The Power of Testing Memory

Basic Research and Implications for Educational Practice

Henry L. Roediger, III, and Jeffrey D. Karpicke

Washington University in St. Louis

ABSTRACT—*A powerful way of improving one's memory for material is to be tested on that material. Tests enhance later retention more than additional study of the material, even when tests are given without feedback. This surprising phenomenon is called the testing effect, and although it has been studied by cognitive psychologists sporadically over the years, today there is a renewed effort to learn why testing is effective and to apply testing in educational settings. In this article, we selectively review laboratory studies that reveal the power of testing in improving retention and then turn to studies that demonstrate the basic effects in educational settings. We also consider the related concepts of dynamic testing and formative assessment as other means of using tests to improve learning. Finally, we consider some negative consequences of testing that may occur in certain circumstances, though these negative effects are often small and do not cancel out the large positive effects of testing. Frequent testing in the classroom may boost educational achievement at all levels of education.*

In contemporary educational circles, the concept of testing has a dubious reputation, and many educators believe that testing is overemphasized in today's schools. By "testing," most commentators mean using standardized tests to assess students. During the 20th century, the educational testing movement produced numerous assessment devices used throughout education systems in most countries, from prekindergarten through graduate school. However, in this review, we discuss primarily the kind of testing that occurs in classrooms or that students engage in while studying (self-testing). Some educators argue that testing in the classroom should be minimized, so that valuable time will not be taken away from classroom instruction. The nadir of testing occurs in college classrooms. In many universities, even the most basic courses have very few tests, and classes with only a midterm exam and a final exam are common. Students do not like to take tests, and teachers and professors do not like to grade them, so the current situation seems propitious to both parties.

The traditional perspective of educators is to view tests and examinations as assessment devices to measure what a student knows. Although this is certainly one function of testing, we argue in this article that testing not only measures knowledge, but also changes it, often greatly improving retention of the tested knowledge. Taking a test on material can have a greater positive effect on future retention of that material than spending an equivalent amount of time restudying the material, even when performance on the test is far from perfect and no feedback is given on missed information. This phenomenon of improved performance from taking a test is known as the *testing effect*, and though it has been the subject of many studies by experimental psychologists, it is not widely known or appreciated in education. We believe that the neglect of testing in educational circles is unfortunate, because testing memory is a powerful technique for enhancing learning in many circumstances.

The idea that testing (or recitation, as it is sometimes called in the older literature) improves retention is not new. In 1620, Bacon wrote: "If you read a piece of text through twenty times, you will not learn it by heart so easily as if you read it ten times while attempting to recite from time to time and consulting the text when your memory fails" (F. Bacon, 1620/2000, p. 143). In the *Principles of Psychology*, James (1890) also argued for the power of testing or active recitation:

> A curious peculiarity of our memory is that things are impressed better by active than by passive repetition. I mean that in learning (by heart, for example), when we almost know the piece, it pays better to wait and recollect by an effort from within, than to look at the book again. If we recover the words in the former way, we shall probably know them the next time; if in the latter way, we shall very likely need the book once more. (p. 646)

Bacon and James were describing situations in which students test themselves while studying. We show later that their hypotheses are correct and that testing greatly improves retention of material. However, we need to make a distinction between two

Address correspondence to Henry L. Roediger, III, or to Jeffrey D. Karpicke, Department of Psychology, Box 1125, Washington University in St. Louis, One Brookings Dr., St. Louis, MO 63130-4899, e-mail: roediger@wustl.edu or karpicke@wustl.edu.

types of effects that testing might have on learning: mediated (or indirect) effects and direct (unmediated) effects. Let us consider mediated effects first, because testing can enhance learning in a variety of ways. To give just a few examples, frequent testing in classrooms encourages students to study continuously throughout a course, rather than bunching massive study efforts before a few isolated tests (Fitch, Drucker, & Norton, 1951). Tests also give students the opportunity to learn from the feedback they receive about their test performance, especially when that feedback is elaborate and meaningful, as is the case in the technique of formative assessment, discussed in a later section. In addition, if students test themselves periodically while they are studying (as Bacon and James advocated long ago), they may use the outcome of these tests to guide their future study toward the material they have not yet mastered. The facts that testing encourages students to space their studying and gives them feedback about what they know and do not know are good reasons to recommend frequent testing in courses, but they are not the primary reasons we focus on in this article. In these cases of mediated effects of testing, it is not the act of taking the test itself that influences learning, but rather the fact that testing promotes learning via some other process or processes. For example, when a test provides feedback about whether or not students know particular items and the students guide their future study efforts accordingly, testing promotes learning by making later studying or encoding more effective; thus, testing enhances learning by means of this mediating process.

These examples of mediated effects of testing serve as additional evidence in favor of the use of frequent testing in education. However, our review is focused on direct effects of testing on learning—the finding that the act of taking a test itself often enhances learning and long-term retention. In many of the experiments we describe, one group of students studied some set of materials and then was given an initial test (or sometimes repeated tests). Retention of the material was assessed on a final criterial test, and the tested group's performance was compared with that of one or two control groups. In one type of control, students studied the material and took the final test just as the tested group did, but were not given an initial test. In a second type of control (a restudy control), students studied the material just as the tested group did, but then studied the material a second time when the tested group received the initial test; in this case, total exposure time to the material was equated for the tested and control groups. The typical finding throughout the literature is that the tested group outperforms both kinds of control groups (the no-test control and the restudy control) on the final test, even when no feedback is given after the initial test. In variations on this prototypical experiment, the effects of several variables have been investigated (e.g., the materials to be learned, the format of the initial and final tests, whether or not subjects receive feedback on the first test, the time interval between studying and initial testing, and the retention interval before the final test, to name but a few). As we show, across a wide variety of contexts, the testing effect remains a robust phenomenon.

The direct effects of testing are especially surprising when exposure time is equated in the tested and study conditions, because although the repeated-study group experiences the entire set of materials multiple times, the students in the tested group can experience on the test only what they are able to produce, at least when the test involves recall. Yet despite the differences in initial exposure favoring the study group, the tested group performs better in the long term. That the testing effect is so counterintuitive helps explain why it remains unknown in education. The direct effects of testing on learning are not purely a result of additional exposure to the material, which indicates that processes other than additional studying are responsible for them. The testing effect represents a conundrum, a small version of the Heisenberg uncertainty principle in psychology: Just as measuring the position of an electron changes that position, so the act of retrieving information from memory changes the mnemonic representation underlying retrieval—and enhances later retention of the tested information.

In this article, we review research from both experimental and educational psychology that provides strong evidence for the direct effect of testing in promoting learning. After presenting two classic studies, we consider evidence from laboratories of experimental psychologists who have investigated the testing effect. As is the experimentalists' predilection, they have typically used word lists as materials, college students as subjects, and standard laboratory tasks such as free recall and paired-associate learning (see Cooper & Monk, 1976; Richardson, 1985; and Dempster, 1996, 1997, for earlier and somewhat more focused reviews). Effects on later retention are usually quite large and reliable. We next consider studies conducted in more educationally relevant situations. Such studies often use prose passages about science, history, or other topics as the subject matter and investigate the effects of tests more like those found in educational settings (e.g., essay, short-answer, and multiple-choice tests). Once again, we show that testing promotes strong positive effects on long-term retention. We also review studies carried out in actual classrooms using even more complex materials, and they again show positive effects of testing on learning.

After concluding our review of basic research findings, we provide an overview of theoretical approaches that have been directed toward explaining the testing effect, although many puzzles about testing have not been satisfactorily explained. We then consider the related approaches of dynamic testing (e.g., Sternberg & Grigorenko, 2002) and formative assessment (e.g., Black & Wiliam, 1998a), which are both aimed at using tests to promote learning by altering instructional techniques on the basis of the results of tests (i.e., mediated effects of testing). Because testing does not always have positive consequences, we next review two possible negative effects (retrieval interference and negative suggestibility) that need to be considered when using tests as possible learning devices. Finally, we discuss

common objections to increased use of testing in the classroom, and we tell why we believe that none of these objections outweighs our recommendations for frequent testing.

TWO CLASSIC STUDIES

Gates (1917) and Spitzer (1939) published two classic studies showing strong positive effects of testing on retention. Both were rather heroic efforts, and so it is unfortunate that neither is accorded much attention in the contemporary literature. Although other research showing the benefits of testing appeared before Gates's work (e.g., Abbott, 1909; Thorndike, 1914), he carried out the first large-scale study. Gates tested groups of children across a range of grades (Grades 1, 3, 4, 5, 6, and 8), and, admirably, he used two different types of materials (nonsense syllables, the classic stimulus of Ebbinghaus, 1885/1964, and brief biographies taken from *Who's Who in America*). The children studied these materials during a two-phase learning procedure. In the first phase, they simply read the materials to themselves, whereas in the second phase, the experimenter instructed them to look away from the materials and try to recall the information to themselves (covert recitation). During the recitation phase, the students were permitted to glance back at the materials when they needed to refresh their memories. Although this feature of the design relaxed experimental control, it probably faithfully captured what students do when using a recitation or testing strategy to study.

Gates (1917) manipulated the amount of time the children spent reciting by instructing them to stop reading and start reciting after different amounts of study time had elapsed. Different groups of children at each age level spent 0, 20, 40, 60, 80, or 90% of the learning period involved in recitation, or self-testing. Finally, at the end of the period, Gates gave the children a test, asking them to write down as many items as they could in order of appearance. He then retested the children 3 to 4 hr later.

Gates's (1917) basic results are shown in Figure 1, which shows that in almost all conditions, he obtained positive effects of recitation. With nonsense syllables, all groups except first graders showed a strong effect of recitation. For the biographical materials, all groups showed a recitation effect, but one that was less dramatic on the initial tests than on the delayed tests. (Note that first graders were not tested with prose passages because their reading abilities were so poor.) With prose passages, the optimal amount of recitation seemed to be about 60% of the total learning period. Gates concluded that recall attempts during learning (recitation with restudy of forgotten material) are a good way to promote learning. He argued that these results had important implications for educational practice and described ways to incorporate recitation into classroom exercises (Gates, 1917, pp. 99–104). However, Gates's work pointed to limitations of recitation/self-testing, too. First graders did not show the effect, which suggests that it may occur only after a certain point in development. Also, with prose passages, the effect of recitation leveled off and even appeared to drop when the amount of time spent on recitation exceeded 60%, and consequently study time was less than 40%. Thus, the data suggest that a certain amount of study may be necessary before recitation or testing can begin to benefit learning.

A second landmark study showing positive effects of testing was carried out by Spitzer (1939) in his dissertation work. His experiment involved testing the entire population of sixth-grade students in 91 elementary schools in nine Iowa cities—a total of 3,605 students. The students studied 600-word articles (on peanuts or bamboo) that were similar to material they might study in school, and then they took tests according to various schedules across the next 63 days. Each test consisted of 25 multiple-choice items with five alternatives (e.g., "To which family do bamboo plants belong? A) trees, B) ferns, C) grasses, D) mosses, E) fungi"). Some students took a single test 63 days later, whereas others also took earlier tests so that Spitzer could see what effect these would have on later tests. Several interesting patterns could be discerned in the results, which are shown in Figure 2. First, the dashed line shows a beautiful forgetting curve in that the longer the first test was delayed, the worse was performance on that test. Second, giving a test nearly stopped forgetting; when students were given a first test and then retested at a later time, their performance did not drop much at all (and sometimes increased). Third, the sooner the initial test was given after study, the better students did on later tests. For example, Group 2 was tested immediately after study and then a week later. When tested again 56 days later (day 63), they showed much better performance than Group 6 (which was not tested initially until Day 21). In fact, because forgetting had reached asymptote by Day 21, the first test taken by Group 6 did not enhance later recall at all. The lesson from Spitzer's study is that a first test (without feedback) must be given relatively soon after study (when the student still can recall or recognize the material) in order to have a positive effect at a later time.

The studies by Gates (1917) and Spitzer (1939) were among the most extensive in their times (although see Jones, 1923–1924, for another impressive study), and in some features the experimental techniques would not hold up to today's standards. However, the essential points Gates and Spitzer made are secure because later researchers replicated their results. For example, Forlano (1936) replicated Gates's work by demonstrating that testing improved children's learning and spelling of vocabulary words, and Sones and Stroud (1940) replicated Spitzer's (1939) research, albeit on a smaller scale. However, around 1940, interest in the effects of testing on learning seemed to disappear. We can only speculate as to why. One reason may be that with the rise of interference theory (McGeoch, 1942; Melton & Irwin, 1940; see Crowder, 1976, chap. 8), interest swung to the study of forgetting. For the purpose of measuring forgetting, repeated testing was deemed a confound to be avoided because, as Figure 2 shows, an initial test interrupts the course of forgetting. McGeoch (1942, pp. 359–360), Hilgard (1951, p. 557), and

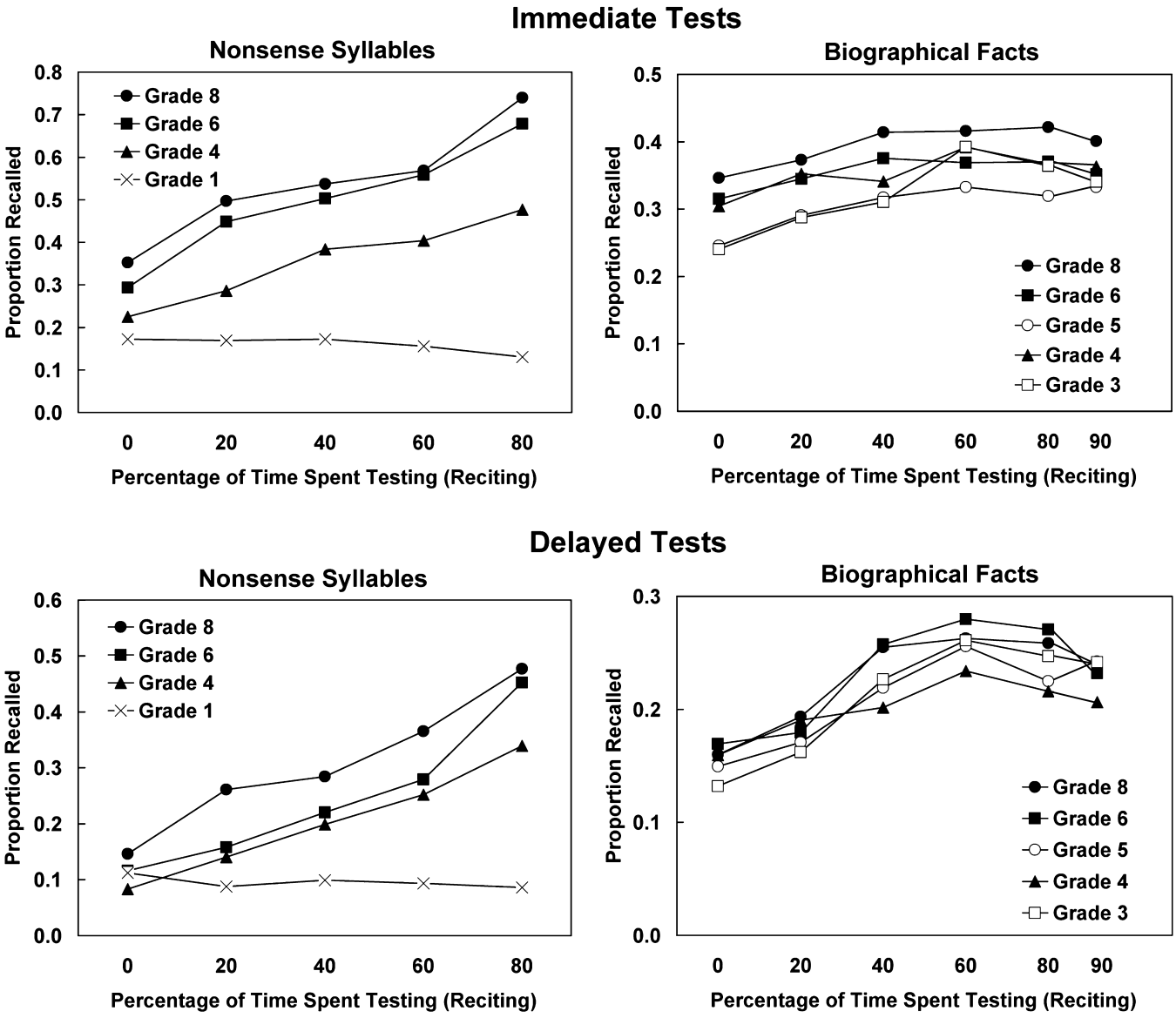

Fig. 1. Proportion of nonsense syllables and biographical facts recalled by children on immediate and delayed tests as a function of the amount of time spent reciting the material. Adapted from data reported by Gates (1917).

Deese (1958) all argued against the use of repeated-testing designs. For example, Deese wrote that "an experimental study of this sort yields very impure measures of retention after the first test, since all subsequent measures are contaminated by the practice the first test allows" (pp. 237–238). This statement is true for the study of forgetting, but of course, for studying the effects of tests per se, repeated testing is necessary, and the "contamination" that Deese referred to is the phenomenon of interest. Nevertheless, leading experimental psychologists' attitude against repeated-testing designs probably halted the study of testing effects (and the study of phenomena such as reminiscence and hypermnesia, which also require repeated testing; W. Brown, 1923; Erdelyi & Becker, 1974; Roediger & Challis, 1989).

