e., (32 – 17)/17 × 100). Looked at another way, immediately after study about 32 items could be recalled. If subjects took three tests just after recall, they could still recall 32 items a week later. The act of taking three tests essentially stopped the forgetting process in its tracks, so testing may be a mechanism to permit memories to consolidate or reconsolidate (Dudai, 2006).

Critics, however, could pounce on a potential flaw in the Wheeler and Roediger (1992) experiment just reported. Perhaps, they would carp, repeated testing simply exposes students to information again. That is, all "testing" does is allow for repeated study opportunities, and so the testing effect is no more surprising than the fact that when people study information two (or more) times they remember it better than if they study it once (e.g., Thompson, Wenger, & Bartling, 1978). This objection is plausible, but has been countered in many experiments that compared subjects who were tested to ones who spent the same amount of time restudying the material. The consistent finding is that taking an initial test produces greater recall on a final test than does restudying material (see Roediger & Karpicke, 2006b). Here we report only one experiment that makes the point.

Roediger and Karpicke (2006a, Experiment 1) had students read brief prose passages about a variety of topics, many having to do with science ("The Sun" or "Sea Otters") and other topics. After reading the passage, students either took a 7-minute test on the passage or read it again. Thus, in one condition, students studied the passage twice, whereas in the other they studied it once and took a test. The test consisted of students being given the title of the passage and asked to recall as much of it as possible. The data were scored in terms of the number of idea units recalled. The students taking the test recalled about 70% of the idea units during the test; on the other hand, students who restudied the passage were of course exposed to all the ideas in the passage. Thus, students who reread the passage actually received a greater exposure to the material than did students who took the test. The final test on the passages was either 5 minutes, 2 days or 7 days later, and was manipulated between subjects.

The results are shown in Figure 1.2 and several notable patterns can be seen. First, on the test given after a short (5-minute) delay, students who had repeatedly studied the material recalled it better than those who had studied it once and taken a test. Cramming (repeatedly reading) does work, at least at very short retention intervals. However, on the two delayed tests, the pattern reversed; studying and taking an initial test led to better performance on the delayed test than did studying the material twice. Testing enhanced long-term retention. Many other experiments, some of which are discussed below, have reported this same pattern (see Roediger & Karpicke, 2006b, for a review).

The results reviewed above, along with many others dating back over a century, establish the reality of the testing effect. However, not all experiments reveal testing effects. In the sections below, we consider variables that

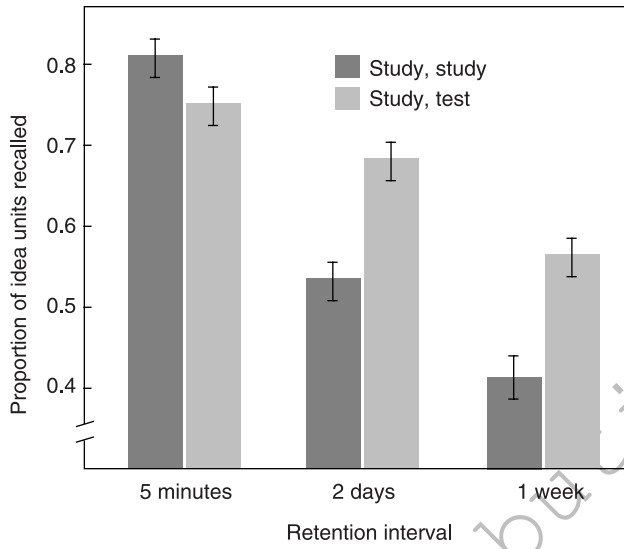

Figure 1.2 Results from Roediger and Karpicke (2006a, Experiment 1). On the 5-minute delayed test, students who had repeatedly studied the material recalled it better than those who had studied it once and taken a test. Cramming (repeatedly reading) does work, at least at very short retention intervals. However, on the two delayed tests, the pattern reversed; studying and taking an initial test led to better performance on the delayed test than did studying the material twice.

modulate the magnitude of the testing effect, beginning with the format of tests.

THE FORMAT OF TESTS

The power of testing to increase learning and retention has been demonstrated in numerous studies using a diverse range of materials; but both study and test materials come in a multitude of formats. Although the use of true/false and multiple-choice exams is now commonplace in high school and college classrooms, there was a time (in the 1920s and 1930s) when these kinds of exams were a novelty and referred to as "new-type," in contrast to the more traditional essay exams (Ruch, 1929). Given the variety of test formats, one question that arises is whether all formats are equally efficacious in improving retention. If we want to provide evidence-based recommendations for educators to utilize testing as a learning tool, it is important to ascertain if particular types of tests are more effective than others.

In a study designed to examine precisely this issue, Kang, McDermott, and Roediger (2007) manipulated the formats of both the initial and final tests—multiple-choice (MC) or short answer (SA)—using a fully-crossed,

within-subjects design. Students read four short journal articles, and immediately afterwards they were given an MC quiz, an SA quiz, a list of statements to read, or a filler task. Feedback was given on quiz answers, and the quizzes and the list of statements all targeted the same critical facts. For instance, after reading an article on literacy acquisition, students in the SA condition generated an answer to "What is a phoneme?" (among other questions), students in the MC condition selected one of four possible alternatives to answer the same question, and students in the read-statements condition read "A phoneme is the basic sound unit of a language." This last condition allowed the effects of testing to be compared to the consequences of focused re-exposure to the target information (i.e., similar to receiving the test answers, without having to take the test). This control condition was a very conservative one, given that students in real life generally do not receive the answers to upcoming exams. A more typical baseline condition (i.e., having a filler task after reading the article) was also compared to the testing and focused rereading conditions. Three days later, subjects took a final test consisting of MC and SA questions.

Figure 1.3 shows that final performance was best in the initial SA condition. The initial MC condition led to the next best performance, followed by the read-statements condition and finally the filler-task condition. This pattern of results held for both final MC and final SA questions. Final test scores

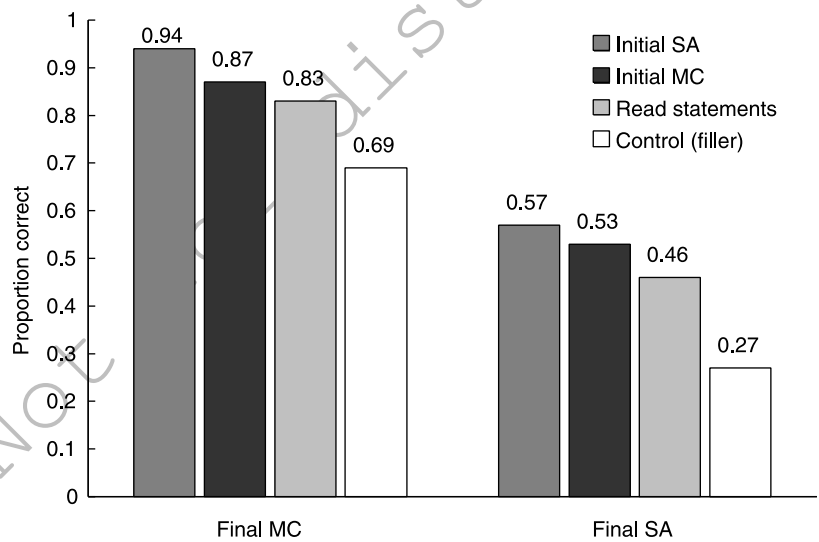

Figure 1.3 Results from Kang, McDermott, and Roediger (2007). Regardless of the format of the final test, the initial test format that required more effortful retrieval (i.e., the SA condition) yielded the best final performance, which was significantly better than being given the test answers without having to take a test. Although taking an initial MC test did benefit final performance relative to the filler-task control condition, the boost was not significantly above the read-statements condition.

were significantly worse in the filler-task condition than the other three conditions, indicating that both testing (with feedback) and focused re-exposure aid retention of the target information. Importantly, only the initial SA condition produced significantly better final performance than the read-statements condition; the initial MC and read-statements conditions did not differ significantly. Retrieval is a potent memory modifier (Bjork, 1975). These results implicate the processes involved in actively producing information from memory as the causal mechanism underlying the testing effect. Regardless of the format of the final test, the initial test format that required more effortful retrieval (i.e., short answer) yielded the best final performance, and this condition was significantly better than having read the test answers in isolation.

Similar results from other studies provide converging evidence that effortful retrieval is crucial for the testing effect (Carpenter & DeLosh, 2006; Glover, 1989). Butler and Roediger (2007), for example, used art history video lectures to simulate classroom learning. After the lectures, students completed short answer or multiple-choice tests, or they read statements as in Kang et al. (2007). On a final SA test given 30 days later, Butler and Roediger found the same pattern of results: (1) retention of target facts was best when students were given an initial SA quiz, and (2) taking an initial MC test produced final performance equivalent to reading the test answers (without taking a test). As discussed in a later section of this chapter, these findings have been replicated in an actual college course (McDaniel, Anderson, Derbish, & Morrisette, 2007b).

Although most evidence suggests that tests that require effortful retrieval yield the most memorial benefits, it should be noted that this depends upon successful retrieval on the initial test (or the delivery of feedback). Kang et al. (2007) had another experiment identical to the one described earlier except that no feedback was provided on the initial tests. Without corrective feedback, final test performance changed: the initial SA condition yielded poorer performance than the initial MC condition. This difference makes sense when performance on the initial tests is considered: accuracy was much lower on the initial SA test (M = .54) than on the initial MC test (M = .86). The beneficial effect of testing can be attenuated when initial test performance is low (Wenger, Thompson, & Bartling, 1980) and no corrective feedback is provided.

