CHAPTER ONE

Ten Benefits of Testing and Their Applications to Educational Practice

Henry L. Roediger III, Adam L. Putnam *and* Megan A. Smith

Contents

Abstract

Testing in school is usually done for purposes of assessment, to assign students grades (from tests in classrooms) or rank them in terms of abilities (in standardized tests). Yet tests can serve other purposes in educational settings that greatly improve performance; this chapter reviews 10 other benefits of testing. Retrieval practice occurring during tests can greatly enhance retention of the retrieved information (relative to no testing or even to restudying). Furthermore, besides its durability, such repeated retrieval produces knowledge that can be retrieved flexibly and transferred to other situations. On open-ended assessments (such as essay tests), retrieval practice required by tests can help students organize information and form a coherent knowledge base. Retrieval of some information on a test can also lead to easier retrieval of related information, at least on

Psychology of Learning and Motivation, Volume 55

ISSN 0079-7421, DOI 10.1016/B978-0-12-387691-1.00001-6

delayed tests. Besides these direct effects of testing, there are also indirect effects that are quite positive. If students are quizzed frequently, they tend to study more and with more regularity. Quizzes also permit students to discover gaps in their knowledge and focus study efforts on difficult material; furthermore, when students study after taking a test, they learn more from the study episode than if they had not taken the test. Quizzing also enables better metacognitive monitoring for both students and teachers because it provides feedback as to how well learning is progressing. Greater learning would occur in educational settings if students used self-testing as a study strategy and were quizzed more frequently in class.

1. Introduction

Benefits of testing? Surely, to most educators, this statement represents an oxymoron. Testing in schools is usually thought to serve only the purpose of evaluating students and assigning them grades. Those are important reasons for tests, but not what we have in mind. Most teachers view tests (and other forms of assessment, such as homework, essays, and papers) as necessary evils. Yes, students study and learn more when given assignments and tests, but they are an ordeal for both the student (who must complete them) and the teacher (who must construct and grade them). Quizzes and tests are given frequently in elementary schools, often at the rate of several or more a week, but testing decreases in frequency the higher a student rises in the educational system. By the time students are in college, they may be given only a midterm exam and a final exam in many introductory level courses. Of course, standardized tests are also given to students to assess their relative performance compared to other students in their country and assign them a percentile ranking. However, for purposes of this chapter, we focus on the testing that occurs in the classroom as part of the course or self-testing that students may use themselves as a study strategy (although surveys show that this practice is not widespread).

Why might testing improve performance? One key benefit is the active retrieval that occurs during tests. William James (1890, p. 646) wrote:

> 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.

James presented no evidence for this statement, apparently basing it on introspection. However, experimental reports appearing in the next 20 years showed he was right (Abbott, 1909; Gates, 1917). The act of retrieving when taking a test makes the tested material more memorable, either relative to no activity or compared to restudying the material. The size of the testing effect, as it has been named, also increases with the number of tests given.

Throughout the twentieth century, examination of the testing effect occurred in fits and starts. Gates (1917) provided the first thorough examination, but other important studies were done by Jones (1923/1924), Spitzer (1939), Tulving (1967), and Izawa (1970). In 1989, Glover bemoaned the fact that the testing effect had not been applied to education and the subtitle of his paper on the testing phenomenon was "not gone, but nearly forgotten." Since this rather gloomy appraisal, interest in testing and retrieval practice has made a great comeback. Carrier and Pashler (1992) developed a particular paired-associate learning paradigm that has been used extensively since then, and their study may serve as a landmark for a resurgence of interest in testing over the past 20 years.

Roediger and Karpicke (2006b) provided a thorough review of the early testing work as well as research conducted since that time. But even in the half-dozen years since that review was published, research on retrieval practice and testing has grown rapidly. Many papers cited in this chapter answer important questions that came after 2006, as will become obvious over the course of the chapter.

1.1. Direct and indirect effects of testing

One critical distinction is between the direct effects tests have on retention and the indirect effects provided by tests (Roediger & Karpicke, 2006b). We will refer to this distinction throughout the chapter. Briefly, as the name implies, direct effects arise from the test itself. So, for example, if a student is asked "Which kings fought in the Battle of Hastings in 1066?" and she correctly answered the question, her retrieval of this fact would lead to it being better recollected again later than if she had no practice or had simply studied the answer. This is an example of the direct effect of testing (e.g., Carrier & Pashler, 1992). Incidentally, in case you need it, the answer is that the forces of Duke William II of Normandy overwhelmed King Harold II's English forces at Hastings, hence "the Norman conquest."