TESTS AS AN AID DURING LEARNING

One venerable topic in experimental-cognitive psychology is how and why learning occurs. The traditional way of studying learning is through alternating study and test trials. For example, in multitrial free-recall learning, students typically study a list of words (a study trial), recall as many as possible in any order (a test trial), study the list again, recall it again, and so on through numerous study-test cycles (e.g., Tulving, 1962). When data are averaged across subjects, a regular, negatively accelerated learning curve is produced (e.g., see Fig. 3, which presents results of a study we discuss in the next section).

A controversy about the nature of learning erupted in the late 1950s and early 1960s. Some theorists believed that learning of

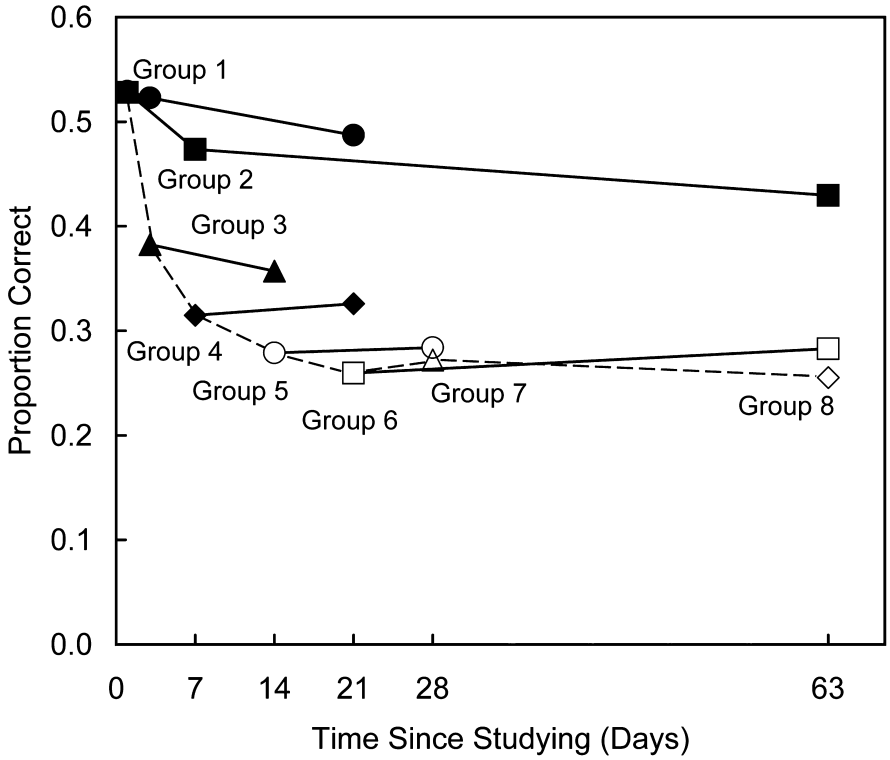

Fig. 2. Proportion correct on multiple-choice tests taken at various delays after studying. After studying the passage, each of the eight groups of subjects was given one, two, or three tests on various schedules across the next 63 days. The solid lines show results for repeated tests for particular groups, and the dashed line represents normal forgetting as the delay between studying and testing increases. Adapted from data reported by Spitzer (1939).

individual items occurs through an incremental process (the standard view), and others argued that learning is all-or-none (Rock, 1957). The incremental-learning position held that each item in the list is represented by a trace that is strengthened a bit by each successive repetition; once enough strength is accrued via repetitions so that some threshold is crossed, an item will be recalled. The all-or-none position held that on each study trial, a subset of items jumps from zero strength to 100% strength in a step function—hence "all or none." In this view, the fact that learning curves appear to be smooth is an artifact of averaging, and performance would actually be all-or-none if the fate of each item could be examined separately. This controversy about the nature of the learning process raged on in some circles throughout the 1950s and into the 1960s and was never completely decided, although the incrementalist assumption is still largely built into today's theories. Tulving (1964) noted that in one sense the controversy was beside the point, because each item in such an experiment is perfectly learned when it is first presented, in the sense that it can be recalled perfectly immediately after its presentation. Thus, learning is always "all," and the critical issue is why students forget items on the subsequent test (i.e., why there is intratrial forgetting).

The reason for bringing up this controversy in the current context is to examine a hidden assumption. Both the incrementalist and the all-or-none positions make the assumption that learning occurs during study trials, when students are exposed to the material, and that the test trials simply permit students to exhibit what they have learned on previous study trials. This is essentially the same attitude that teachers take toward testing in the classroom: Tests simply are assessment devices. An experiment by Tulving (1967) called this assumption into question and helped usher in a new wave of research on testing.

Tulving (1967) had subjects learn lists of 36 words, which were presented in a different random order on every study trial, and then take free-recall tests (subjects recalled out loud as many items as possible in any order, and the experimenter recorded responses). In the standard learning condition, students saw the list, recalled it, saw it, recalled it, and so on for 24 trials. If S stands for a study trial and T stands for a test trial, then the standard condition can be represented as STST STST . . . (for a total of 12 study trials and 12 test trials). Tulving considered every 4 trials a cycle, for reasons that will be clear when the other conditions are described. In the repeated-study condition, each cycle consisted of 3 study trials and 1 test trial (SSST SSST . . .). If subjects learned only during the study trials, then by the end of learning, performance should have been much better in this condition than in the standard condition, because there were 6 more study trials (18 study trials and 6 test trials over the six cycles). In the repeated-test condition, each cycle

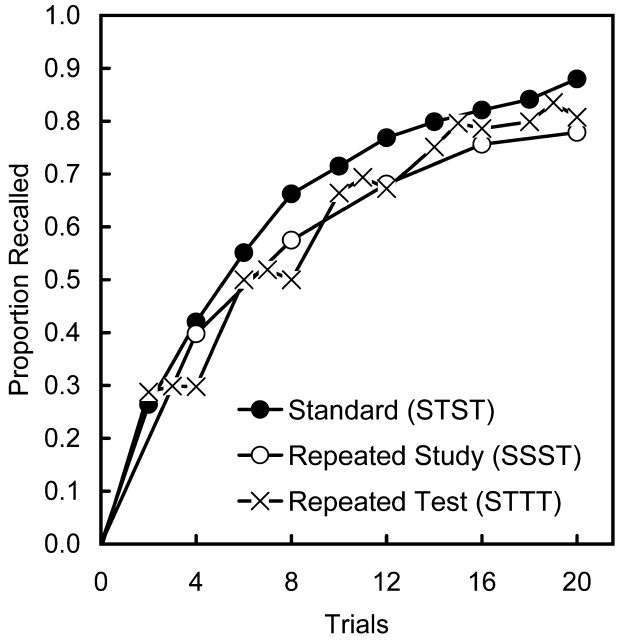

Fig. 3. Proportion of words recalled across trials in standard, repeated-study, and repeated-testing conditions. The shorthand condition labels indicate the order of study (S) and test (T) periods. Data are from Karpicke and Roediger (2006b).

contained 1 study trial followed by 3 consecutive test trials (STTT STTT . . .), leading to a total of only 6 study trials and 18 test trials during the entire learning phase. By the common assumption that learning occurs only during study trials, subjects in the repeated-test condition should have been at a great disadvantage relative to those in the other two conditions.

The surprise in Tulving's (1967) research was that the learning curves of all three conditions looked about the same. For example, by the end of the experiment, subjects recalled about 20 words in the standard and the repeated-study conditions, even though subjects in the repeated-study condition had studied the words six more times. The subjects in the repeated-test condition recalled somewhat fewer words, finishing at about 18.5 words. This slight difference is probably partly explained by the fact that these subjects were deprived of using primary or short-term memory (Glanzer & Cunitz, 1966). That is, subjects in the standard and repeated-study conditions had just heard the list before the very last test trial, so they could use primary memory to recall the last few items. Subjects in the repeated-test condition could not do this, because they had just had two other tests before their last test, and so the short-term component of recall would no longer have been accessible. Given this procedural difference among conditions, it is remarkable that the learning curves of the three conditions were so similar. Apparently, within rather wide limits (6, 12, or 18 study trials), a study trial can be replaced by a test trial. In other words, just as much learning occurs on a test trial as on a study trial. Of course, as a limiting case, there must be some study opportunities before testing can have an effect (as noted by Gates, 1917), but the surprise is how wide the variability is. There were only 6 study trials in the repeated-test condition, and yet final recall was nearly as good as with 18 study trials (in the repeated-study condition). In our own research, which we review later (Karpicke & Roediger, 2006b), we have shown that if long-term retention is measured after a delay, the repeated-test condition actually shows better recall than the repeated-study condition, a finding that is even more counterintuitive given the customary assumptions about the role of study and test trials in learning.

TESTING EFFECTS IN FREE RECALL

Tulving's (1967) results seemed hard to believe when they first appeared, which is probably why so many researchers immediately tried to replicate them with minor variations, creating a boomlet in testing research that lasted briefly in the early 1970s, followed by sporadic work thereafter. In the title of their article, Lachman and Laughery (1968) asked, "Is a test trial a training trial in free recall learning?" and they answered "yes" from their data. Other researchers also replicated Tulving's work, using his conditions or slight variations thereof (Birnbaum & Eichner, 1971; Donaldson, 1971; Rosner, 1970). One methodological detail of Tulving's work and of these replications was unusual. Because Tulving wanted to equate the time of study and test trials, and because he made the presentation rate for words rather fast in the study trials, the duration of the test trials was short. He presented the 36 words at a 1-s rate during study trials, and so he also gave subjects only 36 s to recall the words during test trials. Even with spoken recall, this is a short time to recall 36 words even if they are well learned. In light of later work examining how free recall unfolds over time, tests lasting this long might greatly underestimate the amount of knowledge subjects have acquired (e.g., Roediger & Thorpe, 1978). The short recall time may also explain why subjects were able to recall only about 20 of 36 words after 24 study or test trials; in all probability, they simply did not have time to recall all they knew.

We (Karpicke & Roediger, 2006b) recently conducted an experiment with Tulving's three conditions (standard, repeated-study, and repeated-test), but using 40 words and a 3-s rate of presentation, so that the accompanying tests lasted 2 min and time on study trials and recall tests remained equated. We examined learning curves and compared the conditions on the five common test positions out of the total of 20 study and test trials. That is, every 4th trial was a test trial for all three conditions (standard: STST . . . ; repeated-study: SSST . . . ; and repeated-test: STTT . . .), so we could directly compare recall on the 4th, 8th, 12th, 16th, and 20th trials across the three conditions. We also eliminated short-term memory effects that would normally disadvantage the repeated-test condition by using Tulving and Colotla's (1970) method of separating short-term from long-term memory effects. (Watkins, 1974, concluded that this technique was the best method for this purpose.) Finally, we provided a

delayed test 1 week later to examine lasting effects of the three study schedules on long-term retention.

Our basic results during the learning phase are shown in Figure 3, which indicates recall from secondary memory across tests in the three conditions (Karpicke & Roediger, 2006b). It is clear that subjects in the repeated-test condition were at a disadvantage early in learning (on Trials 4 and 8), but quickly caught up to the repeated-study condition, so that there was little difference between these two conditions later in learning (Trials 12, 16, and 20). However, the standard group performed better than the other two groups over the last four tests (and this difference was statistically significant). Thus, we replicated Tulving's (1967) basic result that learning curves for these three conditions are remarkably similar, although we did find a difference favoring the standard condition. The advantage for the standard condition probably arose because a study trial just after a test trial serves as feedback for what students do not know (they can recognize words they failed to recall and focus their study efforts on these items), and the standard condition had more test trials followed immediately by study trials than the other conditions did. As Izawa (1970) observed, test trials potentiate new learning on the next study trial. We discuss the role of feedback later in this article.

As noted, we (Karpicke & Roediger, 2006b) also measured performance after a 1-week delay. Subjects were given 10 min to recall and at the end of every minute drew a line under the last word recalled, which permitted us to measure how recall cumulates across time (see Wixted & Rohrer, 1994). Figure 4 shows the result, and it is apparent that from the very first minute of the final test period, subjects in the repeated-study condition performed worse than those in the other two conditions. At the end of the recall period, subjects in the standard and repeated-test conditions recalled 68% and 64% of the 40 words, respectively, whereas those in the repeated-study condition recalled only 57% of the words (this was a significant difference from the other two conditions, which did not themselves differ). Thus, despite the fact that the subjects in the repeated-study condition had studied the list 15 times 1 week earlier and those in the repeated-test condition had studied it only 5 times, delayed recall was greater for the latter group. This outcome again shows the power of testing in improving long-term retention.

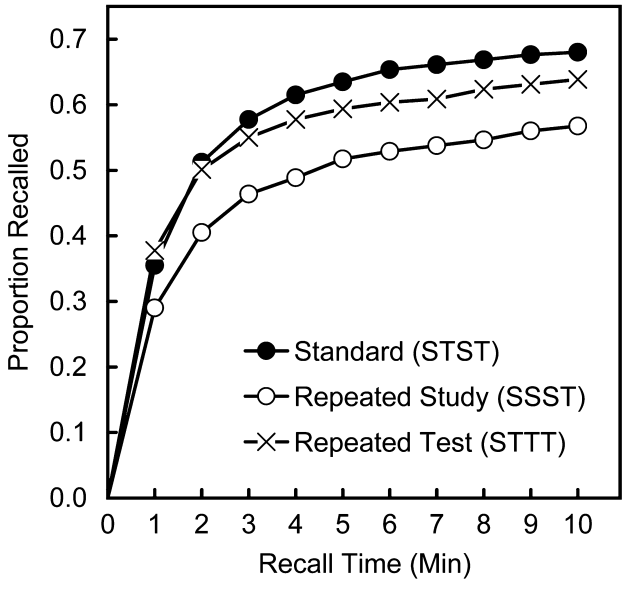

Fig. 4. Cumulative recall on a final retention test given 1 week after initial learning. Results are shown separately for standard, repeated-study, and repeated-testing conditions. The shorthand condition labels indicate the order of study (S) and test (T) periods. Data are from Karpicke and Roediger (2006b).

Although the results just reported are striking, other, earlier experiments also showed testing effects in free recall. For example, Hogan and Kintsch (1971) reported two experiments showing the advantage of test trials over study trials in promoting long-term retention. In one experiment, they had some students study a list of 40 words four times, with only short breaks between presentations of the lists. A second group studied the list once and then took three consecutive free-recall tests (similar to a single cycle in the repeated-test condition of Tulving's, 1967, experiment). Both groups returned 2 days later for a final test. The pure-study group recalled 15% of the words, whereas the group that received only one study trial but three tests recalled 20%. A single study trial and three tests produced significantly better recall than did studying the material four times.

Repeated Testing and Selective Re-Presentation of Forgotten Material

Thompson, Wenger, and Bartling (1978) replicated Hogan and Kintsch's (1971) results, again using 40-word lists, but with two new twists that deserve special mention. In addition to conditions with four study trials (repeated-study condition) and one study trial and three tests (repeated-test condition), they included a condition in which subjects studied the list once, recalled it, studied only those words they failed to recall, recalled the entire list again, and so on for three more study-test episodes with the study lists becoming shorter and shorter. This test/representation condition mimicked a variation of what students are often told to do in study guides: study the material, test themselves, restudy items they missed, and so on until they achieve perfect mastery (this guidance is similar to what Gates's, 1917, subjects were instructed to do). However, note that the subjects of Thompson et al. were instructed to recall the entire list on each test trial, not just the items they restudied in the previous study phase. Besides adding this condition to Hogan and Kintsch's (1971) design, Thompson et al. also included final tests 5 min after the learning phase and 2 days later. (Retention interval was manipulated between subjects, so the 5-min test would not influence the 2-day test.)

Table 1 summarizes the results Thompson et al. (1978) obtained. It is clear that on the 5-min test, the group that had only one study trial but repeated tests had the poorest recall. The group that only studied the lists did next best, but the group that

TABLE 1

Proportion Correct in Immediate and Delayed Recall in Thompson, Wenger, and Bartling's (1978) Experiment 2

Condition	Test		Difference (5 min – 48 hr)	Percentage forgetting
	5 min	48 hr		
Repeated study (SSSS)	.50	.22	.28	56
Repeated test (STTT)	.28	.25	.03	13
Repeated test and re-presentation ($ST^RT^RT^R$)	.60	.44	.16	26

Note. Percentage forgetting was calculated as follows: [(recall at 5 min – recall at 48 hr)/recall at 5 min] × 100. S = study period; T = test; T^R = test with re-presentation of forgotten items.

was tested with re-presentation of the missed items performed best of all. However, 2 days later, the situation changed. Although the test/re-presentation group still did best, the repeated-test group slightly outperformed the repeated-study group. Looking at these results another way, subjects in the repeated-study condition showed dramatic forgetting over 2 days (measured either as the difference between 5-min and 2-day recall or as a percentage of 5-min recall; see Loftus, 1985). Although subjects in the repeated-study condition forgot 56% of what they originally could recall, those in the test/re-presentation condition forgot 26%, and subjects in the repeated-test condition showed the least forgetting, just 13%. This outcome shows that the advice in study guides appears to be accurate: Students should study, test themselves, and then restudy what they did not know on the test. However, in a later experiment, we (Karpicke & Roediger, 2006b, Experiment 2) showed that the fact that Thompson et al. required recall of the entire list during each test was critical to this outcome. If students in the test/re-presentation condition are required to recall only the items that were presented in the preceding re-presentation study phase, they display rather poor recall on a delayed test. Repeated testing of the whole set of material is critical to improve long-term retention.