This same conclusion about the role of level of initial performance can be drawn from Spitzer's (1939) early mega-study involving 3605 sixth-graders in Iowa. Students read an article on bamboo, after which they were tested. Spitzer manipulated the frequency of testing (students took between one and three tests) and the delay until the initial test (which ranged from immediately after reading the article to 63 days later). Most important for present purposes is that the benefit of prior testing became smaller the longer one waited after studying for the initial test; i.e., performance on the initial test declined with increasing retention interval, reducing the boost to performance on

subsequent tests. In a similar vein, it has been shown that items that elicit errors on a cued recall test have almost no chance of being recalled correctly at a later time unless feedback is given (Pashler, Cepeda, Wixted, & Rohrer, 2005). In other words, learning from the test is handicapped when accuracy is low on the initial test (whereas this problem does not occur with rereading, where there is re-exposure to 100% of the target information). For the testing effect to manifest itself fully, feedback must be provided if initial test performance is low. Recent research showing that retrieval failure on a test (i.e., attempting to recall an answer but failing to) can enhance future encoding of the target information (Kornell, Hays, & Bjork, 2009; Richland, Kornell & Kao, 2009) emphasizes further the utility of providing feedback when initial test performance is poor.

A recent study (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2008) delved more deeply into the issue of whether the kind of test influences the testing effect. In a closed-book test, students take the test without having concurrent access to their study materials, and this is the traditional way in which tests have been administered. In recent years, open-book tests – where students are permitted to consult their notes and textbooks during the test – have grown in popularity, with the belief that such tests promote higher-level thinking skills (e.g., Feller, 1994). The final test only involved short answer questions. We examined the issue of whether administering the initial test open- or closed-book made a difference to later retention of target facts. For the sake of brevity, only the second experiment will be described here (the results replicate the first experiment, which was similar in design). Students studied expository passages in various learning conditions: three study-only conditions (read the passage once, twice, or three times) and four test conditions (a closed-book test, a closed-book test with feedback, or an open-book test after reading the passage once, or a test completed simultaneously while reading the passage).

A final closed-book test was given a week later, and the key results are summarized in Table 1.1. Just taking a closed-book test (without feedback) resulted in final performance that was roughly equivalent to reading the passage three times (.55 vs. .54), and significantly better than reading the passage twice (.55 vs. 50), once again demonstrating the power of testing. The two learning conditions that tied for the best performance, however, were the closed-book test with feedback and the open-book test conditions (*M*s = .66), both of which produced significantly more correct responses on the final test than all the other conditions. In our view, these two learning conditions contained two critical components – testing (retrieval) and feedback – that the other learning conditions lacked (or had only one or the other), and this combination contributed to best performance on the delayed test.

Although the current data suggest the equivalence of closed-book tests with feedback and open-book tests in enhancing later retention, further investigation into this issue is warranted, because on theoretical grounds one might expect a closed-book test to involve more retrieval effort than an

Table 1.1 Mean proportion recalled in Agarwal et al.'s (2008) Experiment 2 on test formats. Proportion correct was greater for test conditions than study conditions; however, subjects predicted learning would be greater for study conditions relative to test conditions. Learning conditions that contained both testing and feedback, namely the closed-book test with feedback and the open-book test conditions, contributed to best performance on the delayed test.

	Proportion correct	
Condition	*Initial test*	*One-week delayed test*
Study 1×		.40
Study 2×		.50
Study 3×		.54
Closed-book test	.67	.55
Closed-book test with feedback	.65	.66
Open-book test	.81	.66
Simultaneous answering	.83	.59
Non-studied control		.16

open-book test. Perhaps a difference between these two conditions will emerge with a longer delay for the final test, an outcome that has occurred in other experiments (e.g., Roediger & Karpicke, 2006a). Such a finding would probably further depend on how students approach the open-book tests (e.g., whether they attempt retrieval of an answer before consulting the study material for feedback, or whether they immediately search the study material in order to identify the target information). Future research in our lab will tackle this topic.

In summary, one reason why testing benefits memory is that it promotes active retrieval of information. Not all formats of tests are equal. Test formats that require more effortful retrieval (e.g., short answer) tend to produce a greater boost to learning and retention, compared to test formats that engage less effortful retrieval (e.g., multiple-choice). However, tests that are more effortful or challenging also increase the likelihood of retrieval failure, which has been shown to reduce the beneficial effect of testing. Therefore, to ameliorate low performance on the initial test, corrective feedback should be provided. The practical implications of these findings for improving learning in the classroom are straightforward: instead of giving students summary notes to read, teachers should implement more frequent testing (of important facts and concepts) – using test formats that entail effortful retrieval – and provide feedback to correct errors. We turn now to a greater consideration of the issue of how and when feedback should be given after tests.

TESTING AND FEEDBACK

Broadly defined, feedback is information provided following a response or recollection, which informs the learner about the status of current performance, often leading to improvement in future performance (Roediger, Zaromb, & Butler, 2008). A great deal of laboratory and applied research has examined the conditions under which feedback is, and is not, effective in improving learning and performance. We have already mentioned the critical importance of feedback when initial performance is low, as well as the benefit of feedback regardless of test format. In a later section, we discuss the benefits of feedback in reducing the negative effects of multiple-choice tests. For now, we focus on the effect of feedback type (e.g., corrective or right/wrong) and delay (e.g., immediate vs. delayed) on a learner's future responses and confidence regarding their own performance.

To begin, feedback can encompass a range of information: re-presentation of the original material, information regarding whether the learner is simply correct or incorrect, or specific information regarding the correct response, just to name a few. In a study by Pashler et al. (2005), subjects studied Luganda–English word pairs (e.g., *leero – today*) and received no feedback, right/wrong feedback, or correct answer feedback following each response on an initial test. After one week, performance following the correct answer feedback condition was significantly greater than performance in either of the other conditions. In addition, right/wrong feedback did not benefit retention over and above the absence of feedback, a finding replicated in other studies (e.g., see Bangert-Drowns, Kulik, Kulik, & Morgan, 1991, for a review). Furthermore, Moreno (2004) demonstrated that explanatory feedback (e.g., providing students with the correct answer and an explanation) increased final retention and transfer relative to corrective feedback (e.g., only providing students with the correct answer). In general, in order to improve performance, feedback must include corrective information, and the learner may further benefit when an explanation follows the corrective feedback.

Although Pashler et al. (2005) found the benefits of corrective feedback to be limited to revising errors, Butler, Karpicke, and Roediger (2007) found that feedback benefits initially correct answers as well as incorrect responses. This recent finding is consistent with earlier research, which demonstrated that subjects who receive feedback are more likely to fix initial errors and remember initially correct responses on a final criterial test (Kulhavy, Yekovich, & Dyer, 1976, 1979). More specifically, subjects spend more time processing feedback following errors made with high confidence and correct answers made with low confidence. Butterfield and Metcalfe (2006) held feedback time constant, and still found subjects were more likely to later correct errors made with high confidence than those made with low confidence. They named this the hypercorrection effect, and argued that subjects attend more to feedback that is at odds with their expectations (see also Fazio

& Marsh, 2009). Similarly, Butler, Karpicke, and Roediger (2008) suggested that feedback reinforces the association between a cue and its target response, increasing the likelihood that an initially low-confidence correct response will be produced on a final criterial test.

Regarding the best time to deliver feedback, there exists a great deal of debate and confusion, as immediate and delayed feedback are operationalized differently across studies. For instance, the term "immediate feedback" has been used to imply feedback given just after each test item or feedback provided immediately after a test. On the other hand, "delayed feedback" can take place anywhere from 8 seconds after an item to 2 days after the test (Kulik & Kulik, 1988), although in many educational settings the feedback may actually occur a week or more later.

Butler et al. (2007) investigated the effects of type and timing of feedback on long-term retention. Subjects read prose passages and completed an initial multiple-choice test. For some responses, they received standard feedback (i.e., the correct answer); for others they received feedback by answering until correct (labeled AUC: i.e., each response was labeled as correct or incorrect, and if incorrect they chose additional options until they answered the question correctly). Half of the subjects received the feedback immediately after each question, while the rest received the feedback after a 1-day delay. One week later, subjects completed a final cued recall test, and these data are shown in Figure 1.4. On the final test, delayed feedback led to substantially better performance than immediate feedback, while the standard feedback condition and the answer-until-correct feedback condition resulted in similar performance. Butler et al. discussed why some studies find immediate feedback to be more beneficial than delayed feedback (e.g., see Kulik & Kulik, 1988, for a review). Namely, this inconsistency might occur if learners do not fully process delayed feedback, which would be particularly likely in applied studies where less experimental control is present. That is, even if students receive delayed feedback they may not look at it or look only at feedback on questions that they missed. When feedback processing is controlled, as in the Butler et al. study, a benefit for delayed feedback on long-term retention emerges.

In sum, the provision of feedback leads to substantial increases in long-term learning. Delayed feedback virtually always boosts final performance if the feedback includes both the correct answer as well as an explanation of that answer. Feedback also serves a variety of purposes: correcting errors, improving retention of correct responses, and enhancing metacognition. Feedback should be incorporated into all educational testing.