The indirect effects of testing refer to other possible effects that testing might have. For example, if students are quizzed every week, they would probably study more (and more regularly) during a semester than if they were tested only on a midterm and a final exam. Thus, testing would have

an indirect effect on apportionment of study activities. We return to evidence bearing on this issue later (Section 11).

The above two examples are clear, but in some cases tests may have both direct and indirect benefits. We will revisit this issue from time to time throughout the chapter. We now consider the 10 benefits of testing (see Table 1), but we have a section at the end outlining possible detriments to testing, too.

2. Benefit 1: The Testing Effect: Retrieval Aids Later Retention

In this section, we review several experiments demonstrating the basic testing effect, the fact that information retrieved from memory leads to better performance on a later test. There are perhaps a hundred experiments we could choose from, but we have selected two straightforward ones from our own lab to make the case. The first experiment used easily nameable pictures as materials (the kind of material that experimental psychologists like to use) whereas the second experiment used nonfiction prose materials more relevant to education. However, the basic testing effect has been obtained with many other types of materials, such as foreign language vocabulary, map reading, general knowledge questions, and so on.

Wheeler and Roediger (1992) conducted an experiment in which a strong testing effect occurred, although the experiment was mostly about a different topic. We present selected conditions here from their experiment to make our points about testing. Their subjects saw 60 pictures

Table 1 Ten Benefits of Testing

Benefit 1	The testing effect: retrieval aids later retention
Benefit 2	Testing identifies gaps in knowledge
Benefit 3	Testing causes students to learn more from the next learning episode
Benefit 4	Testing produces better organization of knowledge
Benefit 5	Testing improves transfer of knowledge to new contexts
Benefit 6	Testing can facilitate retrieval of information that was not tested
Benefit 7	Testing improves metacognitive monitoring
Benefit 8	Testing prevents interference from prior material when learning new material
Benefit 9	Testing provides feedback to instructors
Benefit 10	Frequent testing encourages students to study

while they listened to a story, with instructions that they would later be asked to recall the names of the pictures. The pictures were integrated into the story so that when an object was named in the story, the picture appeared on the screen. Subjects were told that paying attention to the story would help them retain the pictures (which was true). After hearing the story and seeing the pictures, subjects were given free recall tests in which they were given a blank sheet of paper and had to recall as many of the names of the 60 pictures as possible.

After hearing the story, one group of subjects was told that they could leave and return a week later for a test. A second group was given a single test that lasted 7 min and then they were excused. The third group was given three successive 7-min tests after the learning phase; that is, they recalled the pictures once, were given a new blank sheet and recalled as many items as possible a second time, and then repeated the process a third time. The group that recalled pictures once recalled about 32 pictures and the group that recalled them three times recalled 32, 35, and 36 pictures (i.e., performance increased across tests, a phenomenon called hypermnesia; Erdelyi & Becker, 1974).

For present purposes, the data of most interest are those on the final retention test 1 week later when the students returned to the lab for more testing. Students in all three groups had heard the story and seen the pictures once, so the only difference among the three groups was how many tests they had taken just after studying the materials (0, 1, or 3). How did this manipulation affect recall? The data to answer this question are shown in Figure 1, where it can be seen that those who had not been tested recalled 17.4 pictures, those who had been tested once recalled 23.3 pictures, and those who had previously been tested three times recalled 31.8 pictures. Thus, taking three tests improved recall by nearly 80% a week later relative to the condition with no tests.

Another way to consider the data is by comparing the scores on the immediate test just after study to those a week later. Recall that on the first test after study, subjects produced about 32 items. We can assume that those subjects who were not tested immediately after study could have recalled 32 had they been tested, yet a week later they could recall only 17, showing 45% forgetting. However, the group that was tested three times immediately were still able to recall 32 items a week after study, thus giving three tests essentially eliminated forgetting after a week. This outcome shows the power of testing.