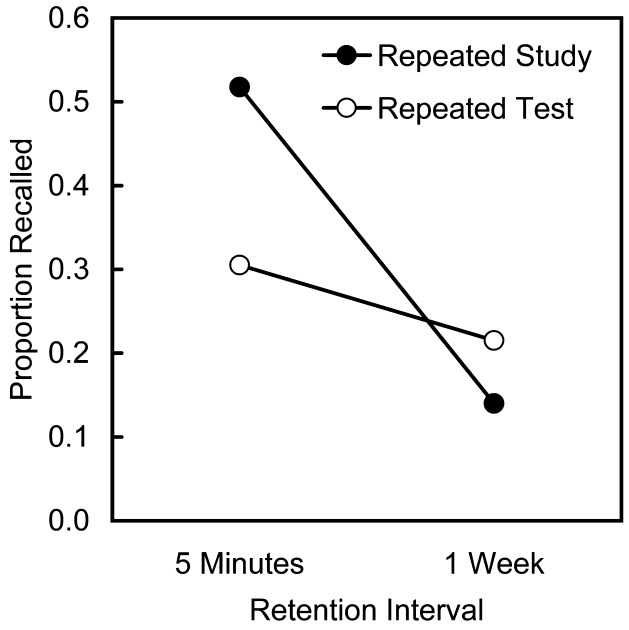

Fig. 5. Proportion of words recalled on immediate (5-min) and delayed (7-day) retention tests after repeated studying or repeated testing. Data are estimated from Wheeler, Ewers, and Buonanno (2003).

In sum, the results of Thompson et al. also show the power of testing for enhancing long-term retention: Both tested groups recalled more on the delayed final test than the group that only studied the word lists, without initial testing. On the delayed test in this experiment, the advantage of repeated testing over repeated studying was rather small (Thompson et al., 1978), probably because of the relatively brief amount of time given to subjects to recall on the initial tests. Nevertheless, the effect has been replicated by Wheeler, Ewers, and Buonanno (2003). In their second experiment, subjects studied a 40-word list either five times (repeated-study condition) or one time with four consecutive recall tests (repeated-test condition). Final free-recall tests were given to different groups of subjects either 5 min or 1 week later. The results are shown in Figure 5, which reveals a huge advantage for massed study on the immediate test, but a significant reversal on the test given a week later. This result and others like it are even more surprising when one considers that in the repeated-study condition, subjects are presented with all 40 words in the list on each trial, whereas in the repeated-test condition, they are reexposed only to those words that they can recall (only about 11 of the 40 words in this experiment). Thus, the overwhelmingly greater number of exposures in the repeated-study condition improved performance only on a relatively immediate test. After a 1-week delay, subjects in the repeated-test condition outperformed those in the repeated-study condition despite having studied the material only once. Once again, the power of testing is clear. In a later section, we review evidence that the same pattern holds for recall of text materials like those used in educational settings (Roediger & Karpicke, 2006).

The experiments we have just discussed compared conditions with several recall tests and conditions in which students repeatedly studied the material. Wheeler and Roediger (1992) investigated whether multiple tests are more beneficial than a single test, and also gave subjects fairly lengthy initial recall tests (unlike most of the experiments reviewed thus far). In some

conditions, subjects heard a story that named 60 particular concrete objects. A picture of each object was shown on a screen the first time the object was named in the story, and subjects were told that they would be tested on the names of the pictures. After presentation, control subjects were dismissed from the lab and asked to return a week later. Another group of subjects took one 7-min recall test and left, and a third group received three recall tests before being permitted to leave. All subjects returned a week later for a final recall test. The results are shown in Table 2. On the initial test, subjects in the single-test condition recalled 53% of the items; control (no-test) subjects would presumably have recalled about the same number of items had they been tested, so this estimate was used to measure forgetting in that condition. Subjects in the three-test condition recalled 61% of the items on their third test; their recall was higher than that of subjects in the one-test condition because recall often increases upon such repeated testing, a phenomenon called hypermnesia (Erdelyi & Becker, 1974; Roediger & Thorpe, 1978). Final recall after a week was 29% in the no-test condition, 39% in the one-test condition, and 53% in the three-test condition. Clearly, forgetting (as either a difference or a proportion) was inversely related to the number of immediate tests, with subjects exhibiting 13% forgetting after three tests, 27% forgetting after one test, and 46% forgetting after no tests. In a sense, subjects who received three tests were completely immunized against forgetting, because they recalled the same number of pictures after a week that subjects in the single-test condition recalled a week earlier (53%). The two extra tests in the repeated-testing condition maintained performance at a high level 1 week later.

Summary

The experiments we have reviewed in this section all involved free-recall tests or slight variations of free-recall tests. Tulving (1967), among other researchers, showed that within very broad limits, a free-recall test permits as much learning as restudying material. However, later research showed a more complicated picture: Repeatedly studying material is beneficial for tests given soon after learning, but on delayed criterial tests with retention intervals measured in days or weeks, prior testing can produce greater performance than prior studying. In the case of delayed recall, test trials produce a much greater gain than study trials. Of course, there must be at least one study opportunity for testing to enhance later recall, but many of the experiments we have discussed used only one study trial followed by several tests and yet demonstrated an advantage in delayed recall for this condition over one in which there were multiple study trials (e.g., five study trials and no tests in Wheeler et al., 2003). Testing reduces forgetting of recently studied material, and multiple tests have a greater effect in slowing forgetting than does a single test (Wheeler & Roediger, 1992). We consider theoretical accounts of these data in a later section, but first we review selected experiments from a different tradition of testing research.

TABLE 2

Proportion of Pictures Recalled Immediately After Study and 1 Week Later in Wheeler and Roediger (1992)

	Test			
Condition	Immediate	Delayed 1 week	Difference (immediate – delayed)	Percentage forgetting
No test	(.53)[a]	.29	.24	46
One test	.53	.39	.14	27
Three tests	.61[b]	.53	.08	13

Note. Percentage forgetting was calculated as follows: [(immediate recall – recall at 1 week)/immediate recall] × 100.

[a]Because subjects in this condition did not take an immediate test, the performance of subjects in the one-test condition was used to estimate their likely performance so that their forgetting could be measured. [b]This proportion is taken from the third test.

TESTING EFFECTS IN PAIRED-ASSOCIATE LEARNING

When a person learns names to go with faces, or that *caballo* means "horse" in Spanish, or that $8 \times 9 = 72$, or that a friend's telephone number is 792-3948, the task is essentially one of paired-associate learning. Of course, in the laboratory, paired-associate learning is often studied using word pairs that may vary in association value (*chair-table* or *chair-donkey*) or nonword-word pairings (*ZEP-house*), among many other variations. This task, first used in experiments by Calkins (1894), has been a favorite for studying testing effects. In addition to mimicking many learning situations with which people are faced in daily life, the task is especially tractable in the laboratory. When used to investigate the testing effect, the task makes it possible to manipulate the interval between study and test of a specific pair, and presentation or withholding of feedback can also be easily accomplished. In this section, we briefly review literature showing testing effects in paired-associate learning and then turn to the issue of spaced testing in continuous paired-associate tasks.

Testing Effects in Cued Recall and Paired-Associate Tests

Estes (1960) began research on testing effects in paired-associate learning, and this work has been carried forward by other researchers. For example, Allen, Mahler, and Estes (1969) had subjects study a list of paired associates either 5 or 10 times and then take no, one, or five tests on the items. One day later, the subjects were given a final retention test in which they were cued with the stimulus (the left-hand member) of the pair and asked to recall the response. Allen et al. found a modest benefit of studying the list 10 times relative to studying it 5 times, but the effects of initial testing were much larger, with final test performance in both study conditions increasing directly as a function of the number of initial tests (see Table 3). Final test performance of subjects who studied the list 5 times and were tested once was equivalent to that of subjects who studied the list

TABLE 3

Proportion of Final Cued Recall on a 24-Hr Retention Test as a Function of Different Levels of Initial Study and Number of Tests on Day 1 (from Allen, Mahler, & Estes, 1969)

	Number of initial tests		
Condition	None	One	Five
5 study trials	.58	.66	.82
10 study trials	.65	.81	.88

10 times and received no initial test. This outcome led Allen et al. to conclude that taking a single test was as effective for long-term retention as 5 additional study trials. Izawa, in particular, has continued this line of research and produced a large body of work (e.g., Izawa, 1966, 1967, 1970; see Izawa, Maxwell, Hayden, Matrana, & Izawa-Hayden, 2005, for a recent summary of this program of research). Izawa has referred to test trials as potentiating future learning and presented a mathematical model of how this process might operate, although this model is specific to repeated study-test trials (Izawa, 1971).

In a rather different tradition, Jacoby (1978) had subjects study word pairs (e.g., *foot-shoe*) and then either restudy the pair (*foot-shoe*) or take a simple test in which they had to generate the right-hand member of the pair when given the left-hand member and a fragmented form of the right-hand member (*foot-s_ _e*). Further, the second occurrence of the pair (either restudied or tested) was either immediately after the pair had initially been studied or after a delay filled with 20 intervening pairs. Many different pairs were presented in these four conditions (restudy or test after either a short or a long delay). At the end of the experiment, subjects received a final test in which they were given only the left-hand cue word and were asked to recall the right-hand target (*foot*-????). The results on this final test showed that prior testing with the fragment (*foot-s_ _e*) led to better retention than restudying the intact word pair (*foot-shoe*), once again demonstrating that testing can be better than restudying material even when the "test" seems quite simple. In addition, recall on the final test was much better when the initial test had been delayed by 20 intervening items than when it occurred immediately after study of the pair. Jacoby argued that when the test occurred immediately after the study phase, the effortful processing that usually occurs during memory retrieval was short-circuited, and the test lost its potency. We return to this issue later.

Jacoby's (1978) experiment is often cited as a pioneering study of the generation effect (the fact that generating material often leads to better recall or recognition than reading the same material; see also Slamecka & Graf, 1978), a phenomenon related to the testing effect. The fragment cues led to high levels of recall (above 90%) on the initial tests in Jacoby's experiment, but other researchers using standard cued-recall tests that do not produce such high initial recall levels have also demonstrated positive effects of testing on later retention of paired-associate material (Carrier & Pashler, 1992; Kuo & Hirshman, 1996; McDaniel & Masson, 1985). Tests during paired-associate learning greatly reduce forgetting (Runquist, 1986), and the effects are increased when feedback is given for items that are missed on the tests (see Cull, 2000; Pashler, Cepeda, Wixted, & Rohrer, 2005). Thus, the testing effects observed in free recall also hold in paired-associate learning.

Spaced Retrieval Practice With Paired Associates

We now focus on a practical question raised by Landauer and Bjork (1978). Given that testing generally improves retention relative to restudying, they asked if the schedule of testing matters. If a subject learns an A-B pair (where A might be *horse* and B *caballo*), what is the best sequence of testing to promote long-term retention? Perhaps testing should occur soon after learning and be repeated in a massed fashion, because multiple tests promote better retention than a single test. Massed testing immediately after study would also permit errorless retrieval on the repeated tests. But perhaps spacing tests over intervals of time is a better schedule, because spaced practice is known to benefit retention in the long term (e.g., Glenberg, 1976; Melton, 1970; for a review, see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). However, if tests are spaced at equal intervals, then delaying an initial test after studying a pair (in a spaced schedule) may lead to forgetting. Thus, Landauer and Bjork made the case for an expanding schedule of testing. In this scheme, a first test occurs immediately after an A-B pair is presented, to ensure that subjects can recall B when given A. Then, a longer span of time (with more studied and tested items presented) occurs before A is presented again for a test, and a yet longer time occurs before a third test, and so on. The idea behind expanding retrieval schedules is to gradually shape production of the desired response so that it can be retrieved out of context, at a long delay (the analogy is to shaping of responses in operant conditioning).

Of course, if an expanding schedule of repeated retrieval shows an advantage over massed testing, this advantage might accrue simply because the expanding schedule, unlike the massed schedule, involves spaced presentations (Rea & Modigliani, 1985). For this reason, Landauer and Bjork (1978) tested expanding and equal-interval schedules matched on the average spacing between tests. For example, if the expanding schedule was 1-5-9 (the numbers refer to the number of trials intervening between successive tests of A-B after its study), then the appropriate equal-interval schedule was 5-5-5, which on average produced the same amount of spacing, but distributed equally. Expanding retrieval practice is thought to be an optimal schedule for long-term retention because success is high on an immediate test and then the spacing implemented on the expanding tests gradually increases the difficulty of retrieval attempts, encouraging better later retention.

Landauer and Bjork (1978) reported two experiments that compared these schedules in paired-associate learning (first

name–surname pairs in one experiment and name-face pairs in the other). No feedback or correction was given to subjects if they made errors or omitted answers. Landauer and Bjork found that the expanding-interval schedule produced better recall than equal-interval testing on a final test at the end of the session, and equal-interval testing, in turn, produced better recall than did initial massed testing. Thus, despite the fact that massed testing produced nearly errorless performance during the acquisition phase, the other two schedules produced better retention on the final test given at the end of the session. However, the difference favoring the expanding retrieval schedule over the equal-interval schedule was fairly small at around 10%.

In research following up Landauer and Bjork's (1978) original experiments, practically all studies have found that spaced schedules of retrieval (whether equal-interval or expanding schedules) produce better retention on a final test given later than do massed retrieval tests given immediately after presentation (e.g., Cull, 2000; Cull, Shaughnessy, & Zechmeister, 1996), although exceptions do exist. For example, in Experiments 3 and 4 of Cull et al. (1996), massed testing produced performance as good as equal-interval testing on a 5-5-5 schedule, but most other experiments have found that any spaced schedule of testing (either equal-interval or expanding) is better than a massed schedule for performance on a delayed test. However, whether expanding schedules are better than equal-interval schedules for long-term retention—the other part of Landauer and Bjork's interesting findings—remains an open question. Balota, Duchek, and Logan (in press) have provided a thorough consideration of the relevant evidence and have shown that it is mixed at best, and that most researchers have found no difference between the two schedules of testing. That is, performance on a final test at the end of a session often shows no difference in performance between equal-interval and expanding retrieval schedules.

For example, Balota, Duchek, Sergent-Marshall, and Roediger (2006) compared expanding-interval retrieval tests with equally spaced tests and massed tests in three groups of subjects: young adults, healthy older adults, and older adults with Alzheimer's disease. They presented items twice (to ensure that patients encoded them) and then employed massed testing for some items (0-0-0), equal-interval testing for others (3-3-3), and expanding-interval testing for still others (1-3-5). A final test occurred at the end of the session. During acquisition, all three groups showed the highest level of performance on the massed tests, the next best performance on the expanding-interval tests, and the worst performance on the equal-interval tests. This last outcome was due to the relatively long lag before the first test for the equal-interval condition. However, despite these differences during acquisition, on the final test at the end of the session, there was no difference between the equal-interval and expanding-interval conditions for any of the three groups (although recall in both these conditions was superior to that in the massed-test condition). Carpenter and DeLosh (2005) showed similar effects in learning of name-face pairs, except that on their final test they found a slight benefit for an equal-interval condition over an expanding-interval condition.

Thus far, we have reviewed studies comparing expanding- and equal-interval retrieval over a relatively narrow range of possible spacing schedules. Logan and Balota (in press) used a variety of expanding schedules and compared them with appropriate equal-interval schedules in younger and older adults. In younger adults, they found that recall at the end of the session was no better for expanding- than for equal-interval testing, but they did find an advantage for expanding-interval retrieval among older adults. However, Logan and Balota also gave subjects a 24-hr delayed test and discovered that initial equal-interval testing produced better recall on this test than did the expanding-interval testing schedule. This outcome occurred despite the fact that expanding-interval retrieval produced better recall during initial acquisition and (for older subjects) on the test at the end of the first day.

We recently obtained a similar result (Karpicke & Roediger, 2006a), using pairs consisting of vocabulary words and their meanings (e.g., *sobriquet-nickname*). We tested subjects in massed (0-0-0), equal-interval (5-5-5), and expanding-interval (1-5-9) conditions during acquisition, and then subjects were given a final test either 10 min or 2 days after the learning session. At both retention intervals, the spaced-practice conditions produced better recall than massed practice. On the 10-min test, we replicated Landauer and Bjork's (1978) results by showing that expanding-interval retrieval produced a modest benefit relative to equal-interval retrieval. However, after 48 hr, we found the opposite pattern of results: Items in the equal-interval condition were recalled better than items studied under an expanding-interval schedule. We replicated this pattern of results in a second experiment in which subjects were given feedback after each test trial during the learning phase.