SCHEDULES FOR TESTING

The great bulk of the literature on testing effects shows the benefit of a single initial test relative to either no test or to a reading control condition. The fact

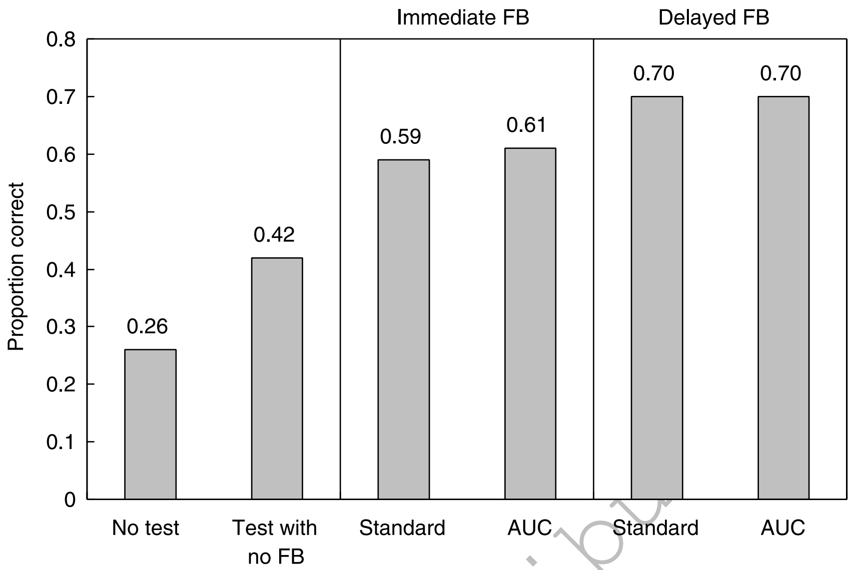

Figure 1.4 Results from Butler, Karpicke, and Roediger (2007). On the final test, delayed feedback (FB) led to substantially better performance than immediate feedback, while the standard feedback condition and the answer-until-correct (AUC) feedback condition resulted in similar performance. When feedback processing is controlled, a benefit for delayed feedback on long-term retention emerges.

that the testing effect occurs robustly in such situations indicates the power of testing, but one can ask whether students would learn better if they received multiple tests on the same material. For example, the Wheeler and Roediger (1992) data shown in Figure 1.1 indicate that three tests shortly after study led to better recall a week later relative to one test. Is this effect general?

Anecdotally, one might expect that it is. For example, when children in the early primary grades are taught their multiplication tables, they often use or construct flashcards. All problems up to 9 × 9 (or even higher) are created so that one side of the card might say 6 × 8 = ?? and the other side has 48. Students are instructed to test themselves repeatedly on the cards, flipping over to the other side when they need feedback. Students are usually instructed to do this until answering the item becomes quick and effortless, but it takes repeated practice to reach this state.

Flashcards are used in many other situations to learn large bodies of factual information, and educational companies make flashcards for a huge number of purposes, including learning foreign language vocabulary, the parts of the skeletal system, birds and their names, and so on. Research on how to effectively use flashcards is relatively new, however. One instruction often given in mastering a set of flashcards is to learn to give the correct answer, then practice it once, and then to drop that card from the pile to

concentrate on others that have not yet been learned. The assumption is that learning to a criterion of one correct recitation means that the item is learned and that further practice on it will be for naught.

Karpicke and Roediger (2008) published a study that questions this common wisdom. In their study, students learned foreign language vocabulary in the form of Swahili–English words pairs. (Swahili was used because students were unfamiliar with the language, and yet the word forms were easily pronounceable for English speakers, such as *mashua–boat*). Students learned 40 pairs under one of four conditions. In one condition, students studied and were tested on the 40 pairs in the usual multitrial learning situation favored by psychologists (study–test, study–test, study–test, study–test, labeled the ST condition). In a second condition, students received a similar first study–test cycle, but if they correctly recalled pairs on the test, these pairs were dropped from the next study trial. Thus, across the four trials, the study list got smaller and smaller as students recalled more items. However, in this condition, labeled S_NT, students were tested on all 40 pairs during each test period. Thus, relative to the ST condition, the S_NT condition involved fewer study opportunities but the same number of tests. In a third condition, labeled ST_N, students studied and were tested the same way as in the other conditions on the first trial, but after the first trial they repeatedly studied the pairs three more times, but once they had recalled a pair, it was dropped from the test. In this condition, the study sequence stayed the same on four occasions, but the number of items tested became smaller and smaller. Finally, in a fourth condition denoted S_NT_N, after the first study–test trial, items that were recalled were dropped both from the study and test phase of the experiment for the additional trials. In this case, then, the study list and the test sequence became shorter over trials. This last condition is most like standard advice for using flashcards – students studied and were tested on the pairs until they were recalled, and then they were dropped so that attention could be devoted to unlearned pairs.

Initial learning on the 40 pairs in the four conditions is shown in Figure 1.5, where it can be seen that all four conditions produced equivalent learning. The data in Figure 1.5 show cumulative performance, such that students were given credit the first time they recalled a pair and not again (for those conditions in which multiple recalls of a pair were required, ST and S_NT). At the end of the learning phase, students were told that they would come back a week later to be tested again and were asked to predict how many pairs they would recall. Students in all four groups estimated that they would recall about 20 pairs, or 50% of the pairs, a week later. After all, the learning curves were equal, so why would we expect students' judgments to differ?

Figure 1.6 shows the proportion of items recalled 1 week later, in each of the four conditions. Students in two conditions did very well (ST and S_NT, around 80%) and in the other two conditions students did much more poorly (ST_N and S_NT_N, around 35%). What do the former two conditions have in common that the latter two conditions lack? The answer is retrieval practice.

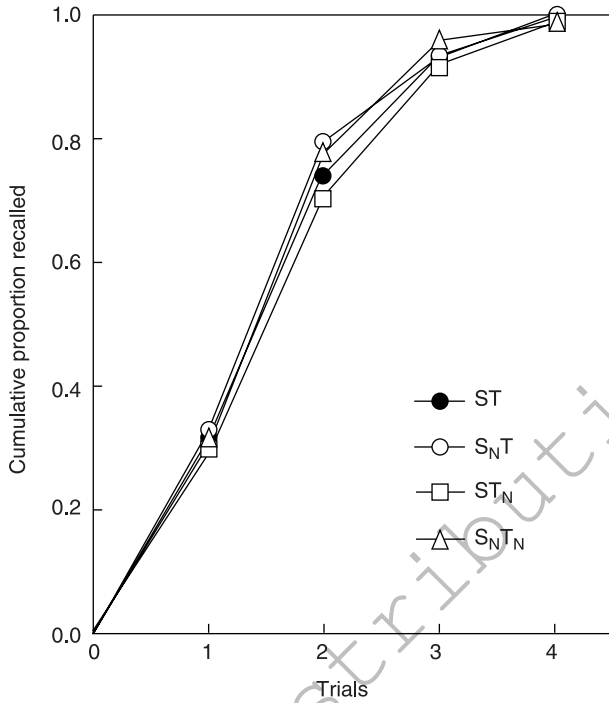

Figure 1.5 Initial cumulative learning performance from Karpicke and Roediger (2008). All four conditions produced equivalent learning.

In both the ST and S_NT conditions, students were tested on all 40 pairs for all four trials. Note that in the ST condition students studied all 40 items four times, whereas in the S_NT condition, the items were dropped from study. However, this reduced amount of study did not matter a bit for retention a week later. Students in the ST_N and S_NT_N conditions had only enough testing for each item to be recalled once and, without repeated retrieval, final recall was relatively poor. Once again, the condition with many more study opportunities (ST_N) did not lead to any appreciably better recall a week later than the condition that had minimal study opportunities (S_NT_N).

The bottom line from the Karpicke and Roediger (2008) experiment is that after students have retrieved a pair correctly once, repeated retrieval is the key to improved long-term retention. Repeated studying after this point does not much matter.

Recall that the students in the four conditions predicted that they would do equally well, and recall about 50% after a week. As can be seen in Figure 1.6, the students who were repeatedly tested actually outperformed their predictions, so they underestimated the power of testing. On the other hand, the students who did not have repeated testing overestimated how well they would do. In a later section, we return to the issue of what students know

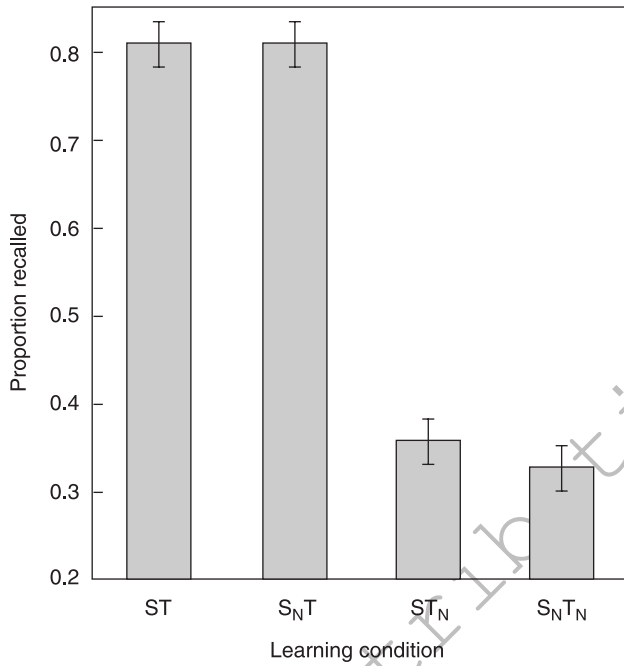

Figure 1.6 Final learning results from Karpicke and Roediger (2008). Students in the ST and S_NT conditions performed very well, and students in the ST_N and S_NT_N conditions did much more poorly. In both the ST and S_NT conditions, students were tested on all 40 pairs for all four trials. Students in the ST_N and S_NT_N conditions had only enough testing for each item to be recalled once and, without repeated retrieval, recall was relatively poor. The condition with many more study opportunities (ST_N) did not lead to any appreciably better recall a week later than the condition that had minimal study opportunities (S_NT_N).

about the effects of testing and whether they use testing as a study strategy when left to their own devices.