Yet a critic might complain that the Wheeler and Roediger (1992) results could be due to an artifact. Perhaps, the critic would maintain, the outcome in Figure 1 has nothing to do with testing per se. Rather, all "testing" did was to permit selective restudy of information. The group that did not take a test did not restudy any material, whereas the group that took the single test restudied 32 of the 60 pictures, and the group with

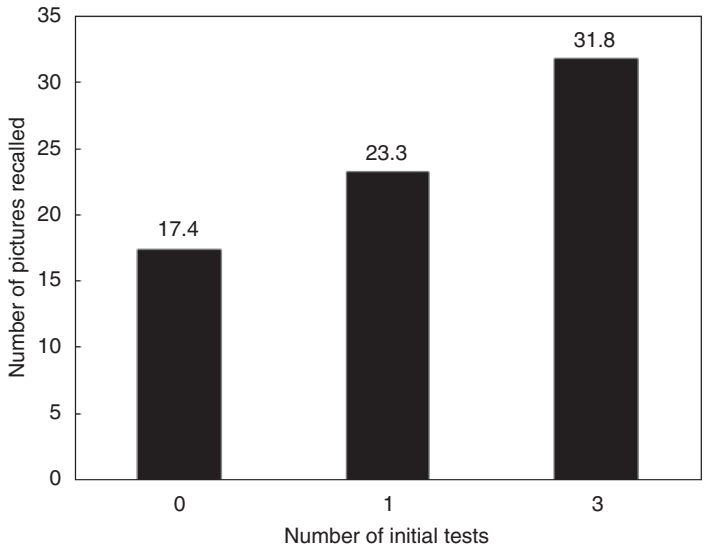

Figure 1 **The number of pictures recalled on a final recall test after a 1-week delay, adapted from Table 1 of Wheeler and Roediger (1992). The number of initial tests strongly influenced final test recall. On the first immediate recall test, subjects recalled, on average, 32.25 pictures. The results indicate that taking three immediate recall tests will effectively eliminate forgetting over a 1-week period.**

three tests restudied 32, then 35, and finally 36 pictures (mostly studying the same items each time). Perhaps it was merely this process of restudying that led to good performance a week later. After all, it is hardly a surprise to find that the more often a person studies material, the better they remember it. Thompson, Wenger, and Bartling (1978) voiced this interpretation of testing research. In a similar vein, Slamecka and Katsaiti (1988) argued that repeated testing may create overlearning on a certain subset of items and that such overlearning is somehow responsible for the effect.

These criticisms of the testing effect are often voiced, but dozens of studies have laid them to rest by including a "restudy" control group in addition to a testing group. That is, in the comparison condition, students restudy the set of material for the same amount of time that others are engaged in taking a test. When this procedure is followed, the testing group is at a disadvantage in terms of restudy of information compared to the restudy group. The reason is that in the testing condition subjects only have the opportunity to restudy the amount of information they can recall (about 53%—$32 \div 60 \times 100$—in the Wheeler and Roediger study), whereas in the restudy condition subjects usually receive the entire set of material again (100%). Thus, if the testing effect were due to restudying,

using such a restudy control should make the testing effects disappear or even reverse. However, this does not happen, at least on delayed tests.

Consider an experiment by Roediger and Karpicke (2006a). They used relatively complex prose passages on such topics as "sea otters" that were full of facts. The test given was free recall; subjects were asked to recall as much as they could from the passage when given its name and the protocols were scored in terms of the number of idea units recalled from the passage. In one condition, subjects studied the passage once and were tested on it three times; on each test, they recalled about 70% of the material. Another group studied the passage three times and was tested once (recalling 77%). Finally, a third group studied the passage four times, so subjects had the greatest study exposure to the material (reading the passage four times) in this condition. Thus, subjects in the three conditions were exposed in one form or another to the material four times via various numbers of studies and test events. We can label the conditions STTT, SSST, and SSSS, where S stands for study of the passage and T stands for its testing.

The data of critical interest were those that occurred on a final criterion test, which was given 5 min or 1 week after the learning session. As can be seen in the left-hand side of Figure 2, when the final test was given shortly after the initial four study/test periods, recall was correlated with the number of study episodes: the SSSS condition led to better performance than the SSST condition that in turn was better than STTT condition. As students have known for generations, cramming does work if a test occurs immediately after studying. However, for subjects given the final test a week later, exactly the opposite ordering of performance emerged: the more students had been tested during the learning session, the better was performance. This outcome occurred despite the fact that subjects who had repeatedly studied the material had received much more exposure to it. Once again, receiving tests greatly slowed down forgetting (see also Karpicke, 2009; Karpicke & Roediger, 2008; Wheeler, Ewers, & Buonanno, 2003). Another point to take from Figure 2 is that a testing effect is more likely to emerge at longer delays after study. On a test given soon after studying, repeated studying can lead to performance greater than that with testing.