Our results (Karpicke & Roediger, 2006a) and those of Logan and Balota (in press) indicate that in some circumstances, equal-interval retrieval practice may promote greater long-term retention than expanding-interval retrieval practice. We have argued that the factor responsible for the advantage of equal-interval practice is the placement of the first retrieval attempt: The longer interval before the first test demands more retrieval effort and leads to better retention (this argument is similar to what Jacoby, 1978, concluded). Other research with paired associates has shown that increasing the delay before an initial test promotes later retention, even though success on the initial test often decreases with increasing delays (e.g., Jacoby, 1978; Modigliani, 1976; Pashler, Zarow, & Triplett, 2003; Whitten & Bjork, 1977). In the equal-spacing conditions used by Logan and Balota (in press) and by us (Karpicke & Roediger, 2006a), as well as by other researchers, the first retrieval attempt occurred after a brief delay. However, the hallmark of expanding-interval retrieval practice is an initial retrieval attempt immediately after studying, to ensure high levels of recall success. Indeed,

performance on this massed initial test is often nearly perfect, most likely because the test involves retrieval from primary or short-term memory. However, retrieval from primary memory usually does not produce benefits for later retention (see also Craik, 1970; Madigan & McCabe, 1971). Thus, equally spaced practice may lead to benefits for long-term retention because of the delayed initial test, and current research is aimed at clarifying why certain spacing conditions are more or less effective for learning (see Balota et al., in press).

Summary

Many of the testing effects found with free-recall tests hold true in paired-associate learning. Tests promote better retention than do additional study trials with paired associates, and repeated tests provide even greater benefits. In addition, paired associates have been used to investigate whether a particular type of testing schedule is optimal for long-term retention. Most of the research has indicated that spaced retrieval practice leads to better retention than massed practice, but the evidence is mixed regarding whether expanding-interval retrieval is a superior form of spaced retrieval. The most recent evidence points to the conclusion that expanding-interval retrieval may not benefit long-term retention, as was originally thought, because the initial test in an expanding schedule appears too soon after study, rendering it ineffective for enhancing learning. Although the efficacy of expanding and equally spaced schedules remains an open issue, the research we have reviewed shows that delaying an initial retrieval attempt and spacing repeated tests often will boost later retention with paired-associate materials.

TESTING EFFECTS WITH EDUCATIONAL MATERIALS

Many of the testing effects we have discussed so far have been observed in psychology laboratories, and the effects have been obtained with materials commonly used in the lab, such as lists of words or unrelated word pairs. Some exceptions do exist. Positive effects of testing have been found in experiments using foreign-language vocabulary words (e.g., Carrier & Pashler, 1992), materials taken from test-preparation books for the Graduate Record Examination (Karpicke & Roediger, 2006a; Pashler et al., 2003), and general knowledge questions (McDaniel & Fisher, 1991). The two classic studies by Gates (1917) and Spitzer (1939) also used educational materials, but these examples aside, the majority of the research on testing effects has used materials that are not found in educational settings. Moreover, the limited range of materials most likely is part of the reason why the testing effect is not widely known in education and has not been incorporated into educational practice. One can therefore wonder, does the testing effect generalize to educationally relevant materials and test formats? The answer to this question is "yes," and in this section, we review research using prose materials and then focus on the effects of different types of tests often used in schools (e.g., short-answer questions and multiple-choice tests).

Testing Effects With Prose Materials

One area of research related to the testing effect has shown that answering questions while reading textbook material often facilitates comprehension and retention of the material. The beginning of research on such adjunct questions is attributed to pioneering studies by Rothkopf (1966), who referred to brief questions placed at different points throughout an instructional text as "test-like events." The effects of adjunct questions on learning were investigated intensively until the 1980s (see Hamaker, 1986), but have received little attention since. Research on adjunct questions showed that they often facilitate retention and comprehension of text material and also pointed to two other important conclusions. First, questions that follow a text promote better retention than questions that appear in advance of the text or interspersed throughout the text. Second, answering questions that accompany a text will often enhance later performance on related questions (see also Chan, McDermott, & Roediger, in press, which is discussed later). We mention the research on adjunct questions only briefly because that literature has been extensively reviewed elsewhere (see R.C. Anderson & Biddle, 1975; Crooks, 1988; Hamaker, 1986; Rickards, 1979). Although the results indicate that these test-like events do facilitate learning of prose materials, it is not clear how often students actually answer questions that accompany texts or how closely adjunct questions approximate the conditions of actual classroom tests.

Recently, we (Roediger & Karpicke, 2006) have investigated the testing effect taking an approach aimed at integrating the research tradition from cognitive psychology, which we have just reviewed (e.g., Hogan & Kintsch, 1971; Thompson et al., 1978; Wheeler & Roediger, 1992), with educational research that has focused on learning of more complex prose materials. In our experiments, we had college students study prose passages covering general scientific topics. Depending on the condition to which a passage was assigned, the students then either restudied the entire passage or took a free recall test in which they were asked to write down as much as they could remember from the passage (this test was similar to an essay test in school contexts). The students were not given any feedback about their test performance (i.e., they did not restudy the material after the test), but were given ample time (7 min) to study the passage in the restudy condition and to take the recall test in the test condition (as mentioned earlier, the brief amount of time given to subjects in previous experiments probably attenuated the positive effects of testing). Finally, 5 min, 2 days, or 1 week after the learning session, different groups of students took a final free-recall test that was just like the recall test given initially. The results of the experiment are shown in Figure 6. After 5 min, restudying produced a modest benefit over testing (81% vs. 75% of the passage recalled), but the opposite pattern of results was

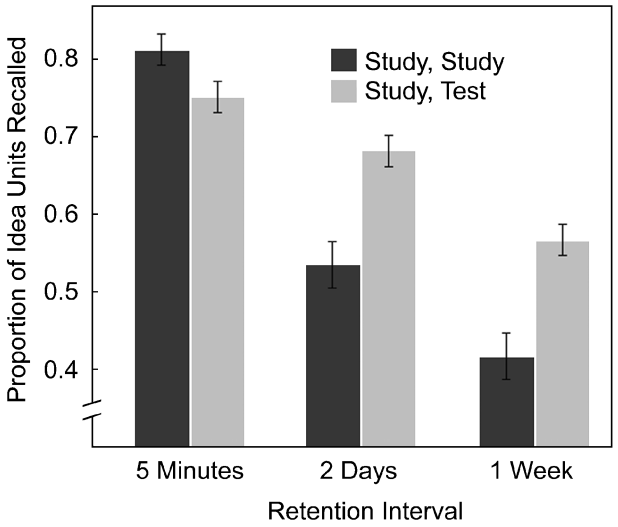

Fig. 6. Mean proportion of idea units recalled from a prose passage after a 5-min, 2-day, or 1-week retention interval as a function of whether subjects studied the passages twice or studied them once before taking an initial test. Error bars represent standard errors of the means. From Roediger and Karpicke (2006).

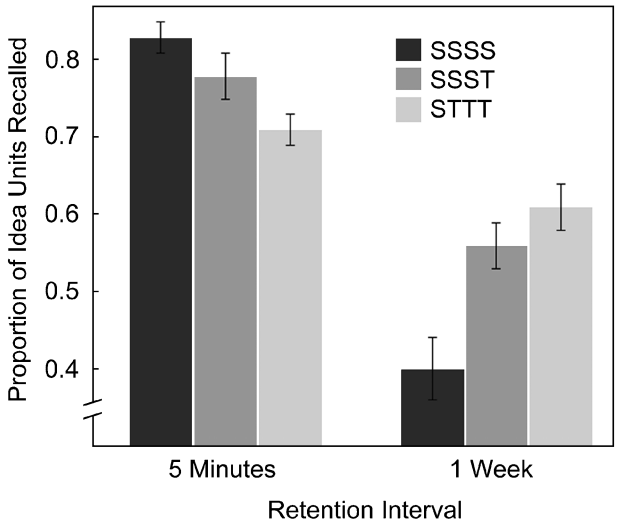

Fig. 7. Mean proportion of idea units recalled on a final test 5 min or 1 week after learning as a function of learning condition. The shorthand condition labels indicate the order of study (S) and test (T) periods. Error bars represent standard errors of the means. From Roediger and Karpicke (2006).

observed on the delayed retention tests. After 2 days, initial testing produced better retention than restudying (68% vs. 54%), and an advantage of testing over restudying was also observed after 1 week (56% vs. 42%). The results conceptually replicate earlier experiments using free recall and paired-associate learning of lists and generalize them to educational materials.

We conducted a second experiment to investigate the effects of repeated studying and repeated testing on later retention (Roediger & Karpicke, 2006). Subjects studied passages during four separate periods (SSSS), studied during three periods and took one recall test (SSST), or studied during one period and took three tests (STTT). They took a final recall test either 5 min or 1 week after this learning session. The results, which are shown in Figure 7, reveal that after 5 min, recall was correlated with repeated studying: The SSSS group recalled more than the SSST group, who in turn recalled more than the STTT group. However, on the 1-week retention test, recall was correlated with the number of initial tests: The STTT group recalled more than the SSST group, who in turn recalled more than the SSSS group. In terms of proportional measures of forgetting (which take into account differences in the level of original learning), the SSSS group showed the most forgetting (52%), followed by the SSST group (28%), and the repeated-testing group (STTT) showed the least amount of forgetting (10%) over 1 week.

Our results (Roediger & Karpicke, 2006) demonstrate the powerful effect testing has in enhancing later retention, and confirm and extend with prose materials the earlier findings with word-list materials. In addition, we investigated the subjects' experience after repeated studying or repeated testing by asking them to predict how well they thought they would remember the passage in the future. These predictions were inflated after repeated study, relative to the testing conditions, even though repeated studying produced the worst long-term retention (see Dunlosky & Nelson, 1992, for a similar result). This finding suggests that students may prefer repeated studying because it produces rapid short-term gains, even though it is an ineffective strategy for long-term retention.

Testing effects have also been found using educationally relevant test formats, such as short-answer and multiple-choice tests. In another experiment (Agarwal, Karpicke, Kang, Roediger, & McDermott, 2006), we had students study textbook passages and then complete short-answer tests on some of the passages. An initial short-answer test enhanced retention on a final short-answer test given 1 week later, relative to studying the passage without taking the test. We also investigated the effects of giving students feedback about their test performance. Providing feedback (by having students restudy the passage) enhanced retention to a greater extent than testing alone, but the effectiveness of feedback depended on when it occurred. In one condition, students were shown the passage while they took the test. This condition was similar to open-book testing commonly used in education and also similar to taking notes while reading. Subjects in this condition had access to feedback continuously during the test. In another condition, students took the test and then were given the passage and instructed to look over their responses (a delayed-feedback condition). Although the immediate-feedback condition produced the best performance on the initial test (not surprisingly), the delayed-feedback condition promoted better long-term retention. The results of this study are analogous to those obtained with motor learning tasks (see Schmidt & Bjork, 1992) and suggest that students should

delay feedback or reviewing their answers until after completing a test in order to optimize later retention.

Nungester and Duchastel (1982) investigated the effects of multiple-choice and short-answer tests on later retention of a prose passage. In their experiment, one group of subjects studied the passage and then took an initial test in which half of the questions were short-answer questions and half were five-alternative multiple-choice questions. Another group of subjects studied the passage and then reviewed portions of it, and a third group studied the passage only once. All the students returned 2 weeks later for a final retention test, in which each question was in the alternate format relative to the initial test (i.e., items that were initially tested in short-answer format were tested in multiple-choice format on the final test, and likewise initial multiple-choice questions were tested as short-answer questions on the final test). Nungester and Duchastel found that reviewing the passage enhanced retention relative to just studying it once, but taking the initial test led to the best retention. This testing effect was found for both the multiple-choice and the short answer-test formats (see also LaPorte & Voss, 1975). In addition, in a follow-up to this original experiment, Nungester and Duchastel had the same subjects take another multiple-choice retention test 5 months after the initial learning session (see Duchastel & Nungester, 1981). The pattern of results was identical on this 5-month test, with the initially tested group performing better than the study-once and study-twice groups. Nungester and Duchastel's work provides a compelling demonstration that the testing effect persists over very long retention intervals (see also Butler & Roediger, in press, and Spitzer, 1939).

Transfer of Testing Effects Across Different Test Formats

The research just described shows that both short-answer and multiple-choice tests produce positive testing effects on later retention. Other research on testing effects with prose materials has investigated whether certain types of tests (e.g., essay, short-answer, or multiple-choice) are more effective than others for enhancing retention, or whether a particular test format facilitates later performance only for that test format. These issues have also been addressed in laboratory research on the effects of recall tests on performance on later recognition tests (e.g., Darley & Murdock, 1971; Lockhart, 1975; Wenger, Thompson, & Bartling, 1980) and the effects of recognition tests on later recall (e.g., Mandler & Rabinowitz, 1981; Runquist, 1983; see also Carpenter & DeLosh, 2006; Hogan & Kintsch, 1971). In this section, we review studies that have used educational materials to investigate the effects of different test formats.

To address the issue of whether testing effects are greater with certain types of tests than with others, we again return to the work of Duchastel and Nungester. Although these researchers carried out several investigations of the testing effect in the early 1980s, their work is rarely cited in discussions of the testing effect. In one study, Duchastel (1981) gave some students an initial short-answer or multiple-choice test on a prose passage and then a final short-answer test 2 weeks later. Both types of initial tests produced better long-term retention than studying alone, but taking the initial short-answer test promoted superior retention 2 weeks later on the final short-answer test. Thus, this work provides evidence that perhaps short-answer tests yield greater testing effects than multiple-choice tests (but see Duchastel & Nungester, 1982, for a somewhat different conclusion).

Glover (1989) had students study a prose passage similar to the one used by Duchastel and Nungester (1982). Two days after studying the passage, the students took a free-recall test, a cued-recall (fill-in-the-blank) test, or a recognition test that involved identifying whether statements had or had not been in the original passage. Two days later, the students took a final free-recall, cued-recall, or recognition test. Glover found that taking the initial free-recall test produced the best final retention, regardless of the format of the final test, and the cued-recall test produced better retention than the recognition test on both the final cued-recall test and the final recognition test. Glover's study indicates that recall tests promote greater retention than recognition tests, which is also a conclusion generally reached by researchers studying testing effects in word-list paradigms. However, one oddity in Glover's study was that scores on the free-recall test were consistently higher than scores on the cued-recall test, which indicates that subjects could recall more in free recall than they did on Glover's cued-recall test, a result directly in contrast to the results of fundamental research on human memory (e.g., Tulving & Pearlstone, 1966). This strange aspect of Glover's data is most likely an artifact of the type of questions asked on the cued-recall test, which somehow led to subjects being able to recall more in free recall than they could express on the cued-recall test. Thus, Glover's results should be interpreted with some caution.

Recently, Kang, McDermott, and Roediger (in press) reexamined the testing effect with short-answer and multiple-choice tests in a study with better control of test content, to try to ensure that the same information was being tested by the two formats. They also examined transfer across test format and examined the role of feedback on a first test in enhancing the testing effect. The students studied articles from *Current Directions in Psychological Science*, and after each article, they took a short-answer or a multiple-choice test. We consider Experiment 2, in which subjects received feedback after the tests, a procedure that equates exposure to information for multiple-choice and short-answer tests. In addition, in a control condition, the students read statements from the articles after reading them; these statements were the same as the items that were tested in the other two conditions, again to equate exposure to the information. Three days later, the students took a final test in either a short-answer or a multiple-choice format. The initial short-answer test produced the best retention for both final-test formats (results consistent with those of Glover, 1989). Butler and Roediger (in press) and McDaniel, Anderson, Derbish, and Morrisette (in press) have reported similar outcomes.

Summary

Clearly, the work using educationally relevant materials has not resolved all the questions concerning the effect of test format. However, some conclusions are warranted. In virtually all the experiments, taking an initial test led to better later retention than not taking a test or than engaging in a period of additional study. The testing effect is secure. Most evidence points to the conclusion that tests involving production of information (essay and short-answer tests) produce greater benefits on later tests than do multiple-choice tests, which involve recognition of a correct answer among alternatives. The literature is not totally consistent on this point, however, so it remains a hypothesis for further investigation. One problem is that performance is usually much higher on initial multiple-choice tests than on initial short-answer tests; unless feedback is given to equate exposure to answers, multiple-choice tests may have an advantage over short-answer tests simply for this reason. Kang et al. (in press) found that a short-answer test (with feedback) produced a greater testing effect than did a multiple-choice test (also with feedback), regardless of the format of the final test. A greater testing effect for production tests than for recognition tests would be similar to the generation effect during study of material. That is, generating or producing material during study usually creates greater retention than reading the material (Jacoby, 1978; Slamecka & Graf, 1978).