Repeated testing seems to be great for consolidating information into long-term memory, but is there an optimal schedule for repeated testing? Landauer and Bjork (1978) argued that a condition they called expanding retrieval was optimal, or at least was better than two other schedules called massed practice and equal interval practice. To explain, let us stick with our foreign-language vocabulary learning example above, *mashua–boat*, and consider patterns in which three retrievals of the target might be carried out. In the immediate test condition, after the item has been presented, *mashua* would be presented three times in a row for boat to be recalled each time. This condition might provide good practice because, of course, performance on each test would be nearly perfect. (In fact, in most experiments, this massed testing condition leads to 98% or higher correct recall with paired-associates.)

The massed retrieval condition will be denoted a 0-0-0 to indicate that the three retrievals occurred back to back, with no other study or test items between retrievals of the target.

A second condition is the equal interval schedule, in which tests are given after a delay from study and at equal intervals after that. So, in a 5-5-5 schedule, a pair like *mashua–boat* would be presented, then five other pairs or tests would occur, and then *mashua – ??* would be given as a cue. This same process would occur two more times. Although distributed retrieval could be beneficial relative to massed retrieval, just as distributed study practice is beneficial relative to massed practice, one looming problem occurs in the case of retrieval – if the first test is delayed, recall on the first test may be low and, as discussed above, low performance on a test can reduce or eliminate the power of the testing effect. To overcome this problem, Landauer and Bjork (1978) introduced the idea of expanding retrieval practice, to insure a nearly errorless early retrieval with a quick first test while at the same time gaining advantages of distributed testing or practice. So, to continue with our example, in an expanding schedule of 1-4-10, students would be tested with *mashua* – ?? after only 1 intervening item, then again after 4 intervening items, and then after 10 intervening items. The idea behind the expanding schedule is familiar to psychologists because it resembles the idea of shaping behavior by successive approximations (Skinner, 1953, Chapter 6); just as schedules of reinforcement (Ferster & Skinner, 1957) exist to shape behavioral responses, so schedules of retrieval may shape the ability to remember. If a student wants to be able to retrieve a vocabulary word long after study, the expanding retrieval schedule may help to shape its retrieval.

Landauer and Bjork (1978) conducted two experiments pitting massed, equal interval, and expanding interval schedules of retrieval against one another. For the latter conditions, they used 5-5-5 spacing and 1-4-10 spacing. Note that this comparison equates the average interval of spacing at 5. The materials in one experiment were fictitious first and last names, such that students were required to produce a last name when given the first name. Landauer and Bjork measured performance on the three initial tests and then on a final test given at the end of the experimental session. The results for the expanding and equal interval retrieval sequences are shown in Table 1.2 for

Table 1.2 Mean proportion recalled in Landauer and Bjork's (1978) experiment on schedules of testing; data are estimated from Figures 1.1 and 1.2. Expanding retrieval schedules were better than equal interval schedules on both the initial three tests and the final criterial test.

	Initial tests			*Final test*
	1	*2*	*3*	
Expanding	.61	.55	.50	.47
Equal	.42	.42	.43	.40

the three initial tests and then the final, criterial, test. Expanding retrieval schedules were better than equal interval schedules, as Landauer and Bjork predicted, on both the initial three tests and then the final criterial test. The 7% advantage of expanding interval retrieval to equally spaced retrieval on the final test was small but statistically significant, and this is the comparison that the authors emphasized in the paper. They replicated the effect in a separate experiment with face–name pairs. However, note a curious fact about the data in Table 1.2: Over the four tests shown, performance drops steadily in the expanding interval condition (from .61 to .47) whereas in the equal interval condition performance is essentially flat (.42 to .40). This pattern suggests that on a more delayed final test, the curves might cross over and equal interval retrieval might prove superior to expanding retrieval.

Strangely, for some years researchers did not investigate Landauer and Bjork's (1978) intriguing findings, perhaps because they made such good sense. Most of the studies on retrieval schedules compared expanding and massed retrieval, but did not include the critical equal interval condition needed to compare expanding retrieval to another distributed schedule (e.g., Rea & Modigliani, 1985). All studies making the massed versus expanding retrieval comparison showed expanding retrieval to be more effective, and Balota, Duchek, and Logan (2007) have provided an excellent review of this literature. They show conclusively that massed testing is a poor strategy relative to distributed testing, despite the fact that massed testing produces very high performance on the initial tests (much higher than equal interval testing). Although this might seem commonplace to cognitive psychologists steeped in the literature of massed versus spaced presentation (the spacing effect), from a different perspective the outcome is surprising. Skinner (1958) promoted the notion of errorless retrieval as being the key to learning, and he implemented this approach into his teaching machines and programmed learning. However, current research shows that distributed retrieval is much more effective in promoting later performance than is massed retrieval, even though massed retrieval produces errorless performance.

On the other hand, when comparisons are made between expanding and equal interval schedules, the data are much less conclusive. The other main point established in the Balota et al. (2007) review is that no consistent evidence exists for the advantage of expanding retrieval schedules over equal interval testing sequences. A few studies after Landauer and Bjork's (1978) seminal study obtained the effect, but the great majority did not. For example, Cull (2000) reported four experiments in which students learned difficult word pairs. Across experiments, he manipulated variables such as intertest interval, feedback or no feedback after the tests, and testing versus restudying the material. The general conclusion drawn from the four experiments was that distributed retrieval produced much better retention on a final test than did massed retrieval, but that it did not matter whether the schedule had uniform or expanded spacing of tests.

Recent research by Karpicke and Roediger (2007) and Logan and Balota

(2008) actually shows a more interesting pattern. On tests that occur a day or more after original learning, equal interval schedules of initial testing actually produce greater long-term retention than do expanding schedules (just the opposite of Landauer and Bjork's findings). Recall the data in Table 1.2 and how the expanding retrieval testing condition showed a steady decline with repeated tests whereas the equal interval schedule showed essentially no decline. Because the final test in these studies occurred during the same session as initial learning, the retention interval for the final test was fairly short, leaving open the possibility that on a long-delayed test the functions would actually cross. This is just what both Karpicke and Roediger and Logan and Balota found.

Karpicke and Roediger (2007) had students learn word pairs taken from practice tests for the Graduate Record Exam (e.g., *sobriquet–nickname, benison–blessing*) and tests consisted of giving the first member of the pair and asking for the second. Their initial testing conditions were massed (0-0-0), expanding (1-5-9) and equal interval (5-5-5). In addition, they included two conditions in which students received only a single test after either 1 intervening pair or 5. The design and initial test results are shown on the left side of Table 1.3. Initial test performance was best in the massed condition, next in the expanding condition, and worst in the equally spaced condition, the usual pattern, and students in the single test condition recalled less after a delay of 5 intervening items than after 1. There are no surprises in the initial recall data. Half the students took a final test 10 minutes after the initial learning phase, whereas the rest received the final test 2 days later. These results are shown in the right side of Table 1.3. First consider data at the 10-minute delay. The top three rows show a very nice replication of the pattern reported by Landauer and Bjork (1978): Expanding retrieval in the initial phase produced better recall on the final test than the equal interval schedule, and both of these schedules were better than the massed retrieval

Table 1.3 Mean proportion recalled in Karpicke and Roediger's (2007) experiment on schedules of testing. Expanding retrieval in the initial phase produced better recall than the equal interval schedule on the 10-minute delayed test, and both of these schedules were better than the massed retrieval schedule. However, after a 2-day delay, recall was best in the equal interval condition relative to the expanding condition, although both conditions still produced better performance than in the massed condition.

	Initial tests			*Final tests*	
	1	*2*	*3*	*10 min*	*48 h*
Massed (0-0-0)	.98	.98	.98	.47	.20
Expanding (1-5-9)	.78	.76	.77	.71	.33
Equal (5-5-5)	.73	.73	.73	.62	.45
Single-immediate (1)	.81			.65	.22
Single-delayed (5)	.73			.57	.30

schedule. Also, the single-immediate test produced better delayed recall than the single-delayed test. However, the startling result in this experiment appeared for those subjects who took the test after a 2-day delay. Recall was now best in the equal interval condition (M = .45) relative to the expanding condition (M = .33), although both conditions still produced better performance than in the massed condition (M = .20). Interestingly, performance also reversed across the delay for the two single test conditions: recall was better in the single-immediate condition after 10 minutes, but was reliably better in the single-delayed condition after 2 days.

Karpicke and Roediger (2007) argued that, congenial as the idea is, expanding retrieval is not conducive to good long-term retention. Instead, what seems to be important for long-term retention is the difficulty of the first retrieval attempt. When subjects try to retrieve an item after 5 intervening items, they exert more effort during retrieval than after 1 intervening item, which in turn is more difficult than after no intervening items (the massed condition). Notice that the same pattern occurs for items that were tested only once, after either 1 or 5 intervening items. Karpicke and Roediger (2007) replicated these results in a second experiment in which feedback was given after the initial tests. A third experiment confirmed that delaying the first test is the critical ingredient in enhancing long-term retention and that the method of distributing retrieval (expanding or equal interval) does not matter. Logan and Balota (2008) reported similar data and reached the same general conclusion.