TESTING EFFECTS IN THE CLASSROOM

The experiments we have described show that the testing effect generalizes to educationally relevant materials (e.g., prose passages) and to test formats like those used in education (e.g., short-answer and multiple-choice tests). Nonetheless, most of the studies described so far have been carried out in the laboratory, and one can still ask whether the testing effect generalizes to actual classroom situations. Several differences between the laboratory and the classroom may lead to different results in these two contexts. For example, the amount of information that students are responsible for learning is much greater in the classroom than in the laboratory (even when the laboratory materials include prose passages taken from educational textbooks). Also, the to-be-learned materials in the classroom are presented in a variety of ways—in textbooks, in lectures, in class discussions, and so on. Students also differ greatly in the amount of studying they do before exams, in how soon they begin studying (relative to when exams occur), in their interest in the course material, and in their motivation to learn. All these factors are typically controlled in well-designed experiments, but they are free to vary in the classroom. In this section, we review evidence from classroom studies of the testing effect. This evidence shows that despite the differences between psychology laboratories and school classrooms, the testing effect is a robust phenomenon in educational settings, and frequent testing in the classroom improves students' learning.

Although classroom studies of frequent testing date back to the 1920s (Deputy, 1929; Maloney & Ruch, 1929), relatively few systematic studies have been carried out since that time. Bangert-Drowns, Kulik, and Kulik (1991) conducted a meta-analysis of 35 classroom studies (22 published, 13 unpublished), carried out from 1929 through 1989, that manipulated the number of tests given to students during a semester. All of the studies compared a frequently tested group of students against a control group of students who received fewer tests. Bangert-Drowns et al. obtained the studies from the Educational Resources Information Center (ERIC) and Dissertation Abstracts databases, and only studies in which the frequent-testing and control groups received identical instructions were included in the meta-analysis. Twenty-eight of the studies were carried out in college classrooms, and 7 were carried out in high school classrooms. Most of the classes covered math and science, but some covered other topics (e.g., reading, government, law), and the tests were conventional classroom tests, such as multiple-choice and short-answer tests (though Bangert-Drowns et al. did not analyze different test formats separately). The criterial measure for all studies was performance on a final examination given at the end of the class.

The majority of the studies Bangert-Drowns et al. (1991) included (29 of 35, 83%) found positive effects of frequent testing, and the mean effect size (standardized mean difference, *d*) was .23. Five of the studies found negative effects, and 1 study found no difference between frequent testing and the control condition. There was great variation in the number of tests given during the semester, with the number of tests in the control group ranging from 0 to 15, and number of tests in the frequent-testing group ranging from 3 to 75. To investigate the effects of increasing the number of tests during a semester-long class, Bangert-Drowns et al. fit the data from the frequent-testing and control conditions to a regression equation predicting the size of the effect (indicating gains in learning due to testing) from frequency of testing. The function they obtained, showing the relation between the number of tests given during the semester-long class and the expected effect size, is displayed in Figure 8, which shows that performance on the final test increased as a negatively accelerated function of the number of tests given in class. Most notably, giving just 1 test produced a big gain relative to giving no tests at all, and subsequent repeated tests added to these gains in learning. (Of course, unlike the experimental studies described earlier, the repeated-testing studies in this meta-analysis involved testing different sets of material, not the same set of material repeatedly.) Bangert-Drowns et al. noted that in 11 studies in which the control group received no tests, the effect size comparing the frequent-testing and control conditions was .54. However, when the control group received at least 1 test, the effect size dropped to .15. The implication is that including a single test in a class produces a large improvement in final-exam scores, and Figure 8 shows that gains in learning continue to increase as the frequency of classroom testing increases.

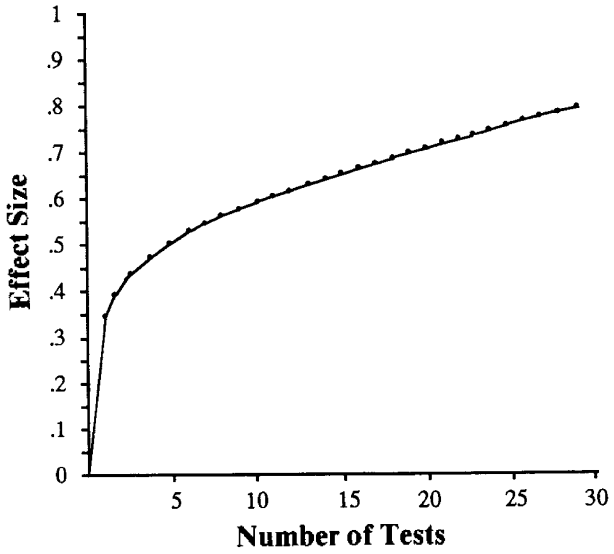

Fig. 8. Expected effect size for classroom testing as a function of the number of tests given during a semester-long course. From "Effects of Frequent Classroom Testing," by R.L. Bangert-Drowns, J.A. Kulik, and C.L.C. Kulik, 1991, *Journal of Educational Research*, *85*, p. 96. Copyright 1999 by Heldref Publications. Reprinted with permission of the Helen Dwight Reid Educational Foundation.

One other result from this meta-analysis (Bangert-Drowns et al., 1991) is worth noting. Four of the studies reported students' attitudes toward the amount of testing in their classes, and all four studies found that the students who were tested frequently rated their classes more favorably (in course ratings at the end of the semester) than the students who were tested less frequently. We return to this point later.

The meta-analysis of Bangert-Drowns et al. (1991) is lacking in some important respects. For instance, the authors did not analyze possible differences between test formats, nor did they include any information about what kind of feedback students received on their tests. In addition, most (29) of the studies included in the analysis did not randomly assign students to the frequent-testing or control conditions. Nevertheless, the implications of this meta-analysis are important: The testing effect works in the classroom, and students react favorably to frequent testing in their courses.

Leeming (2002) recently reported that giving a brief test each day in college courses on introductory psychology and on learning and memory improved students' final grades (relative to the grades in other courses he taught without daily testing). Leeming began each class period with a 10- to 15-min test that included about seven short-answer questions. After each test, he spent 2 to 3 min discussing the correct answers with the students (i.e., giving immediate feedback) before starting the lecture. Thus, a typical semester-long class that met 2 days a week could involve 22 to 24 exams. Leeming reported that the final grades in his courses with this exam-a-day procedure were better than the final grades in previous versions of the same courses that he had taught without daily testing (80% vs. 74% for the introductory psychology course and 89% vs. 80% for the learning and memory course). In addition, near the end of the course, Leeming had some introductory psychology students take a retention test covering material that had not been discussed in class for at least 6 weeks (thus, the retention interval before the test was approximately 6 weeks). Leeming compared students in the frequent-testing course with students in other sections of the introductory psychology course that did not involve daily testing and found that students in the frequent-testing course performed better on this test than did students in the other sections.

Leeming's (2002) report provides yet another example of how frequent testing in the classroom can enhance students' learning. Also, students in the frequent-testing classes completed a questionnaire about the procedure at the end of the course. The responses indicated that, overall, students liked the frequent-testing procedure. Although the majority of students agreed that they were skeptical about the procedure at the beginning of the course, they also indicated that they studied more frequently in this class than in other classes with fewer tests and believed that they learned more. The majority of students also said they liked daily testing and would choose frequent testing over fewer exams.

One problem with these classroom studies, noted earlier, is that they lack some of the controls included in laboratory experiments, such as random assignment of students to tested versus nontested conditions. Recently, McDaniel et al. (in press) were able to overcome the problem of random assignment by instead randomly assigning different items to the tested and nontested conditions in a within-subjects design. They had volunteer students enrolled in a brain and behavior course take weekly 10-min quizzes during the semester. The quizzes were administered and scored over the Internet and included short-answer questions or multiple-choice questions. Some individual statements or facts that were not included on the quizzes were presented to the students for them to reread, and other items were not reexposed to the students at all (no-exposure control condition). After completing each quiz, the students were given feedback about their performance on each question. The students also took two unit tests during the semester and then a cumulative final exam at the end of the course. Some items that appeared on the quizzes were repeated on the later criterial tests, and other items on the criterial tests had not been on a quiz (the items in the no-exposure control condition). However, the items that were repeated from the quizzes were worded differently when they appeared on the criterial tests.

McDaniel et al. (in press) observed similar patterns of results on the unit tests and final exams. Being reexposed to the facts (restudy, multiple-choice quiz, or short-answer quiz) produced a modest benefit over not being reexposed to them (no-exposure control condition), and both of the quiz conditions produced better performance on the unit and final tests than the restudy condition. Although taking multiple-choice quizzes produced better performance than studying the statements, short-answer quizzes produced even greater gains on the criterial tests. Thus,

the results of this classroom experiment converge with the results of other experiments in demonstrating the effectiveness of frequent testing for enhancing learning. Further, they confirm that short-answer tests produce greater testing effects than multiple-choice tests, supporting the results of the laboratory studies of Butler and Roediger (in press), Glover (1989), and Kang et al. (in press).

Summary

Classroom studies often lack the control over variables found in laboratory studies. Nonetheless, the meta-analytic study by Bangert-Drowns et al. (1991) reviewing the literature on frequency of classroom testing, Leeming's (2002) work in his own courses, and the within-subjects, within-course experiment of McDaniel et al. (in press) all point to the same conclusion, that the testing effect does generalize to the classroom.

THEORIES OF THE TESTING EFFECT

Prior reviews of the literature by Dempster (1996, 1997) identified two theories to account for the positive effects of testing on learning. He referred to these theories as the *amount-of-processing hypothesis* and the *retrieval hypothesis* (see also Glover, 1989). In this section, we evaluate and expand upon these two theories and provide additional explanations to account for the data we have reviewed. We first consider the idea that the testing effect is merely a result of additional exposure to material during the test (i.e., the amount-of-processing hypothesis), or more specifically, that testing simply leads to overlearning of a portion of the to-be-learned materials. As we have noted throughout this review, the bulk of evidence about the testing effect leads us to reject these ideas. Next, we discuss several ideas emphasizing that tests enhance learning via retrieval processes that reactivate and operate on memory traces either by elaborating mnemonic representations or by creating multiple retrieval routes to them (Bjork, 1975; McDaniel & Masson, 1985), and we discuss the related notion of creating "desirable difficulties" for learners, an idea championed by Bjork (1994, 1999; see also Bjork & Bjork, 1992). Finally, we consider the concept of transfer-appropriate processing (e.g., Blaxton, 1989; Morris, Bransford, & Franks, 1977; Roediger, 1990) and how it can be applied to the testing effect.

Additional Exposure and Overlearning

One idea that we sketched at the outset of our review is that a test provides additional exposure to the tested material, and that this extra exposure is responsible for the testing effect (an idea suggested by Thompson et al., 1978). We believe that the evidence is inconsistent with this simple explanation. The probable reason this idea arose is that many experiments on the testing effect have compared a condition in which students study material and then take a delayed final test with a condition in which subjects study, take an initial test, and then take the delayed final test. The latter condition shows better performance on the final criterial test—the testing effect—but this design confounds the effects of testing with the effects of total exposure time. Other experiments we have reviewed have equated exposure to the material in the two conditions (by re-presenting material for study in the control condition) and have still obtained robust testing effects. In fact, the usual restudy control condition provides a greater (rather than equal) exposure to the material, because in the testing condition subjects are reexposed only to the material that they could produce on the test. This suggests that some process other than additional exposure is responsible for the effect.

Nevertheless, some authors have argued that the testing effect simply reflects overlearning of items practiced on the test (e.g., Slamecka & Katsaiti, 1988; Thompson et al., 1978), concluding that it is not the process of retrieval per se that promotes later retention, but rather overlearning of a subset of the materials. This explanation, however, encounters problems explaining why additional studying produces better retention in the short term than repeated testing does, even though testing produces better long-term retention (e.g., Roediger & Karpicke, 2006; Wheeler et al., 2003). That is, repeated studying apparently leads to "overlearning" on immediate tests, but this initial overlearning does not translate into greater long-term retention because the testing conditions show better recall than the repeated-study conditions on delayed tests. In short, the additional-exposure, or overlearning, account predicts a main effect at all retention intervals and cannot explain the interaction that has been obtained in several experiments. Finally, an account of the testing effect based on additional exposure to, or overlearning of, the material practiced on the test does not provide an explanation for how tests can facilitate later retention of related material that was not tested (Chan et al., in press). We agree with previous researchers (Dempster, 1996; Glover, 1989) that accounts of the testing effect based on additional processing or overlearning are not satisfactory.

One other problem related to the exposure-overlearning account of the testing effect is also worth addressing. Some investigators may worry that the testing effect is nothing more than the result of some sort of item-selection artifact because subjects themselves select which items are recalled on an initial test. The logic would be as follows: Some items are inherently easier than other items (for whatever reason), and those easy items are recalled on an initial test and then again on the final test, producing the illusion that the test has caused learning when all it did was show that easy items can be recalled twice. That is, the "easy" items receive additional practice through the test and are better recalled later than items in the nontested control condition, in which they were not selected and practiced (see Modigliani, 1976, for discussion). However, this account cannot explain many important phenomena in the literature, such as the crossover interactions observed as a function of retention interval (e.g., Roediger & Karpicke, 2006; Wheeler et al., 2003). Moreover, procedures developed to estimate and remove item-

selection effects (when initial test performance differs across conditions) demonstrate that testing facilitates learning even when item-selection effects are present in the data. For example, Modigliani (1976) showed that increasing the delay before an initial test led to increasingly greater effects of testing (Jacoby, 1978; Karpicke & Roediger, 2006a), and when the enhancement effects due to testing were mathematically separated from item-selection effects, the positive effects of delaying the initial test were attributed entirely to enhancement effects, whereas item-selection estimates remained invariant across the delays (and were quite negligible to begin with). Other procedures for handling item-selection problems were developed by Lockhart (1975) and Bjork, Hofacker, and Burns (1981) and show similar results. To conclude, the testing effect is not simply a result of additional exposure, or overlearning, or item-selection artifacts.

Effortful Retrieval and Desirable Difficulties

If additional exposure and overlearning cannot explain the testing effect, then the alternative is that some aspect of the retrieval process itself must be at work. This is what Dempster (1996) called the retrieval hypothesis. A variety of ideas about how retrieval may affect later retention have been advanced, although they may be describing the same process in somewhat different words. Various writers have argued that retrieval effort causes the testing effect (e.g., Gardiner, Craik, & Bleasdale, 1973; Jacoby, 1978). Alternatively, retrieval may increase the elaboration of a memory trace and multiply retrieval routes, and these processes may account for the testing effect (e.g., Bjork, 1975, 1988; McDaniel, Kowitz, & Dunay, 1989; McDaniel & Masson, 1985). We consider these ideas in turn, but note that they need not be mutually exclusive.

One explanation for why tests that require production, or recall, of material lead to greater testing effects than tests that involve identification, or recognition, is that recall tests require greater retrieval effort or depth of processing than recognition tests (Bjork, 1975; Gardiner et al., 1973). Bjork (1975) argued that depth of retrieval may operate similarly to depth of processing at encoding (e.g., Craik & Tulving, 1975), and that deep, effortful retrieval may enhance the testing effect. As already discussed, increasing the spacing of an initial test—which can be assumed to increase retrieval effort—promotes better retention (Jacoby, 1978; Karpicke & Roediger, 2006a; Modigliani, 1976), so long as material is still accessible and able to be recalled on the test (Spitzer, 1939) or feedback is provided after the test (Pashler et al., 2003). This positive testing effect probably reflects greater retrieval effort on delayed tests.

Other evidence from different sorts of research also leads to the general conclusion that retrieval effort enhances later retention. Gardiner et al. (1973) asked students general knowledge questions and measured the amount of time it took them to answer the questions. At the end of the session, they gave subjects a final free-recall test on the answers. The longer it took subjects to produce the answer to a question (indicating greater retrieval effort), the more likely they were to recall the answer on the final test (see also Benjamin, Bjork, & Schwartz, 1998). In a similar line of research, Auble and Franks (1978) gave subjects sentences that were initially incomprehensible (e.g., *The home was small because the sun came out*) and varied the amount of time before they provided a key word that made the sentences comprehensible (*igloo*). They found that the longer subjects puzzled over the incomprehensible sentences (making an "effort toward comprehension"), the greater their retention of the sentence on a final test. These studies demonstrate the positive effects of retrieval effort on later retention, and the testing effect reflects another example of retrieval effort promoting retention.

Other experiments have examined the multiplexing of retrieval routes by using the technique of varying cues given on a first test to examine how the type of retrieval on the first test affects performance on a second test given later (e.g., Bartlett, 1977; Bartlett & Tulving, 1974; McDaniel et al., 1989; Mc Daniel & Masson, 1985). The general finding is that the nature of the cues on the first test can affect how much that test enhances performance on the second test (although in some case, the exact nature of the experimental design matters; see McDaniel et al., 1989, p. 434). For example, McDaniel and Masson (1985) manipulated whether studied words were processed with semantic or phonemic encoding tasks, the typical levels-of-processing manipulation (Craik & Tulving, 1975). Soon after study, subjects were given cued-recall tests with phonemic or semantic cues, and the cues either matched or mismatched the type of initial encoding. Subjects took a final cued-recall test 24 hr later. (There were also conditions in which items were tested only on the second test, to assess the testing effect.) McDaniel and Masson found that the testing effect that appeared on the second test was greater when the cues for the first test mismatched the original encoding and yet successful retrieval occurred than when the cues on the first test and the type of encoding matched. This result can be understood as due to an increase in the types of retrieval routes that permit access to the memory trace (or perhaps a multiplexing of the features of the memory trace itself).