The conclusion that retrieval difficulty is the key element promoting better long-term retention for equal interval (relative to expanding or massed) schedules fits well with data reviewed in previous parts of the chapter, such as recall tests producing a greater testing effect than recognition tests. However, as Balota et al. (2007) have noted, the literature on expanding retrieval sequences has used a relatively small array of conditions (three tests, paired associate learning of one sort or another, relatively immediate tests). One can imagine that an expanding retrieval schedule could be better than equal intervals if there were more presentations and these occurred over longer periods of time. However, this conjecture awaits future testing.

DANGERS OF MULTIPLE-CHOICE AND TRUE/FALSE TESTS

We have established that students learn from tests, and that this learning seems to be especially durable. However, the fact that we learn from tests can also pose a danger in some situations. Although professors would never knowingly present wrong information during lectures or in assigned readings, they do it routinely when they give certain types of tests, viz., true/false and multiple-choice tests. If students learn from tests, might they learn wrong information if it is presented on the tests? On true/false tests it is common for

half the statements to be right and half to be wrong, and normally false items are plausible in order to require rather fine discriminations. Similarly, for multiple-choice tests, students receive a question stem and then four possible completions, one of which is correct and three others that are erroneous (but again, statements that might be close to correct). Because erroneous information is presented on the tests, students might learn that incorrect information, especially if no feedback is given (as is often the case in college courses). If the test is especially difficult (meaning a large number of wrong answers are selected), the students may actually leave a test more confused about the material than when they walked in. However, even if conditions are such that students rarely commit errors, it might be that simply reading and carefully considering false statements on true/false tests and distractors on multiple-choice tests can lead later to erroneous knowledge. Several studies have shown that having people simply read statements (whether true or false) increases later judgments that the statements are true (Bacon, 1979; Begg, Armour, & Kerr, 1985; Hasher, Goldstein, & Toppino, 1977). This effect underlies the tactics of propagandists using "the big lie" technique by repeating a statement over and over until the populace believes it, and is also a favored tactic in most US presidential elections. If you repeat an untruth about an opponent repeatedly, the statement comes to be believed.

Remmers and Remmers (1926) first discussed the idea that incorrect information on tests might mislead students, when the "new" techniques of true/false and multiple-choice testing were introduced into education (Ruch, 1929). They called this outcome the negative suggestibility effect, although not much research was done on it for many years. Much later Toppino and his colleagues showed that statements presented as distractors on true/false and multiple-choice tests did indeed accrue truth value from their mere presentation, because these statements were judged as more true when mixed with novel statements in appropriately designed experiments (Toppino & Brochin, 1989; Toppino & Luipersbeck, 1993). In a similar vein, Brown (1988) and Jacoby and Hollingshead (1990) showed that exposing students to misspelled words increased misspelling of those words on a later oral test.

Roediger and Marsh (2005) asked whether giving a multiple-choice test (without feedback) would lead to a kind of misinformation effect (Loftus et al., 1978). That is, if students take a multiple-choice test on a subset of facts, and then take a short answer test on all facts, will prior testing increase intrusions of multiple-choice lures on the final test? Roediger and Marsh conducted a series of experiments to address this question, manipulating the difficulty of the material (and hence level of performance on the multiple-choice test) and the number of distractors given on the multiple-choice test. Three experiments were submitted using these tactics, but the editor asked us to drop our first two experiments (which established the phenomenon) and report only a third, control, experiment that showed that the negative suggestibility effect occurred under tightly controlled but not necessarily realistic conditions. We complied and published the third experiment, but the two

most interesting experiments (in our opinion) were not reported. We present them here to show that negative effects of testing do accrue from taking multiple-choice tests and to have the experiments published, albeit in terse form.

Our first experiment was exploratory, just to make sure we could get the effects we sought. We predicted that we would see a positive testing effect to the extent students were able to answer the multiple-choice questions. Of interest was whether selecting distractors on the multiple-choice test would lead to their intrusion on a later test. We selected 80 easy and 80 hard general knowledge questions from the Nelson and Narens (1980) norms. Subjects in the norming study correctly answered an average of 72% of easy questions ("What sport uses the terms 'gutter' and 'alley'? [bowling]) and 13% of the difficult items ("Which union general defeated the Confederate Army at the Civil War battle of Gettysburg?" [Meade]). Because the norms are for short answer questions, we generated three plausible distractors for each item. Four sets of 40 items (20 easy and 20 hard) were created and rotated through four multiple-choice test conditions: 40 items were not tested, 40 were tested with one distractor, 40 with two distractors, and 40 with 3 distractors. Thus the multiple-choice test consisted of 120 questions, and no feedback was given as to the correctness of the answers. Following this test, the 40 subjects in the experiment spent 5 minutes doing a visuospatial filler task before they took a final short answer (cued recall) test. They were given the 160 general knowledge questions (120 items previously tested with multiple-choice and the 40 nontested items).

Performance on the multiple-choice test is shown in the top panel of Table 1.4, in the section devoted to Experiment 1 data. Not surprisingly, performance was better on easy than difficult items and declined with the number of alternatives. However, we did succeed in manipulating the level of

Table 1.4 Proportion correct on a multiple-choice test as a function of question difficulty and number of alternatives (including correct) for each question in Experiments 1 and 2 of Roediger and Marsh (reported in this chapter). Performance was better on easy than difficult items and declined with the number of alternatives. Similarly, performance on unread passages was lower than for read passages, and performance generally declined with the number of distractors (albeit more for unread than read passages).

	Number of alternatives		
	Two	*Three*	*Four*
Experiment 1			
Easy questions	.91	.85	.85
Hard questions	.66	.55	.48
Experiment 2			
Passages read	.86	.86	.84
Passages not read	.68	.62	.51

multiple-choice performance, and this allowed us to see if any negative effects of testing were limited to conditions in which subjects made more errors on the multiple-choice test (i.e., difficult items and relatively many distractors).

The interesting data are contained in Tables 1.5 and 1.6 (again, the top panels devoted to Experiment 1). Table 1.5 shows the proportion of short answer questions answered correctly ("bowling" in response to "What sport uses the terms 'gutter' and 'alleys'?"). A strong positive testing effect appeared: Relative to the nontested questions, subjects correctly answered more previously-tested questions, for both easy and hard items. When previously tested with more multiple-choice distractors, the size of the positive effect dropped a bit for the hard items. However, the positive testing effect was robust in all conditions. Subjects had been required to guess on the short answer test (and they provided confidence ratings), but when we removed the "not sure" responses from the data, the same pattern held (for both high confidence and medium confidence answers). These data with "not sure responses removed" are shown in parentheses in Table 1.5.

Table 1.6 shows errors committed on the short answer test, and we found that prior multiple-choice testing also led to a negative effect (in addition to the positive testing effect just documented). The prior multiple-choice test led to more multiple-choice lure answers on the final test and this effect grew larger when more distractors had been presented on the multiple-choice test. Again, removing the "not sure" responses reduced the size of the negative suggestibility effect, but left the basic pattern intact. Those data are again in parentheses.

Table 1.5 Proportion correct on the cued recall test as a function of question difficulty and number of alternatives (including the correct answer) on the prior multiple-choice test. Non-guess responses are in parentheses (proportion correct not including those that received a "not sure" rating). A positive testing effect is evident for both easy and hard questions, and read and unread passages, although the effect declined with the number of distractors on the prior multiple-choice test for the unread items. In both experiments, a positive testing effect was observed under all conditions, and the effect was maintained even when "not sure" responses were removed.

	Number of previous alternatives			
	Zero (not-tested)	*Two*	*Three*	*Four*
Experiment 1				
Easy questions	.69	.86	.84	.84
	(.65)	(.83)	(.81)	(.80)
Hard questions	.18	.44	.41	.38
	(.16)	(.40)	(.36)	(.33)
Experiment 2				
Read passages	.56	.79	.79	.75
	(.52)	(.73)	(.72)	(.70)
Non-read passages	.23	.56	.53	.44
	(.14)	(.43)	(.40)	(.31)

Table 1.6 Proportion target incorrect answers on the cued recall test as a function of question difficulty and number of alternatives (including the correct answer) on the prior multiple-choice test. Non-guess responses are in parentheses (proportion correct not including those that received a "not sure" rating). The prior multiple-choice test led to more errors on the final test and this effect grew larger when more distractors had been presented on the multiple-choice test. Removing the "not sure" responses reduced the size of the negative suggestibility effect, but left the basic pattern intact.

	Number of previous alternatives			
	Zero (not-tested)	*Two*	*Three*	*Four*
Experiment 1				
Easy questions	.08	.09	.11	.11
	(.05)	(.06)	(.09)	(.08)
Hard questions	.17	.26	.34	.36
	(.10)	(.20)	(.21)	(.24)
Experiment 2				
Read passages	.05	.09	.10	.11
	(.01)	(.06)	(.05)	(.07)
Non-read passages	.11	.24	.25	.37
	(.03)	(.13)	(.12)	(.15)

The data show clearly that taking a multiple-choice test can simultaneously enhance performance on a later cued recall test (a positive testing effect) and harm performance (a negative suggestibility effect). The former effect comes from questions answered correctly on the multiple-choice test, whereas the latter effect arises from errors committed on the multiple-choice test. In fact, 78% of the multiple-choice lure answers on the final test had been selected erroneously on the prior multiple-choice test. This result is noteworthy because it suggests that any negative effects of multiple-choice testing require selection of an incorrect answer, and that simply reading the lures (and then selecting the correct answer) is not problematic.