Recently, Jacoby and his colleagues have obtained direct experimental evidence for different depths of retrieval in a memory-for-foils paradigm (Jacoby, Shimizu, Daniels, & Rhodes, 2005; Jacoby, Shimizu, Velanova, & Rhodes, 2005). In this type of experiment, subjects encode material under shallow or deep encoding conditions. During a first recognition test, subjects discriminate between old words that were studied under either the shallow or the deep conditions and new items (foils or lures). They are later given a second recognition test that assesses memory for the foils on the first test. For college students, having taken the first recognition test with the meaningfully studied (or deeply studied) items enhanced recognition of foils on the later test, compared with having taken the first test with the shallowly studied items. Interestingly, older adults did

not show this difference (Jacoby, Shimizu, Velanova, & Rhodes, 2005), but for present purposes, the critical aspect of these studies is that manipulation of the depth of retrieval on the first test produced a large effect on recognition of the foils on the later test among younger adults.

Bjork and Bjork (1992) developed a theory to explain the testing effect and other effects of retrieval effort. They distinguished between *storage strength*, which reflects the relative permanence of a memory trace or permanence of learning, and *retrieval strength*, which reflects the momentary accessibility of a memory trace and is similar to the concept of retrieval fluency, or how easily the memory represented by the trace can be brought to mind. Their model assumes that retrieval strength is negatively correlated with increments in storage strength; that is, easy retrieval (high retrieval strength) does not enhance storage strength, whereas more effortful retrieval practice does enhance storage strength and promotes more permanent, long-term learning. However, because students often use the fluency of their current processing (retrieval strength) as evidence about the status of their current learning (e.g., see Jacoby, Bjork, & Kelly, 1994), they may elect poor study strategies. That is, students may choose strategies to maximize fluency of their current processing, even though conditions that involve non-fluent processing may be more beneficial to long-term learning. For example, students may prefer massed study (or repeated rereading) because it leads to fluent processing, although other strategies (such as spaced processing or effortful self-testing) would lead to greater long-term gains in knowledge.

Bjork (1994, 1999) has referred to techniques that promote long-term retention even though they slow initial learning as desirable difficulties and has argued that teachers should focus on creating desirable difficulties for students in order to enhance their learning. Techniques such as spaced practice (relative to massed practice) and delayed feedback (relative to immediate feedback) constitute desirable difficulties. We have argued that relative to studying, testing also constitutes a desirable difficulty (Roediger & Karpicke, 2006). Repeated testing tends to slow initial learning relative to repeated studying (as evidenced on final tests at a short retention interval), but testing promotes far greater long-term retention (e.g., see Fig. 7).

Not surprisingly, people often do not voluntarily engage in difficult learning activities, even though such activities may improve learning. To give but one relevant example, Baddeley and Longman (1978) trained postal workers on typing and keyboard skills under massed- or spaced-practice conditions. The subjects reported that they preferred the massed-practice condition (and some refused to participate in further spaced-practice training), even though spaced practice promoted far better retention than massed practice. In many contexts, conditions that lead to rapid gains in initial learning will produce poor long-term retention, and likewise, conditions that make learning slower or more effortful often enhance long-term retention, with the testing effect being an example of the latter scenario. To the extent that students monitor and guide their learning on the basis of the fluency of their current processing, they may fall prey to illusions of competence, believing that their future performance will be greater than it really will be (see Bjork, 1999; Jacoby et al., 1994; Koriat & Bjork, 2005, in press). Because repeated testing is more effortful than repeated studying, students may choose not to test themselves while learning, and likewise, teachers may choose not to give many tests in their classes. Implementing test-enhanced learning as a desirable difficulty remains a challenge for education.

Transfer-Appropriate Processing

The concept of transfer-appropriate processing is also useful in understanding the testing effect, although it should be seen as perhaps incorporating some of the ideas discussed earlier in this section at a more general level. Encoding may emphasize many different strategies and types of processing, such as rote or meaningful processing, as described in the levels-of-processing tradition (Craik & Tulving, 1975), or item-specific (focused on isolated facts) or relational (focused on relating ideas) processing, as described in a different framework (Hunt & McDaniel, 1993). The idea behind transfer-appropriate processing is that performance on a test of memory benefits to the extent that the processes required to perform well on the test match encoding operations engaged during prior learning (Morris et al., 1977; see also Kolers & Roediger, 1984; McDaniel, Friedman, & Bourne, 1978). Thus, the same study strategies or processes of encoding that may greatly aid performance on one type of test may have no effect or even an opposite effect on a different type of test that emphasizes different types of information or processing (e.g., Blaxton, 1989; Fisher & Craik, 1977). The idea is similar to the encoding-specificity principle (Tulving & Thomson, 1973) and emphasizes the critical relation between encoding and retrieval processes. The concept of transfer-appropriate processing has been applied to a wide array of phenomena. For example, Roediger, Weldon, and Challis (1989) argued that transfer-appropriate processing is critical for understanding differences between performance on explicit and implicit memory tests (see also Blaxton, 1989; Roediger, 1990).

McDaniel (in press) pointed out that all situations in which information is learned and then expressed through tests or actions involve transfer. He noted that although the idea of transfer-appropriate processing seems obvious in prospect, in practice it is often violated. He used the example of a teacher who encourages excellent classroom study strategies that permit deep understanding of the core concepts of the subject and how they relate to one another, but then gives students a multiple-choice test emphasizing recognition of isolated facts and wonders why the students perform so poorly. In this case, relational processing strategies (although they may be good for long-term retention) are poor for the specific test that the instructor gives. Thomas and McDaniel (in press) provided experimental evidence to bolster this point. Educators make the same point about

standardized tests; such tests may assess what is easy to measure rather than the complex skills students may develop in class.

In applying transfer-appropriate processing to education, the key question is what knowledge and skills the instructor wants the students to know when they leave the course. One goal would be being able to retrieve the information when it is needed, and retrieval practice is critical to developing this skill. Taking tests allows students to engage in retrieval operations during learning and thus to practice the same skills needed to enhance subsequent retrieval. Such retrieval practice in taking tests permits greater retention than does engaging in additional encoding operations such as repeated reading (Roediger & Karpicke, 2006). Transfer-appropriate processing provides an explanation for why taking memory tests often enhances performance on later memory tests, especially when effortful retrieval is required. The results we have reviewed show that testing under conditions of effortful retrieval has a greater transfer effect on later test performance than testing under conditions of easy retrieval. Of course, another educational goal is to have students transfer information learned in courses to new problems they face later in their jobs, but this kind of distant transfer is more difficult to study although it remains a target for future research (see Barnett & Ceci, 2002, for a review).

We believe that the concept of transfer-appropriate processing offers an intuitive explanation for the somewhat counterintuitive testing effect, and for this reason, the concept may be useful in helping educators understand why taking tests should benefit learning—testing leads students to engage in retrieval processes that transfer in the long term to later situations and contexts. However, we note one drawback to this approach. One prediction that may be drawn from transfer-appropriate processing is that performance on a final test should be best when that test has the same format as a previous test. As we have shown, the general finding is that recall tests promote learning more than recognition tests, regardless of the final test's format (e.g., Kang et al., in press). This result needs confirmation through additional experiments, but if it is true, it would seem to be good news for educators, because it would lead to a straightforward recommendation for educational practice. Nonetheless, the same outcome (e.g., better transfer from a short-answer test than from a multiple-choice test to a later multiple-choice test) may be construed as inconsistent with transfer-appropriate processing. However, it may not be inconsistent with the broader idea embodied in transfer-appropriate processing. If, for example, a final multiple-choice test requires effortful retrieval and a prior short-answer test fostered such effortful processes more than a prior multiple-choice test did, then it could be understandable that the prior short-answer test leads to better final performance than the prior multiple-choice test. We realize that such reasoning can quickly become circular and invulnerable to disconfirmation; the real challenge for the future is to specify how transfer-appropriate processing ideas apply to educational contexts so that they can be tested, as Thomas and McDaniel (in press) have recently done in one situation.

Summary

The testing effect cannot be explained by additional exposure to the material. This suggests that retrieval processes engaged in during a test are responsible for enhancing learning. More specifically, elaboration of encoding, more effortful or deeper encoding, and creation of different routes of access can account for the basic effect. Further, proponents of each of these ideas can point to evidence consistent with their viewpoint. The concept of transfer-appropriate processing is also congenial, albeit at a general level, to explaining the testing effect. It seems safe to say that empirical efforts to understand the testing effect have outstripped theoretical understanding, but the database is now firm enough to permit deeper understanding of the effect at a theoretical level and does permit the conclusive rejection of at least one prominent theory, that the testing effect is due to additional study, or overlearning.

DYNAMIC TESTING AND FORMATIVE ASSESSMENT

We have emphasized that the act of testing memory can have powerful effects on learning and later retention of material. Other perspectives on testing have also emphasized that learning can occur during testing or that tests may be used to promote learning through mediated effects. In this section, we describe two approaches that are complementary but that developed in different contexts to serve different aims. In the area of mental-abilities testing, dynamic testing uses tests and test feedback to assess students' learning potential rather than the products of previous learning, providing a more accurate measure of students' ability to learn. In education, the practice of formative assessment involves the use of testing to give feedback to teachers and students that may guide future classroom practices. The common thread between these techniques is using testing to generate feedback that can be used to assess learning potential or to promote future learning.

Dynamic Testing

Many tests of mental abilities, like IQ tests or the SAT, measure developed abilities and the results of prior learning but are aimed at assessing general learning capabilities. These tests can be considered static tests, because they involve an examiner giving the test to the examinee without providing feedback about performance during the test (except perhaps for overall results at some later point in time). Grigorenko and Sternberg (1998) and Sternberg and Grigorenko (2001, 2002) have advocated the use of dynamic testing procedures instead of static testing to measure individuals' strengths and weaknesses in cognitive skills, as well as their learning potential. Dynamic testing is another example of using tests to promote learning in addition to merely assessing learning.

The key difference between static and dynamic testing is that dynamic testing uses feedback to measure learning during the test. In both static and dynamic testing, the examiner gives students a series of problems that typically become progressively more difficult. However, in dynamic testing, the examiner gives the students feedback about their performance after each problem on an initial test in order to help improve their scores on a second test. Feedback is typically aimed at helping the students understand the principles underlying their errors. When the students are retested, performance gains from the pretest to the posttest reflect their ability to learn from the feedback, which is indicative of their learning potential. Thus, individuals learn during dynamic testing, and the assessment procedure is used to improve learning and simultaneously to measure learning potential.

Sternberg and Grigorenko (2001, 2002) have argued that dynamic tests not only serve to enhance students' learning of cognitive skills, but also provide more accurate measures of ability to learn than do traditional static tests. For example, some students may not have had sufficient educational experience to perform well on traditional static tests. Although these students would show poor performance on such tests, they may respond well to the feedback given in dynamic testing and thus demonstrate their learning potential. Sternberg and Grigorenko gave an example of administering mental-abilities tests to children in rural Tanzania, who had not received the same levels of education as do children in Western cultures. Not surprisingly, the Tanzanian children performed poorly on static tests relative to the typical performance levels of Western children of the same ages, because in their schooling they had not learned the skills required to do well on the tests. On the basis of the results of only static tests, one might conclude that the Tanzanian children were simply not intelligent or perhaps mentally retarded. However, when the children were given dynamic tests involving feedback during the initial test and a second assessment of their performance, they demonstrated their underlying capacities to learn by improving their scores when retested (see Sternberg et al., 2002). Static and dynamic testing procedures clearly led to different conclusions about the learning abilities of these Tanzanian children. The implication of dynamic testing is that standardized testing can be used to promote learning if meaningful feedback about test performance is given to students.

Formative Assessment

An idea similar to that embodied in dynamic testing is gaining currency in education. *Formative assessment* refers to the general procedure of using the results of classroom assessments as feedback for teachers to guide future instruction and also for students to guide their future studying (Black & Wiliam, 1998a, 1998b). Formative assessment is often contrasted with summative assessment, a distinction similar to that between dynamic and static testing. Most tests in education are summative: The tests are used to summarize performance, to measure prior learning, and often to assign grades and to rank students. Summative assessments become formative assessments when teachers use the results of their classroom assessments to change classroom instruction and to promote further learning (see Leahy, Lyon, Thompson, & Wiliam, 2005; McTighe & O'Connor, 2005). Proponents of formative assessment define the concept of assessment broadly to include formal classroom testing, the teacher's interactions with students in class (how the students answer questions), and other more dynamic classroom activities (e.g., working in groups). In all these cases, feedback is used to guide future teaching practices. Thus, formative assessment is often referred to as assessment *for* learning, in contrast to assessment *of* learning.

The evidence shows that formative assessment promotes learning. Black and Wiliam (1998a) carried out an extensive review of 250 studies of formative assessment and showed that classrooms that used formative assessments promoted better student performance than those that used only summative assessments. In addition, Black and Wiliam identified several areas in which teachers may improve their use of formative assessment. These areas included giving students elaborate and detailed feedback about how to improve their work and then giving them the opportunity to improve, providing students with clear performance goals, and instructing students on how to use the feedback on their tests to improve their performance.

Despite the benefits of formative assessment, many teachers may choose not to use its techniques because they can be difficult to implement. Obviously, it is much easier to administer tests to students as summative measures, grade the tests, and not change one's teaching practices in response to the test results. (In fact, this is probably what most teachers do.) Just as the concept of desirable difficulties may be useful for students (because often the study strategies that produce more difficult initial learning are better for long-term retention; Bjork, 1994, 1999), so too might the concept of desirable difficulties be useful for teachers. Often the best instruction may require teachers to implement the difficult process of using tests to assess performance and then changing the style and content of their teaching on the basis of the outcome of the tests. Even though using formative-assessment techniques may be difficult for teachers, the evidence shows that they benefit students' learning.

Summary

The techniques of formative assessment and dynamic testing are examples of how tests can be used to enhance learning. However, the learning benefits gained from use of these techniques are not examples of the direct effects of testing on learning, which we emphasized earlier. Formative assessment and dynamic testing illustrate mediated effects of testing on learning: The test gives knowledge about current levels of performance, and improvements in learning (as indexed by later tests) occur because of the studying or instruction that occurs between tests—and the instruction is guided by the performance on the test. Many educators have argued against simply using

summative assessments. For example, McTighe and O'Connor (2005) wrote, "By themselves, summative assessments are insufficient tools for maximizing learning. Waiting until the end of a teaching period to find out how well students have learned is simply too late" (p. 11). Quite to the contrary, the evidence for the testing effect, which we have described in this review, suggests that taking tests, even without feedback, can enhance later retention of material. Frequent classroom testing can both directly aid students' learning and give teachers continuous assessments of how well students are learning, so that they can review material that many students did not understand and change teaching strategies appropriately. In short, frequent testing in the classroom can serve dual purposes of direct enhancement of learning and mediated enhancement in the form of dynamic testing and formative assessment.

POSSIBLE NEGATIVE CONSEQUENCES OF TESTING

We have emphasized the positive effects of testing, but there can also be negative effects in some situations. In this section, we describe two classes of problems: how the act of recall during a test can sometimes impair recall of material that is tested later and how certain types of tests can produce a negative influence on a person's knowledge as expressed on a later test.

Interfering Effects of Recall

The act of recall increases the probability of later recall for the tested material, but sometimes can impair recall of other studied material. Using a paired-associate learning paradigm and varying the order in which items were tested, Tulving and Arbuckle (1963, 1966) observed a decline across output positions, such that items tested later in a sequence were recalled worse than those tested earlier. This pattern held over rather short time intervals in their experiments, and so arguably could be only a short-term effect. That is, the act of recalling some pairs could serve as a distractor task, and distractor tasks reduce recall over the short term (e.g., Glanzer & Cunitz, 1966). However, Smith (1971) reported similar findings when subjects recalled lists of categorized items (types of fruit, birds, articles of furniture, etc.) in situations in which short-term memory effects did not play a role, and this finding has been replicated (e.g., Roediger & Schmidt, 1980). Analogous processes can occur in recognition memory (e.g., Neely, Schmidt, & Roediger, 1983). To our knowledge, such output-interference experiments have never been performed with educationally relevant tests in which output order of questions (say, in a multiple-choice test) is counterbalanced across subjects. (The experiment could be done in classroom conditions without students even being aware of the manipulation.) The inhibitory effects observed in output-interference paradigms are often not large, and we expect that even if they occur in the classroom, they simply reduce the size of the positive testing effect for items tested late in the sequence. That is, we expect that the overall effect of testing will be positive despite output-interference effects, but this hypothesis needs to be tested.

Another type of interference phenomenon occurs when people are given cues for part of a set of material with instructions to recall the entire set (Slamecka, 1968). For example, Roediger (1978) gave students categorized lists that contained five words from each of 10 categories. On the test, the students were instructed to recall the entire set of items from all the categories; different groups of students were given no, three, five, or seven category names to aid recall. The groups receiving category names recalled well the items from the cued categories (relative to the free-recall subjects, who received no category-name cues); however, they did worse than the free-recall subjects in recalling categories for which they received no cues. That is, subjects first recalled items from the cued categories, and this act of recall inhibited their ability to access the other categories that had been part of the list; further, the more categories subjects were given as cues, the greater was the inhibition. Recall thus seems to be a self-limiting process; the act of recalling some items inhibits recall of others (see also J. Brown, 1968).

M.C. Anderson, Bjork, and Bjork (1994) developed a similar within-category paradigm, sometimes called the retrieval practice paradigm. The experiment involves three phases—study, retrieval practice, and final test—and thus has a design similar to that of experiments on the testing effect. During the study phase, subjects are exposed to items such as *fruit-orange* and *fruit-banana* (always a category name and category member). During the practice phase, they are cued with items such as *fruit-o____* and are asked to recall the missing words. Some of the study items are tested and others are not. During the final test, subjects are given the category name by itself and are asked to recall all the items from the studied list belonging to that category. Many experiments have shown that final recall of the items that received retrieval practice is greatly enhanced (a version of the testing effect), whereas recall of items that were not practiced is inhibited.

There is by now a large literature on the phenomenon of retrieval-induced forgetting (see M.C. Anderson, 2003, for a review). For present purposes, one boundary condition is that well-integrated materials do not show the effect (M.C. Anderson & McCulloch, 1999). Therefore, retrieval-induced forgetting and related phenomena may not occur with highly interrelated materials such as textbook chapters and lectures, although this is an open question. In fact, there is some evidence that in testing of prose materials, retrieval of some facts from the text may facilitate recall of other, untested facts. Chan et al. (in press) had subjects read passages with interrelated facts and concepts about geographic and historical topics. For example, one of the passages was about toucans, and the body of the text stated that toucans sleep in tree holes at night and that woodpeckers create these tree holes (because toucans have soft bills, they cannot make tree holes and must sleep in those made by woodpeckers). Chan et al. prepared two tests containing some items that were

related across the tests. For example, the first test asked, "Where do toucans sleep at night?" and the second test asked, "What other bird species is the toucan related to?" Because answering the first question ("tree holes") might have activated information about woodpeckers (the answer to the question on the second test), a positive testing effect might have been expected in this situation, and this is just what Chan et al. found. Further studies are needed to determine how well this finding generalizes to other types of prose materials, but clearly, testing does not always cause retrieval interference, as M.C. Anderson (2003) has also noted.

In sum, although various types of recall interference are quite real (and quite interesting) phenomena, we do not believe that they compromise the notion of test-enhanced learning. At worst, interference of this sort might dampen positive testing effects somewhat. However, the positive effects of testing are often so large that in most circumstances they will overwhelm the relatively modest interference effects. The types of interference we discuss next may prove more problematic for certain types of tests.

Negative Suggestion Effects

If people learn from tests (and we have reviewed copious evidence showing that they do), then what happens if people are presented erroneous information on tests? Although teachers would never deliberately provide false or misleading information during class or in reading materials, they routinely do so on some of the most popular kinds of tests that they give: multiple-choice and true/false tests. Many multiple-choice tests present three erroneous answers along with one correct answer. For some items, students may pick the wrong alternative, and because the act of retrieval enhances later retrieval, they may acquire erroneous information. Similarly, in true/false tests, typically half the items are true and half are false. Students may sometimes endorse false items as being true and thereby learn erroneous information. However, even if they read a false item and know it is false, the mere act of reading the false statement may make it seem true at a later point in time. Hasher, Goldstein, and Toppino (1977) showed that when people were asked to judge the truth of statements on a rating scale, they judged statements they had previously read repeatedly as more likely to be true than new statements of the same sort, regardless of whether they were actually true. This mere-truth effect has been replicated and extended (e.g., F.T. Bacon, 1979; Begg, Armour, & Kerr, 1985) and is presumably due to the fact that a familiar statement has a true ring to it ("Yes, I think I remember hearing that somewhere").

Many years ago, Remmers and Remmers (1926) coined the term *negative suggestion effect* to refer to the increased belief in erroneous information that students may acquire from tests. The topic is quite important for practical reasons, so it is rather surprising that it has not been more thoroughly studied. However, the studies that have been reported all show that the effect is quite real. We briefly review what is known.

Toppino and Brochin (1989) gave students true/false tests and later asked them to judge the truth of objectively false statements to which they had been exposed, mixed in with similar false items to which they had not been exposed. The repeated items were judged as truer than the new items. Toppino extended this finding to the distractor items from multiple-choice tests in a later experiment (Toppino & Luipersbeck, 1993; see also Rees, 1986). A.S. Brown, Schilling, and Hockensmith (1999) exposed subjects to misinformation after an original test and showed negative effects on later cued-recall and multiple-choice tests. In some cases, they told subjects that the erroneous information was false, yet the negative effects were still obtained. In other work, A.S. Brown (1988) and Jacoby and Hollingshead (1990) showed that if students were exposed to misspelled words, this experience caused them to misspell the words later on an oral spelling test. Jacoby and Hollingshead pointed out that this sort of negative suggestion effect does not affect only students; teachers' spelling may get worse from reading frequent misspellings in students' papers.

Roediger and Marsh (2005) examined the negative suggestion effect from multiple-choice tests in a design in which testing effects could also be measured. At issue was the question of whether negative suggestion effects are so pernicious as to overcome the positive effects of testing. Subjects read 18 short nonfiction passages about a wide variety of topics, including science, geography, famous people, and animals. The students then took a multiple-choice test covering both these passages and 18 other passages that were not read. The two types of items (read vs. nonread) provided conditions analogous to students' having studied versus not having studied for a test. The items on the multiple-choice test had two, four, or six alternatives (one correct answer and one, three, or five lures). The number of incorrect answers was varied to test whether being exposed to more incorrect answers would increase the negative suggestion effect, if one were found. Results on the multiple-choice test showed, not surprisingly, that students did much better answering items on passages they had read than answering items on passages they had not read, and that their performance became worse as the number of distractors on the multiple-choice test increased.

The results of primary interest come from the third phase of the experiment, which involved a cued-recall test that asked the same questions that had been on the multiple-choice test but without any alternatives from which to choose. This final test also asked questions about information that had not appeared on the multiple-choice test, so that a baseline could be established to examine whether the multiple-choice test produced a testing effect. Instructions on the final test warned students not to guess and to be sure of any answers they produced. The results are presented in Table 4. There was a large testing effect for both the read passages and the nonread passages. For the read passages, correct recall was 63% for tested items (averaged across the number-of-distractors variable) and only 40% for nontested items. For nonread passages, the corresponding figures were 29% and 16%, so students learned from testing even when they

TABLE 4

Proportion of Final Cued Recall as a Function of Experimental Condition in Roediger and Marsh (2005)

Measure and condition	Number of multiple-choice alternatives on the initial test			
	Zero (not tested)	Two	Four	Six
Proportion correct				
Read passages	.40	.67	.61	.61
Nonread passages	.16	.34	.28	.26
Proportion of lures recalled				
Read passages	.04	.06	.08	.09
Nonread passages	.06	.09	.13	.15

had not read the relevant material. However, a negative suggestion effect is also apparent in the data, because correct responding on the tested items decreased as a function of the number of prior distractors on the multiple-choice test. Yet even after subjects took a six-alternative test, their performance on the cued-recall test was better for the previously tested items than for the items that had not been tested; this was true for both read and nonread passages.

The error data at the bottom of Table 4 tell a similar story. Erroneous recall was greater for the items that had been tested on the multiple-choice test than for the nontested items, and the error rate increased with the number of distractors on the prior test. These errors occurred despite stringent instructions to subjects not to guess. These results probably underestimate the negative suggestion effect of multiple-choice tests in educational settings, because students taking exams usually are not penalized for guessing and likely make more errors than in this experiment, at least under similar sorts of conditions (Marsh, Fazio, & Roediger, 2006).

Butler, Marsh, Goode, and Roediger (in press) sought to determine why the negative suggestibility effect arises from multiple-choice tests and to reconcile this effect with list-learning experiments by Whitten and Leonard (1980) that showed positive effects from the number of distractors on a multiple-choice test on later free recall. Butler et al. concluded from three experiments that the level of performance on the multiple-choice test is the key factor. If the multiple-choice test is very easy, so that the correct answer can almost always be selected, then the number of lures on the test seems to exert a positive effect—recall of the target item on the later test increases as the number of lures on the initial test increases. However, under more realistic conditions, when multiple-choice performance is far from perfect, negative suggestion effects occur and become larger as the number of distractors on the multiple-choice test increases. In related research, Butler and Roediger (2006) showed that if students are given feedback soon after taking the multiple-choice test, the negative suggestion effect is eliminated. The educational implication is clear—students should receive feedback promptly after taking a test—but often this practice is not followed in classrooms, either because of practical difficulties (large classes with many tests to be graded) or because teachers do not want banks of test items to be distributed among students.

Negative suggestion effects are not restricted to multiple-choice and true/false tests. For example, on short-answer and essay tests, students who do not know the correct answer may try to write as intelligently as possible about the subject at hand in hopes of earning some points. Classroom tests implicitly use what Jacoby (1991) has called inclusion instructions, because in answering a question, students can include what they learned in class and what they knew before taking the class, and they can also guess the answer for purposes of trying to do well on the test. Penalties for guessing are rarely imposed. Laboratory evidence shows that subjects who erroneously recall information on one test are quite likely to do so again on a later test (e.g., McDermott, 2006; Meade & Roediger, 2006; Roediger, Wheeler, & Rajaram, 1993). Because laboratory studies show a testing effect for erroneously recalled information, as well as for correct information, we assume similar effects occur in classroom settings.

Summary

In this section, we have considered two types of negative effects of testing: interfering effects of recall and negative suggestion effects. Both are real, and both are interesting, but in our opinion, neither undermines our advocating the frequent use of testing in classrooms. Most of the results show that the magnitude of negative suggestion effects is not so great as to undercut the large positive effects of testing. Of course, there may be other negative effects of testing, such as test anxiety and stereotype threat (see Steele, 1997), but these have been shown to apply to standardized tests such as the SAT and may not apply to classroom testing. We suspect that if classroom testing were made more routine and there were few "big tests" that counted for most of a student's grade, phenomena such as test anxiety and stereotype threat would diminish through habituation (see Leeming, 2002, for some evidence supporting this speculation). Future research is needed to put these conclusions on a firmer foundation.

OBJECTIONS AND CAVEATS

When we and our colleagues have proposed our ideas on test-enhanced learning to various audiences in recent years, we have met with varying reactions, from enthusiastic endorsement to stunned disbelief that anyone could be seriously suggesting increased testing in the schools. People who have the latter reaction, who are often in schools of education, raise several points that we now consider in turn.

First, these critics say that there is already too much testing in the schools and that increasing the amount would be even worse. However, what they usually mean is that there is too much standardized testing in schools. As we have discussed, our aim

in encouraging testing is not to increase the number of standardized assessment tests given (although we do believe they can be useful indicators of students' knowledge, or lack thereof). The one exception is that we do advocate using standardized testing in ways envisioned by Grigorenko and Sternberg in their dynamic-testing program. This use of standardized tests seems excellent, as the gathering evidence shows (Sternberg & Grigorenko, 2002).

Second, critics object that taking valuable classroom time for testing will deprive students of other activities, such as lectures, exercises, creative use of materials, group discussion, and so on. After all, learning material and then being tested on it smacks of the "drill and practice" routines that seem to foster rote learning and bored students. We have several replies to this objection. We certainly encourage the current emphasis on creativity in the classroom, but we do not view testing as inimical to creative uses of knowledge. If students have not mastered basic knowledge of the subject matter, they have no chance of thinking critically and creatively about the subject, and testing can help students acquire this body of knowledge. Also, we note that teachers in certain situations know that testing works, and they recommend it. (In fact, when we discuss our ideas and evidence with teachers in elementary schools, they are usually enthusiastic.) When multiplication tables are taught in the primary grades, teachers have students create flash cards with a problem on one side and the answer on the other side. Students are taught to test themselves at home to prepare for the test of this exact nature that they will take in class. Such self-testing works, as shown by Gates (1917) long ago. The same is true for learning foreign-language vocabulary, another task for which flash cards and similar learning strategies are routine. The testing we advocate is simply an extension of these strategies that are already used as study tools, in the classroom and at home, in certain circumstances.

Of course, we do not mean to imply that testing works only for multiplication and foreign-language vocabulary, and that the general strategy cannot be adapted for more complex learning situations and materials. We believe frequent testing (or more neutrally, frequent assignments to be handed in) will increase learning at all grade levels. If the principle of transfer-appropriate processing is applied to educational settings, the types of assignments and tests given in class can be determined depending on the nature of the class and the type of learning desired. The fundamental question is what knowledge the teacher would like the students to take from the class and be able to use in (transfer to) other situations. Key principles should be emphasized in class and should be tested repeatedly. Test formats should be appropriate to the knowledge structures that are desired. Exclusive reliance on multiple-choice tests or true/false tests that examine only specific items and tidbits of information (say, only names and dates in a history class) will lead students to study and retain only such item-specific information (see Hunt & McDaniel, 1993; McDaniel & Einstein, 1989; Thomas & McDaniel, in press). If teachers are interested in fostering creativity, then they can construct creative essay questions for daily (or weekly) tests that cause students to combine domains of knowledge. Even multiple-choice tests need not assess knowledge of rote facts; teachers can create questions requiring more complex reasoning according to Bloom's (1956) taxonomy of types of knowledge and types of questions. However, creating such thought-provoking multiple-choice questions is difficult.

A third issue, which relates to the second, is whether our proposal of testing is really appropriate for courses with complex subject matters, such as the philosophy of Spinoza, Shakespeare's comedies, or creative writing. Certainly, we agree that most forms of objective testing would be difficult in these sorts of courses, but we do believe the general philosophy of testing (broadly speaking) would hold—students should be continually engaged and challenged by the subject matter, and there should not be merely a midterm and final exam (even if they are essay exams). Students in a course on Spinoza might be assigned specific readings and thought-provoking essay questions to complete every week. This would be a transfer-appropriate form of weekly "testing" (albeit with take-home exams). Continuous testing requires students to continuously engage themselves in a course; they cannot coast until near a midterm exam and a final exam and begin studying only then.

Finally, critics ask us whether using frequent testing in the classroom (and encouraging students to use self-testing to study) can work at all levels of education and with all types of students. Of course, this is an empirical question, and it is too early to answer it with certainty. However, many elementary schools do have frequent testing (e.g., spelling and vocabulary tests every Friday). Although we are not aware of any concrete data on the subject, we suspect from talking to teachers at various levels in the educational system that the frequency of classroom tests declines throughout the years in American education, with classes in colleges and universities representing a nadir. In many large college classes, there may be a midterm and a final exam using only multiple-choice or other objective questions.

Some critics wonder if college students would not rebel in shock at the introduction of weekly or even daily testing. We suspect not. Frank Leeming (2002) at the University of Memphis described a system of frequent testing that worked very well and that students enjoyed. Roberto Cabeza at Duke University also employs daily testing in his cognitive psychology course with positive results. David Pisoni at Indiana University has students use the Internet to answer questions about the main points covered in class after each lecture in courses on cognitive psychology and language and cognition.

Our colleague Kathleen McDermott gives daily tests in her undergraduate human memory course. The class meets twice a week for 1.5 hr, and the last 10 min of each class are used to quiz students on the assigned reading for that day and the lecture material that they have just heard. In addition, three longer exams are given, each covering a third of the course content. The process requires students to keep up with the reading

assignments, to attend class, and to pay attention, and the ratings of the course are very high. One student's comment is representative: "I liked having quizzes at the end of each class, because they didn't add too much pressure and caused me to focus and to retain information during each class better."

In short, we see no reason why students at all levels of education cannot profit from a system of frequent testing. Of course, the form of testing would depend on the nature of the course, as we discussed earlier. In the case of very large introductory courses, quizzes with an objective format might be given once a week, with immediate feedback (so as to correct errors and overcome negative suggestion effects). As previously noted, small upper-level courses might replace frequent testing with frequent assignments (e.g., written essays on the material) that require students to remain continuously engaged.

Will frequent testing work as a strategy with all types of students? Again, this is an empirical question, but we suspect that the answer is "yes." Strong students should thrive because they will prepare for class (and, in fact, they may use self-testing as a study method already). Weaker students who know that they will be quizzed frequently may try to keep up with the course more systematically than if they had tests only once or twice during a semester. One reason to expect that poorer students might benefit from testing is that several researchers have successfully used repeated testing as a way to teach information to memory-impaired individuals (e.g., Camp, Bird, & Cherry, 2000; Schacter, Rich, & Stampp, 1985).

No single change to educational practice is a panacea, but from the evidence we have reviewed in this article, we believe that testing (or continuous assignments that function as tests) has the important effect of enhancing learning of the tested material. We also believe that testing causes students to study more in preparation for the tests. Tests serve as a motivator to keep up with course assignments and to engage in study activities.