In Experiment 2 we examined whether students would show the same effects when learning from prose materials. We used 20 nonfiction passages on a wide variety of topics (the sun, Mt. Rainier, Louis Armstrong). For each passage we constructed four questions, each of which could be tested in both multiple-choice and short answer formats. Students read half the passages and not the other half, and then took a multiple-choice test where the number of multiple-choice lures was manipulated from zero (the item was not tested) through three alternatives. Thus, the design conformed to a 2 (studied, non-studied passages) × 4 (number of distractors on the test, 0–3) design. Five minutes after completing the multiple-choice test, the students took the final short answer test that contained 80 critical items (60 from the previous multiple-choice test and 20 previously nontested items). As in the first experiment, they were required to answer all questions and to rate their confidence in each answer.

The multiple-choice data are displayed in the bottom part of Table 1.4. Again, the results are straightforward: Not surprisingly, performance on unread passages was lower than for read passages, and performance generally declined with the number of distractors (albeit more for unread than read passages). Once again, the manipulations succeeded in varying multiple-choice performance across a fairly wide range.

The consequences of multiple-choice testing can be seen in the bottom of Table 1.5, which shows the proportion of final short answer questions answered correctly. A positive testing effect occurred for both read and unread passages, although for unread passages the effect declined with the number of distractors on the prior multiple-choice test. Still, as in the first experiment, a positive testing effect was observed in all conditions, even when "not sure" responses were removed (the data in parentheses).

As can be seen in Table 1.6, the negative suggestibility effect also appeared in full force in Experiment 2, although it was greater for the nonread passages, with their corresponding higher rate of errors on the multiple-choice test than for the read passages. For the read passages, the error rate nearly doubled after the multiple-choice test, from 5% to 9%. When the "not sure" responses were removed, the difference grew from 1% (not tested) to 6% (tested). This is not large, but is statistically significant. Also, note that students in this experiment were tested under conditions that usually do not hold in actual educational settings – they had carefully read the relevant passages only moments before the multiple-choice test and only 20 or so minutes before taking the final criterial test. The data from the nonread passages with their higher error rates may in some ways be more educationally informative, as unfortunately students are often unprepared for exams. The negative suggestibility effect was much larger in the nonread condition whether or not the "not sure" responses were included.

The generalization that may be taken from these two experiments is that when multiple-choice performance is relatively high, a large positive testing effect and a relatively small negative suggestibility effect will be found. Correspondingly, under conditions of relatively poor multiple-choice performance, the positive testing effect will be diminished and the negative effect will be increased. These conclusions hold over more recent experiments, and also agree with the third experiment conducted, the one that did appear in the Roediger and Marsh (2005) paper. In this study, we replicated Experiment 2 but changed the instruction on the final short answer test from forced recall with confidence ratings to recall with a strong warning against guessing. That is, subjects were told to give an answer on the final test only if they were reasonably sure they were right. Under these stricter conditions, we still obtained the negative suggestibility effect (and, of course, the positive testing effect). These strict conditions are unlike those used in educational settings, though, where students are typically free to respond without fear of penalty for wrong answers.

The subjects in the three experiments just described were all from

Washington University, a highly selective university, and therefore these expert test takers are unrepresentative of test takers in general. To take a step towards a more typical sample, we (Marsh, Agarwal, & Roediger, 2009) recently tested high school juniors at a suburban high school in Illinois, on practice SAT II questions on chemistry, biology, and history. SAT II tests (now renamed SAT subject tests) are often taken by high school juniors and used for college admissions and class placement decisions. We examined the effects of answering SAT II questions on a later short answer test, in both the high school students and a Duke University sample tested with similar procedures. Duke University students are expert test takers, who took tests similar to the SAT II for admission. Not surprisingly, the undergraduates did much better on the initial multiple-choice questions than did the high school students; undergraduates answered 55% correctly whereas high schoolers only answered 34% correctly. High schoolers also endorsed far more multiple-choice lures than did the university students; SAT II questions always offer a "don't know" option as the test penalizes wrong answers. So even if high school students didn't know the answers, they still could have responded "don't know" rather than endorsing a distractor – but they endorsed distractors for 56% of the multiple-choice questions! The results on the final short answer test were consistent with what we predicted – the negative testing effect was much larger in the group (high school students) who endorsed more multiple-choice lures. Testing led to a smaller positive testing effect in high school students, and a larger negative testing effect, emphasizing the need for future research to include populations other than undergraduates.

None of the experiments described thus far in this section provided any corrective feedback to students. Importantly, Butler and Roediger (2008) showed that feedback given shortly after a multiple-choice test enhanced the positive testing effect and neutralized the negative suggestibility effect. However, it is critical that feedback be provided under conditions in which it is carefully processed to have this positive impact (see too Butler et al., 2007). Giving feedback is thus one obvious method of preventing the negative suggestibility effect. However, in our experience feedback is rarely given in university and college settings and when provided it occurs under suboptimal conditions. In large introductory courses using multiple-choice and short answer tests, professors often want to protect items in their test banks (so they do not have to create new tests and can refine their old tests with the data students provide). Even when professors do give feedback on tests, it is often given relatively long after taking the test (due to time for grading) and/or the feedback is provided under conditions in which students may not attend to it (e.g., just giving back the marked tests or requiring students to stop by the professor's office to see the corrected tests).

Most professors we know give credit (and partial credit) as deserved, but do not deduct points for bad answers – the worst possible score on an item is a zero, not some negative number. However, giving a penalty for wrong

answers sounds more interesting when one thinks about the importance of endorsing multiple-choice lures for the negative suggestibility effect. We examined this more directly in another experiment using SAT II practice questions and Duke undergraduates (Marsh et al., 2009). One group of undergraduates was warned they would receive a penalty for wrong answers and that they should choose a "don't know" option if they were not reasonably sure of their answer. Another group was required to answer all of the multiple-choice questions. Both groups showed large positive testing effects, and smaller negative testing effects. Critically, the penalty instruction significantly reduced the negative testing effect, although it was still significant.

Research on negative suggestibility is just beginning, and only a few variables have been systematically investigated. Three classes of variables are likely to be interesting: ones that affect how likely subjects are to select multiple-choice lures (e.g., reading related material, a penalty for wrong answers on the MC test), ones that affect the likelihood that selected multiple-choice lures are integrated with related world knowledge (e.g., corrective feedback), and ones that affect monitoring at test (e.g., the warning against guessing on the final test used in Roediger & Marsh, 2005). The negative testing effect could change in size for any of these reasons. For example, consider one recent investigation involving the effects of adding a "none of the above" option to the MC test (Odegard & Koen, 2007). When "none of the above" was the correct answer on the MC test, the negative testing effect increased. It turned out that subjects were less willing to endorse "none of the above" than a specific alternative, meaning that MC performance was worst for items containing a "none of the above" option (and MC lure endorsements increased), with consequences for later performance.

In summary, most experiments show that the positive testing effect is larger than any negative testing effect; even if subjects learn some false facts from the test, the net effect of testing is positive (see Marsh, Roediger, Bjork, & Bjork, 2007, for a review). The exception may be when students are totally unprepared for the test and endorse many multiple-choice lures – this is a scenario that needs further research. The best advice we can give is to make sure the students receive corrective feedback, and to consider penalizing students for wrong answers.

METACOGNITION AND SELF-REGULATED LEARNING

While we have focused on testing as a pedagogical tool to enhance learning in the classroom, we are mindful that the bulk of learning in real life takes place outside the classroom. More often than not learning is self-regulated – the learner has to decide what information to study, how long to study, the kind of strategies or processing to use when studying, and so on. All these decisions depend on the learner's goals (e.g., the desired level of mastery), beliefs (e.g., that a particular type of study strategy is more effective), external

constraints (e.g., time pressure), and online monitoring during the learning experience (i.e., subjective assessments of how well the material has been learned; Benjamin, 2007). In other words, a student's beliefs about learning and memory and his or her subjective evaluations during the learning experience are vital to effective learning (Dunlosky, Hertzog, Kennedy, & Thiede, 2005). In this section we shall discuss the metacognitive factors concomitant with testing, how testing can improve monitoring accuracy, as well as the use of self-testing as a study strategy by students.

Research on metacognition provides a framework for examining how students strategically monitor and regulate their learning. *Monitoring* refers to a person's subjective assessment of their cognitive processes, and *control* refers to the processes that regulate behavior as a consequence of monitoring (Nelson & Narens, 1990). One indirect way in which testing can enhance future learning is by allowing students to better monitor their learning (i.e., discriminate information that has been learned well from that which has not been learned). Enhanced monitoring in turn influences subsequent study behavior, such as having students channel their efforts towards less well-learned materials. A survey of college students' study habits revealed that students are generally aware of this function of testing (Kornell & Bjork, 2007). In response to the question "If you quiz yourself while you study, why do you do so?" 68% of respondents chose "To figure out how well I have learned the information I'm studying," while only 18% selected "I learn more that way than through rereading," suggesting that relatively few students view testing as a learning event (see too Karpicke, Butler, & Roediger, 2009).

To gain insight into subjects' monitoring abilities, researchers ask them to make judgments of learning (JOLs). Normally done during study, students predict their ability to remember the to-be-learned information at a later point in time (usually on a scale of 0–100%), and then these predictions are compared to their actual performance. Usually people are moderately accurate when making these predictions in laboratory paradigms (e.g., Arbuckle & Cuddy, 1969), but JOLs are inferential in nature and can be based on a variety of beliefs and cues (Koriat, 1997). The accuracy of one's metacognitive monitoring depends on the extent to which the beliefs and cues that one uses are diagnostic of future memory performance – and some of students' beliefs about learning are wrong. For example, subjects believe that items that are easily processed will be easy to retrieve later (e.g., Begg, Duft, Lalonde, Melnick, & Sanvito, 1989), whereas we have already discussed that more effortful retrieval is more likely to promote retention. Similarly, students tend to give higher JOLs after repeated study than after receiving initial tests on the to-be-remembered material, but actual final memory performance exhibits the opposite pattern (i.e., the testing effect; Agarwal et al., 2008; Kang, 2009a; Roediger & Karpicke, 2006a). Repeated studying of the material probably engenders greater processing fluency, which leads to an overestimation of one's future memory performance.

Students' incorrect beliefs about memory mean that they often engage in

suboptimal learning strategies. For example, JOLs are often negatively correlated with study times during learning, meaning that students spend more time studying items that they feel are difficult and that they still need to master (Son & Metcalfe, 2000; although see Metcalfe & Kornell [2003] for conditions that produce an exception to this generalization). Not only is testing a better strategy, but sometimes substantial increases in study time are not accompanied by equivalent increases in performance, an outcome termed the "labor-in-vain" effect (Nelson & Leonesio, 1988).

Consider a study by Karpicke (in press) that examined subjects' strategies for learning Swahili–English word pairs. Critically, the experiment had repeated study–test cycles (multi-trial learning) and once subjects were able to correctly recall the English word (when cued with the Swahili word) they were given the choice of whether to restudy, test, or drop an item for the upcoming trial, with the goal of maximizing performance on a final test 1 week later. Subjects chose to drop the majority of items (60%), while about 25% and 15% of the items were selected for repeated testing and restudy, respectively. Subjects also made JOLs before making each choice, and items selected for restudy were subjectively the most difficult (i.e., lowest JOLs), dropped items were perceived to be the easiest, and items selected for testing were in between. As expected, final performance increased as a function of the proportion of items chosen to be tested, whereas there was no relationship between the proportion of items chosen for restudy and final recall. Finally, there was a negative correlation between the proportion of items dropped and final recall, indicating that subjects dropped items before they had firmly registered the pair.

These results suggest that learners often make suboptimal choices during learning, opting for strategies that do not maximize subsequent retention. Also, the tendency to drop items once they were recalled the first time reflects overconfidence and under-appreciation of the value of practicing retrieval. Follow-up research in our lab (Kang, 2009b) is investigating whether experiencing the testing effect (i.e., performing well on a final test for items previously tested, relative to items that were previously dropped or restudied) can induce learners to select preferentially self-testing study strategies that enhance future recall. We suspect this may be possible, given that testing can help improve metacognitive monitoring and sensitize learners to retrieval conditions, as described in the next two experiments.

Comparisons between immediate and delayed JOLs suggest an important role for testing in improving monitoring accuracy. Delayed JOLs refer to ones solicited at some delay after the items have been studied, whereas immediate JOLs are solicited immediately after each item has been studied. Delayed JOLs are typically more accurate than immediate JOLs (e.g., Nelson & Dunlosky, 1991). This delayed JOL effect is obtained only under certain conditions, specifically when the JOLs are "cue-only" JOLs. This term refers to the situation in which studied items are A–B pairs and subjects are provided only with A when asked to make their predictions for later recall of the

target B; the effect does not occur when JOLs are sought with intact cue-target pairs presented (Dunlosky & Nelson, 1992). One explanation for this finding is that subjects attempt retrieval of the target for cue-only delayed JOLs, and success or failure at retrieval then guides subjects' predictions (i.e., a high JOL is given if the target is successfully retrieved; if not then a low JOL is given). This enhanced ability to distinguish well-learned from less well-learned items, coupled with the testing effect on items retrieved successfully during the delayed JOL, has been proposed to account for the increased accuracy of delayed JOLs (Spellman & Bjork, 1992; Kelemen & Weaver, 1997).

Testing can also augment monitoring accuracy by sensitizing learners to the actual conditions that prevail at retrieval. Consider one study where JOLs were not always accurate: Koriat and Bjork (2005) had subjects learn paired associates, including forward-associated pairs (e.g., *cheddar–cheese*), backwards-associated pairs (e.g., *cheese–cheddar*), and unrelated pairs. During learning, subjects were asked to judge how likely it was that they would remember the 2nd word in the pair. Subjects over-predicted their ability to remember the target in the backwards-associated pairs, and the authors dubbed this an "illusion of competence." When subjects see "*cheese–cheddar*" they think they will easily remember *cheddar* when they later see *cheese* because the two words are related. However, when *cheese* occurs on the later test, it does not cue *cheddar* because the association between them is asymmetric and occurs in the opposite direction (*cheddar* reminds people of *cheese*, but *cheese* is much less likely to remind people of *cheddar*). Castel, McCabe, and Roediger (2007) reported the same overconfidence in students believing they would remember identical pairs of words (*cheese–cheese*). Critically, Koriat and Bjork (2006) found that study–test experience could alleviate this metacognitive illusion. On the first study–test cycle, subjects showed the same overconfidence for the backward-associated pairs, but JOLs and recall performance became better calibrated with further study–test opportunities. This finding suggests that prior test experience can enhance learners' sensitivity to retrieval conditions on a subsequent test, and can be a way to improve metacognitive monitoring.

The studies discussed in this section converge on the general conclusion that the majority of college students are unaware of the mnemonic benefit of testing: When students monitor their learning they feel more confident after repeated reading than after repeated testing, and when allowed to choose their study strategies self-testing is not the dominant choice. Students' self-reported study strategies mirror these laboratory findings (Karpicke et al., 2009). When preparing for tests, rereading notes or textbooks was by far the most preferred strategy (endorsed by almost 55% of students completing the survey), whereas strategies that involve retrieval practice (e.g., doing practice problems, using flashcards, self-testing) were preferred by fewer than 20% of students. It is clear that, when left to their own devices, many students engage in suboptimal study behavior. Even though college students might be

expected to be expert learners (given their many years of schooling and experience preparing for exams), they often labor in vain (e.g., rereading the text) instead of employing strategies that contribute to robust learning and retention. Self-testing may be unappealing to many students because of the greater effort required compared to rereading, but this difficulty during learning turns out to be beneficial for long-term performance (Bjork, 1994). Therefore, the challenge for future research is to uncover conditions that encourage learners to set aside their naïve intuitions when studying and opt for retrieval-based strategies that yield lasting results.

APPLICATIONS OF TESTING IN CLASSROOMS

Recently, several journal articles have highlighted the importance of using tests and quizzes to improve learning in real educational situations. The notion of using testing to enhance student learning is not novel, however, as Gates employed this practice with elementary school students in 1917 (see too Jones [1923] and Spitzer [1939]). One cannot, however, assume that laboratory findings necessarily generalize to classroom situations, given that some laboratory parameters (e.g., relatively short retention intervals, tight experimental control) do not correspond well to naturalistic contexts. This distinction has garnered interest recently and we will outline a few studies that have evaluated the efficacy of test-enhanced learning within a classroom context.

Leeming (2002) adopted an "exam-a-day" procedure in two sections of Introductory Psychology and two sections of his summer Learning and Memory course, for a total of 22 to 24 exams over the duration of the courses. In comparable classes taught in prior semesters, students had received only four exams. Final retention was measured after 6 weeks. Leeming found significant increases in performance between the exam-a-day procedure and the four exam procedure in both the Introductory Psychology sections (80% vs. 74%) and Learning and Memory sections (89% vs. 81%). In addition, the percentage of students who failed the course decreased following the exam-a-day procedure. Leeming's students also participated in a survey, and students in the exam-a-day sections reported increased interest and studying for class.

McDaniel et al. (2007a) described a study in an online Brain and Behavior course that used two kinds of initial test questions, short answer and multiple-choice, as well as a read-only condition. Weekly initial quizzes were administered via the Web; questions were followed by immediate feedback and in the read-only condition, the facts were re-presented. Two unit examinations in multiple-choice format were given after 3 weeks of quizzes, and a final cumulative multiple-choice assessment at the end of the semester covered material from both units. Although facts targeted on the initial quizzes were repeated on the unit/final exams, the question stems were

phrased differently so that the learning of concepts was assessed rather than memory for a prior test response. On the two unit exams, retention for quizzed material was significantly greater than that for non-quizzed material, regardless of the initial quiz format. On the final exam, however, only short answer (but not multiple-choice) initial quizzes produced a significant benefit above non-quizzed and read-only material. The results from this study provide further evidence of the strength of the testing effect in classroom settings, as well as replicating prior findings showing that short answer tests produce a greater testing effect than do multiple-choice tests (e.g., Butler & Roediger, 2007; Kang et al., 2007).

McDaniel and Sun (2009) replicated these findings in a more traditional college classroom setting, in which students took two short-answer quizzes per week. The quizzes were emailed to the students, who had to complete them by noon the next day. After emailing their quiz back to the professor, students received an email with the quiz questions and correct answers. Retention was measured on unit exams, composed of quizzed and non-quizzed material, and administered at the end of every week. Performance for quizzed material was significantly greater than performance on non-quizzed material.