CONCLUSION

Testing is a powerful tool to enhance learning. Many laboratory studies have demonstrated this point, and the few systematic applications in the classroom have been successful in improving performance. Of course, much remains to be learned, both about basic cognitive mechanisms that lead to the testing effect and about practical applications in classrooms at all levels of education. Although cognitive and educational psychologists have studied testing off and on over the years, we believe the time is ripe for a dedicated and thorough examination of issues surrounding testing and its application in the classroom. The broad ideas of transfer-appropriate processing and creating desirable difficulties provide a guide to how testing may be implemented in the classroom. If teachers determine what critical knowledge and skills they want their students to know after leaving the class, these points can be emphasized in class and tested repeatedly at spaced intervals to ensure that students acquire this knowledge. Frequent testing not only has a direct effect on learning, but also should encourage students to study more, to be continuously engaged in the material, to experience less test anxiety, and probably even to score better on standardized tests. However, this last point remains a promissory note for future research. Direct effects of testing, as well as mediated effects from the use of dynamic testing and formative assessment, have the potential to greatly improve learning in the schools.

Acknowledgments—The writing of this article, as well as some of the research reported, was supported by grants from the Institute of Education Sciences and the James S. McDonnell Foundation. We thank Jane McConnell for her help with preparing the manuscript and for her suggestions. Elena Grigorenko, Sean Kang, and Robert Sternberg provided helpful comments on an earlier draft of the article.

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., III, & McDermott, K.B. (2006). *Examining the testing effect with open- and closed-book tests*. Unpublished manuscript, Washington University in St. Louis, St. Louis, MO.

Allen, G.A., Mahler, W.A., & Estes, W.K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *8*, 463–470.

Anderson, M.C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415–445.

Anderson, M.C., Bjork, E.L., & Bjork, R.A. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 1063–1087.

Anderson, M.C., & McCulloch, K.C. (1999). Integration as a general boundary condition on retrieval-induced forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 608–629.

Anderson, R.C., & Biddle, W.B. (1975). On asking people questions about what they are reading. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 9, pp. 90–132). New York: Academic Press.

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

Bacon, F. (2000). *Novum organum* (L. Jardine & M. Silverthorne, Trans.). Cambridge, England: Cambridge University Press. (Original work published 1620)

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

Baddeley, A.D., & Longman, D.J.A. (1978). The influence of length and frequency of training sessions on the rate of learning to type. *Ergonomics*, *21*, 627–635.

Balota, D.A., Duchek, J.M., & Logan, J.M. (in press). 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. New York: Psychology Press.

Balota, D.A., Duchek, J.M., Sergent-Marshall, S.D., & Roediger, H.L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging*, *21*, 19–31.

Bangert-Drowns, R.L., Kulik, J.A., & Kulik, C.L.C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99.

Barnett, S.M., & Ceci, S.J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*, 612–637.

Bartlett, J.C. (1977). Effects of immediate testing on delayed retrieval: Search and recovery operations with four types of cue. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 719–732.

Bartlett, J.C., & Tulving, E. (1974). Effects of temporal and semantic encoding in immediate recall upon subsequent retrieval. *Journal of Verbal Learning and Verbal Behavior*, *13*, 297–309.

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

Benjamin, A.S., Bjork, R.A., & Schwartz, B.L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*, 55–68.

Birnbaum, I.M., & Eichner, J.T. (1971). Study versus test trials and long-term retention in free-recall learning. *Journal of Verbal Learning and Verbal Behavior*, *10*, 516–521.

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

Bjork, R.A. (1988). Retrieval practice and the maintenance of knowledge. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), *Practical aspects of memory: Current research and issues* (Vol. 1, pp. 396–401). New York: Wiley.

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

Bjork, R.A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 435–459). Cambridge, MA: MIT Press.

Bjork, R.A., & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

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

Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice*, *5*, 7–74.

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*, 139–147.

Blaxton, T.A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 657–668.

Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Essex, England: Harlow.

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

Brown, A.S., Schilling, H.E.H., & Hockensmith, M.L. (1999). The negative suggestion effect: Pondering incorrect alternatives may be hazardous to your knowledge. *Journal of Educational Psychology*, *91*, 756–764.

Brown, J. (1968). Reciprocal facilitation and impairment in free recall. *Psychonomic Science*, *10*, 41–42.

Brown, W. (1923). To what extent is memory measured by a single recall? *Journal of Experimental Psychology*, *6*, 377–382.

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

Butler, A.C., & Roediger, H.L., III. (2006). *Feedback neutralizes the detrimental effects of multiple choice testing*. Unpublished manuscript, Washington University in St. Louis, St. Louis, MO.

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

Calkins, M.W. (1894). Association: I. *Psychological Review*, *1*, 476–483.

Camp, C.J., Bird, M.J., & Cherry, K.E. (2000). Retrieval strategies as a rehabilitation aid for cognitive loss in pathological aging. In R.D. Hill, L. Backman, & A.S. Neely (Eds.), *Cognitive rehabilitation in old age* (pp. 224–248). New York: Oxford University Press.

Carpenter, S.K., & DeLosh, E.L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*, 619–636.

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

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

Cepeda, N.J., Pashler, H., Vul, E., Wixted, J.T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380.

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

Cooper, A.J.R., & Monk, A. (1976). Learning for recall and learning for recognition. In J. Brown (Ed.), *Recall and recognition* (pp. 131–156). New York: Wiley.

Craik, F.I.M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*, 143–148.

Craik, F.I.M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*, 268–294.

Crooks, T.J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*, 438–481.

Crowder, R.G. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

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

Cull, W.L., Shaughnessy, J.J., & Zechmeister, E.B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic:

Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, *2*, 365–378.

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

Deese, J. (1958). *The psychology of learning*. New York: McGraw-Hill.

Dempster, F.N. (1996). Distributing and managing the conditions of encoding and practice. In E.L. Bjork & R.A. Bjork (Eds.), *Human memory* (pp. 197–236). San Diego, CA: Academic Press.

Dempster, F.N. (1997). Using tests to promote classroom learning. In R.F. Dillon (Ed.), *Handbook on testing* (pp. 332–346). Westport, CT: Greenwood Press.

Deputy, E.C. (1929). Knowledge of success as a motivating influence in college work. *Journal of Educational Research*, *20*, 327–334.

Donaldson, W. (1971). Output effects in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*, 577–585.

Duchastel, P.C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology*, *6*, 217–226.

Duchastel, P.C., & Nungester, R.J. (1981). Long-term retention of prose following testing. *Psychological Reports*, *49*, 470.

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

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

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H.A. Ruger & C.E. Bussenius, Trans.). New York: Dover. (Original work published 1885)

Erdelyi, M.H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology*, *6*, 159–171.

Estes, W.K. (1960). Learning theory and the new "mental chemistry." *Psychological Review*, *67*, 207–223.

Fisher, R.P., & Craik, F.I.M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory*, *3*, 701–711.

Fitch, M.L., Drucker, A.J., & Norton, J.A. (1951). Frequent testing as a motivating factor in large lecture courses. *Journal of Educational Psychology*, *42*, 1–20.

Forlano, G. (1936). *School learning with various methods of practice and rewards* (Teachers College Contributions to Education No. 688). New York: Teachers College, Columbia University, Bureau of Publications.

Gardiner, J.M., Craik, F.I.M., & Bleasdale, F.A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition*, *1*, 213–216.

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

Glanzer, M., & Cunitz, A.R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, *5*, 351–360.

Glenberg, A.M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, *15*, 1–16.

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

Grigorenko, E.L., & Sternberg, R.J. (1998). Dynamic testing. *Psychological Bulletin*, *124*, 75–111.

Hamaker, C. (1986). The effects of adjunct questions on prose learning. *Review of Educational Research*, *56*, 212–242.

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

Hilgard, E.R. (1951). Methods and procedures in the study of learning. In S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 517–567). New York: Wiley.

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

Hunt, R.R., & McDaniel, M.A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, *32*, 421–445.

Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, *18*, 879–919.

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology*, *75*, 194–209.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology*, *83*, 340–344.

Izawa, C. (1971). The test trial potentiating model. *Journal of Mathematical Psychology*, *8*, 200–224.

Izawa, C., Maxwell, S., Hayden, R.G., Matrana, M., & Izawa-Hayden, A.J.E.K. (2005). Optimal foreign language learning and retention: Theoretical and applied investigations on the effects of presentation repetition programs. In C. Izawa & N. Ohta (Eds.), *Human learning and memory: Advances in theory and application: The 4th Tsukuba International Conference on Memory* (pp. 107–134). Mahwah, NJ: Erlbaum.

Jacoby, L.L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649–667.

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

Jacoby, L.L., Bjork, R.A., & Kelley, C.M. (1994). Illusions of comprehension, competence, and remembering. In D. Druckman & R.A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance* (pp. 57–80). Washington, DC: National Academy Press.

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

Jacoby, L.L., Shimizu, Y., Daniels, K.A., & Rhodes, M.G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*, 852–857.

Jacoby, L.L., Shimizu, Y., Velanova, K., & Rhodes, M.G. (2005). Age differences in depth of retrieval: Memory for foils. *Journal of Memory and Language*, *52*, 493–504.

James, W. (1890). *The principles of psychology*. New York: Holt.

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

Kang, S.H.K., McDermott, K.B., & Roediger, H.L., III. (in press). Test format and corrective feedback modulate the effect of testing on memory retention. *European Journal of Cognitive Psychology*.

Karpicke, J.D., & Roediger, H.L., III. (2006a). *Expanding retrieval promotes short-term retention, but equal interval retrieval enhances long-term retention*. Unpublished manuscript, Washington University in St. Louis, St. Louis, MO.

Karpicke, J.D., & Roediger, H.L., III. (2006b). *Repeated retrieval during learning is the key to enhancing later retention*. Unpublished manuscript, Washington University in St. Louis, St. Louis, MO.

Kolers, P.A., & Roediger, H.L., III. (1984). Procedures of mind. *Journal of Verbal Learning and Verbal Behavior*, *23*, 425–449.

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. (in press). Illusions of competence during study can be remedied by manipulations that enhance learners' sensitivity to retrieval conditions at test. *Memory & Cognition*.

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

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

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

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

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership*, *63*, 18–24.

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

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

Loftus, G. (1985). Evaluating forgetting curves. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 397–406.

Logan, J.M., & Balota, D.A. (in press). Expanded vs. equal interval spaced retrieval practice: Exploration of schedule of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*.

Madigan, S.A., & McCabe, L. (1971). Perfect recall and total forgetting: A problem for models of short-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*, 101–106.

Maloney, E.L., & Ruch, G.M. (1929). The use of objective tests in teaching as illustrated by grammar. *School Review*, *37*, 62–66.

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

Marsh, E.M., Fazio, L., & Roediger, H.L., III. (2006). *The negative suggestion effect in multiple choice tests*. Unpublished manuscript, Duke University, Durham, NC.

McDaniel, M.A. (in press). Transfer. In H.L. Roediger, III, Y. Dudai, & S.M. Fitzpatrick (Eds.), *The science of learning and memory: Concepts*. Oxford, England: Oxford University Press.

McDaniel, M.A., Anderson, J.L., Derbish, M.H., & Morrisette, N. (in press). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*.

McDaniel, M.A., & Einstein, G.O. (1989). Material appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review*, *1*, 113–145.

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

McDaniel, M.A., Friedman, A., & Bourne, L.E. (1978). Remembering the levels of information in words. *Memory & Cognition*, *6*, 156–164.

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

McDaniel, M.A., & Masson, M.E.J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 371–385.

McDermott, K.B. (2006). Paradoxical effects of testing: Repeated retrieval attempts enhance the likelihood of later accurate and false recall. *Memory & Cognition*, *34*, 261–267.

McGeoch, J.A. (1942). *The psychology of human learning*. New York: Longmans, Green and Co.

McTighe, J., & O'Connor, K. (2005). Seven practices for effective learning. *Educational Leadership*, *63*, 10–17.

Meade, M.L., & Roediger, H.L., III. (2006). The effect of forced recall on illusory recollection in younger and older adults. *American Journal of Psychology*, *119*, 433–462.

Melton, A.W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior*, *9*, 596–606.

Melton, A.W., & Irwin, J.M. (1940). The influence of degree of interpolated learning on retroactive inhibition and the overt transfer of specific responses. *American Journal of Psychology*, *53*, 173–203.

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory*, *2*, 609–622.

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

Neely, J.H., Schmidt, S.R., & Roediger, H.L., III. (1983). Inhibitory priming effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 196–211.

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

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

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1051–1057.

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

Rees, P.J. (1986). Do medical students learn from multiple-choice examinations? *Medical Education*, *20*, 123–125.

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

Richardson, J.T.E. (1985). The effects of retention tests upon human learning and memory: An historical review and an experimental analysis. *Educational Psychology*, *5*, 85–114.

Rickards, J.P. (1979). Adjunct postquestions in text: A critical review of methods and processes. *Review of Educational Research*, *49*, 181–196.

Rock, I. (1957). The role of repetition in associative learning. *American Journal of Psychology*, *70*, 186–193.

Roediger, H.L., III. (1978). Recall as a self-limiting process. *Memory & Cognition*, *6*, 54–63.

Roediger, H.L., III. (1990). Implicit memory: Retention without remembering. *American Psychologist*, *45*, 1043–1056.

Roediger, H.L., III, & Challis, B.H. (1989). Hypermnesia: Increased recall with repeated tests. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Floweree Symposium on Cognition* (pp. 175–199). Hillsdale, NJ: Erlbaum.

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

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

Roediger, H.L., III, & Schmidt, S.R. (1980). Output interference in the recall of categorized and paired associate lists. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 91–105.

Roediger, H.L., III, & Thorpe, L.A. (1978). The role of recall time in producing hypermnesia. *Memory & Cognition*, *6*, 296–305.

Roediger, H.L., III, Weldon, M.S., & Challis, B.H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H.L. Roediger, III, & F.I.M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 3–41). Hillsdale, NJ: Erlbaum.

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

Rosner, S.R. (1970). The effects of presentation and recall trials on organization in multitrial free recall. *Journal of Verbal Learning and Verbal Behavior*, *9*, 69–74.

Rothkopf, E.Z. (1966). Learning from written instructive materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal*, *3*, 241–249.

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

Runquist, W.N. (1986). Changes in the rate of forgetting produced by recall tests. *Canadian Journal of Psychology*, *40*, 282–289.

Schacter, D.L., Rich, S.A., & Stampp, M.S. (1985). Remediation of memory disorders: Experimental evaluation of the spaced retrieval technique. *Journal of Clinical and Experimental Neuropsychology*, *7*, 79–96.

Schmidt, R.A., & Bjork, R.A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Slamecka, N.J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology*, *76*, 504–513.

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

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

Smith, A.D. (1971). Output interference and organized recall from long-term memory. *Journal of Verbal Learning and Verbal Behavior*, *10*, 400–408.

Sones, A.M., & Stroud, J.B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology*, *31*, 665–676.

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

Steele, C.M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*, 613–629.

Sternberg, R.J., & Grigorenko, E.L. (2001). All testing is dynamic testing. *Issues in Education*, *7*, 137–170.

Sternberg, R.J., & Grigorenko, E.L. (2002). *Dynamic testing: The nature and measurement of learning potential*. Cambridge, England: Cambridge University Press.

Sternberg, R.J., Grigorenko, E.L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., Jukes, M., & Bundy, D.A. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence*, *30*, 141–162.

Thomas, A.K., & McDaniel, M.A. (in press). The negative cascade of incongruent generative study-test processing in memory and metacomprehension. *Memory & Cognition*.

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

Thorndike, E.L. (1914). Repetition versus recall in memorizing vocabularies. *Journal of Educational Psychology*, *5*, 596–597.

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 Psychology*, *86*, 357–362.

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review*, *69*, 344–354.

Tulving, E. (1964). Intratrial and intertrial retention: Notes towards a theory of free recall verbal learning. *Psychological Review*, *71*, 219–237.

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

Tulving, E., & Arbuckle, T.Y. (1963). Sources of intratrial interference in immediate recall of paired associates. *Journal of Verbal Learning and Verbal Behavior*, *1*, 321–334.

Tulving, E., & Arbuckle, T.Y. (1966). Input and output interference in short-term associative memory. *Journal of Experimental Psychology*, *72*, 145–150.

Tulving, E., & Colotla, V.A. (1970). Free recall of trilingual lists. *Cognitive Psychology*, *1*, 86–98.

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

Tulving, E., & Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*, 352–373.

Watkins, M.J. (1974). The concept and measurement of primary memory. *Psychological Bulletin*, *81*, 695–711.

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

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

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

Whitten, W.B., & Bjork, R.A. (1977). Learning from tests: Effects of spacing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 465–478.

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

Wixted, J.T., & Rohrer, D. (1994). Analyzing the dynamics of free recall: An integrative review of the empirical literature. *Psychonomic Bulletin & Review*, *1*, 89–106.