Finally, Roediger, McDaniel, McDermott, and Agarwal (2010) conducted various test-enhanced learning experiments at a middle school in Illinois. The experiments were fully integrated into the classroom schedules and used material drawn directly from the school and classroom curriculum. In the first study, 6th grade social studies, 7th grade English, and 8th grade science students completed initial multiple-choice quizzes over half of the classroom material. The other half of the material served as the control material. The teacher in the class left the classroom during administration of quizzes, so she did not know the content of the quizzes and could not bias her instruction toward (or against) the tested material. The initial quizzes included a pre-test before the teacher reviewed the material in class, a post-test immediately following the teacher's lecture, and a review test a few days after the teacher's lecture. Upon completion of a 3- to 6-week unit, retention was measured on chapter exams composed of both quizzed and non-quizzed material. At all three grade levels, and in all three content areas, significant testing effects were revealed such that retention for quizzed material was greater than for non-quizzed material, even up to 9 months later (at the end of the school year). The results from 8th grade science, for example, can be seen in Figure 1.7.

This experiment was replicated with 6th grade social studies students who, instead of completing in-class multiple-choice quizzes, participated in games online using an interactive website at their leisure. This design was implemented in order to minimize the amount of class time required for a test-enhanced learning program. Despite being left to their own devices, students still performed better on quizzed material available online than non-quizzed material on their final chapter exams. Furthermore, in a subsequent

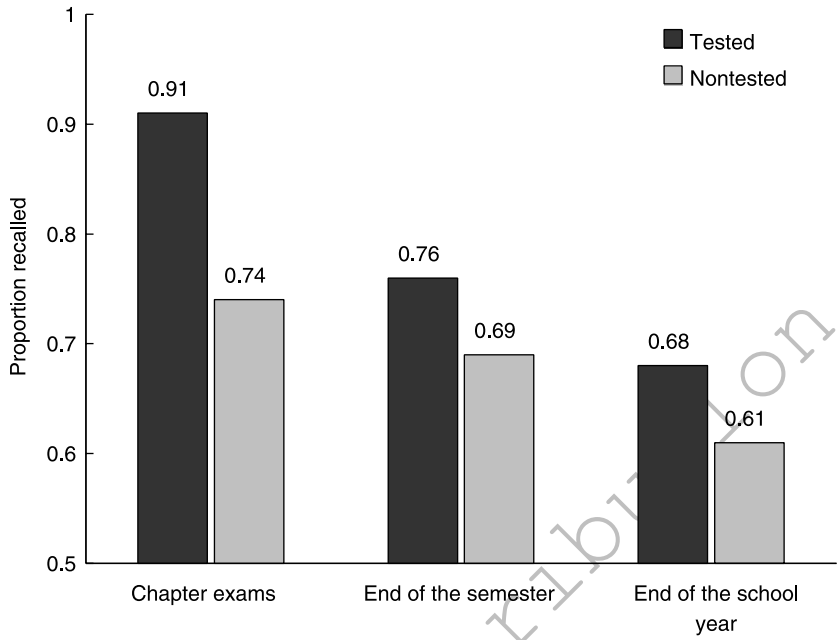

Figure 1.7 Science results from Roediger, McDaniel, McDermott, and Agarwal (2009). Significant testing effects in a middle school setting were revealed such that retention for quizzed material was greater than for non-quizzed material, even up to 9 months later (at the end of the school year).

experiment with 6th grade social studies students, a read-only condition was included, and performance for quizzed material was still significantly greater than read-only and non-quizzed material, even when the number of exposures were equated between the quizzed and read-only condition.

In sum, recent research is beginning to demonstrate the robust effects of testing in applied settings, including middle school and college classrooms. Future research extending to more content areas (e.g., math), age groups (e.g., elementary school students), methods of quizzing (e.g., computer-based and online), and types of material (e.g., application and transfer questions), we expect, will only provide further support for test-enhanced learning programs.

CONCLUSION

In this chapter, we have reviewed evidence supporting test-enhanced learning in the classroom and as a study strategy (i.e., self-testing) for improving student performance. Frequent classroom testing has both indirect and direct benefits. The indirect benefits are that students study for more time and with

greater regularity when tests are frequent, because the specter of a looming test encourages studying. The direct benefit is that testing on material serves as a potent enhancer of retention for this material on future tests, either relative to no activity or even relative to restudying material. Providing correct answer feedback on tests and insuring that students carefully process this feedback greatly enhances this testing effect. Feedback is especially important when initial test performance is low. Multiple tests produce a larger testing effect than does a single test. In addition, tests requiring production of answers (short answer or essay tests) produce a greater testing effect than do recognition tests (multiple-choice or true/false). The latter tests also have the disadvantage of exposing students to erroneous information, but giving feedback eliminates this problem. Test-enhanced learning is not limited to laboratory materials; it improves performance with educational materials (foreign language vocabulary, science passages) and in actual classroom settings (ranging from middle school classes in social studies, English, and science, to university classes in introductory psychology and biological bases of behavior). We believe that the application of frequent testing in classrooms can greatly improve academic performance across the curriculum.

References

Abbott, E. E. (1909). On the analysis of the factors of recall in the learning process. *Psychological Monographs*, *11*, 159–177.

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L., & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology*, *22*, 861–876.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*, 126–131.

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 241–252.

Balota, D. A., Duchek, J. M., & Logan, J. M. (2007). Is expanded retrieval practice a superior form of spaced retrieval? A critical review of the extant literature. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger, III* (pp. 83–105). New York: Psychology Press.

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

Begg, I., Armour, V., & Kerr, T. (1985). On believing what we remember. *Canadian Journal of Behavioral Science*, *17*, 199–214.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*, 610–632.

Benjamin, A. S. (2007). Memory is more than just remembering: Strategic control of encoding, accessing memory, and making decisions. In A. S. Benjamin & B. H. Ross (Eds.), *The psychology of learning and motivation: Skill and strategy in memory use* (Vol. 48, pp. 175–223). London: Academic Press.

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). New York: Wiley.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Brown, A. S. (1988). Encountering misspellings and spelling performance: Why wrong isn't right. *Journal of Educational Psychology*, *4*, 488–494.

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

Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 918–928.

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

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory and Cognition*, *36*, 604–616.

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.

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin and Review*, *14*, 107–111.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

Dudai, Y. (2006). Reconsolidation: The advantage of being refocused. *Current Opinion in Neurobiology*, *16*, 174–178.

Dunlosky, J., Hertzog, C., Kennedy, M. R. T., & Thiede, K. W. (2005). The self-monitoring approach for effective learning. *Cognitive Technology*, *10*, 4–11.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory and Cognition*, *20*, 374–380.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin and Review*, *16*, 88–92.

Feller, M. (1994). Open-book testing and education for the future. *Studies in Educational Evaluation*, *20*, 235–238.

Ferster, C. B., & Skinner, B. F. (1957). *Schedules of reinforcement*. New York: Appleton-Century-Crofts.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*(40).

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

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112.

Jacoby, L. L., & Hollingshead, A. (1990). Reading student essays may be hazardous to

your spelling: Effects of reading incorrectly and correctly spelled words. *Canadian Journal of Psychology*, *44*, 345–358.

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

Kang, S. H. K. (2009a). Enhancing visuo-spatial learning: The benefit of retrieval practice. Manuscript under revision.

Kang, S. H. K. (2009b). The influence of text expectancy, test format and test experience on study strategy selection and long-term retention. Unpublished doctoral dissertation, Washington University, St Louis, MO, USA.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Karpicke, J. D. (in press). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*.

Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*, 471–479.

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

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

Kelemen, W. L., & Weaver, C. A. (1997). Enhanced memory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1394–1409.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*, 349–370.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one's knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 187–194.

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory and Cognition*, *34*, 959–972.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin and Review*, *14*, 219–224.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learninig, Memory, and Cognition*, *35*, 989–998.

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

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

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

Landauer, T. K., & Bjork, R. A. (1978). Optimal rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Harris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). New York: Academic Press.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in Psychology classes. *Teaching of Psychology*, *29*, 210–212.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19–31.

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*, 257–280.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007a). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*, 494–513.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007b). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin and Review*, *14*, 200–206.

McDaniel, M. A., & Sun, J. (2009). The testing effect: Experimental evidence in a college course. Manuscript under revision.

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., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin and Review*, *14*, 194–199.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*, 530–542.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, *32*, 99–113.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect." *Psychological Science*, *2*, 267–270.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the "labor-in-vain effect." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 676–686.

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 and Verbal Behavior*, *19*, 338–368.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). New York: Academic Press.

Odegard, T. N., & Koen, J. D. (2007). "None of the above" as a correct and incorrect alternative on a multiple-choice test: Implications for the testing effect. *Memory*, *15*, 873–885.

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.

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning: Journal of Practical Research and Applications*, *4*, 11–18.

Remmers, H. H., & Remmers, E. M. (1926). The negative suggestion effect on true-false examination questions. *Journal of Educational Psychology*, *17*, 52–56.

Richland, L. E., Kornell, N. & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied*, *15*, 243–257.

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

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

Roediger, H. L., McDaniel, M. A., McDermott, K. B., & Agarwal, P. K. (2010). Test-enhanced learning in the classroom: The Columbia Middle School project. Manuscript in preparation.

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.

Roediger, H. L., Zaromb, F. M., & Butler, A. C. (2008). The role of repeated retrieval in shaping collective memory. In P. Boyer and J. V. Wertsch (Eds.), *Memory in Mind and Culture* (pp. 29–58). Cambridge: Cambridge University Press.

Ruch, G. M. (1929). *The objective or new-type examination: An introduction to educational measurement*. Chicago: Scott, Foresman, and Co.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Skinner, B. F. (1958). Teaching machines. *Science*, *128*, 969–977.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 204–221.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, *3*, 315–316.

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.

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.

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