We could add dozen more experiments to this section on the basic testing effect (e.g., Carpenter & DeLosh, 2005, 2006; Cull, 2000; Pyc & Rawson, 2007), but we will desist. Many experiments will be reviewed later that have the same kind of design and establish conditions in which testing memory produces a mnemonic boost relative to a restudy control condition (as in Roediger & Karpicke, 2006a) or relative to a condition with no further exposure (as in Wheeler & Roediger, 1992). However, even in the latter case, we can rest assured that the testing effect is mostly due to causes other than restudying the material.

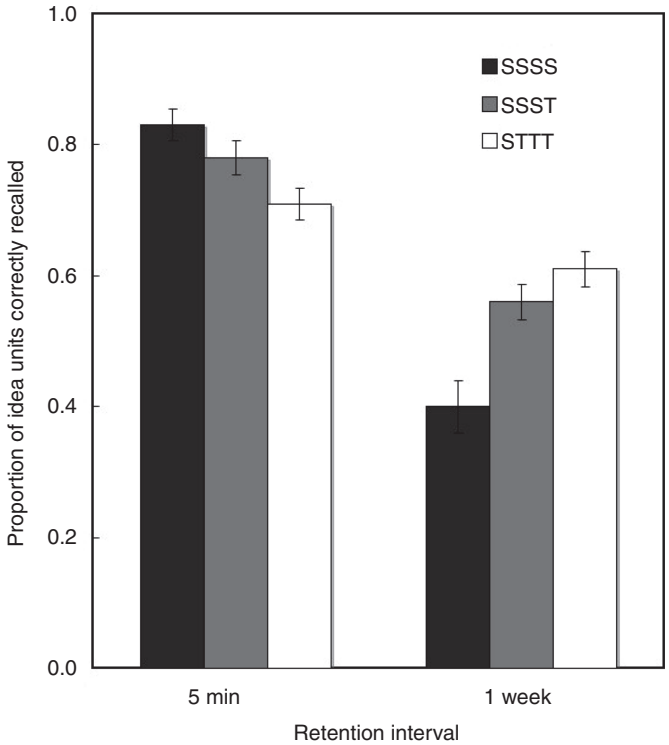

Figure 2 **Mean number of idea units recalled on the final test taken 5 min or 1 week after the initial learning session. During learning, subjects studied prose passages and then completed a varying number of study (S) and test (T) periods. Error bars represent standard errors of the mean (estimated from Figure 2 of Roediger and Karpicke (2006a)).**
Adapted from Experiment 2 of Roediger and Karpicke (2006a).

3. Benefit 2: Testing Identifies Gaps in Knowledge

The testing effect represents a direct benefit of testing; the second benefit is indirect. Taking a test permits students to assess what they know and what they do not know, so that they can concentrate study efforts on areas in which their knowledge is deficient. Students may take a practice quiz, realize which questions or items they got wrong, and then spend more time studying the items they missed. For example, Amlund, Kardash, and Kulhavy (1986) found that subjects corrected errors on a second test if they had an intervening study session after the first test. Other research shows that when students receive opportunities to restudy material after a test, they spend longer on restudying items that were

missed than those that were correctly retrieved (see Son & Kornell, 2008).

Kornell and Bjork (2007) provided evidence from a laboratory experiment that students are typically unaware that learning can occur during testing. In one experiment, students learned a set of Indonesian–English vocabulary words by repeated trials. They had the option of studying the pairs or being tested on them (with feedback) on each occasion and could switch between the two modes at any point. Most students began in the study mode, although nearly everyone changed to the test mode after the first two trials. Kornell and Bjork interpreted this outcome as indicating that students wanted to achieve a basic level of knowledge before testing themselves. In addition, Kornell and Bjork also reported the results of a survey in which students were asked whether they quizzed themselves while studying (using a quiz at the end of a chapter, a practice quiz, flashcards, or something else); 68% of respondents replied that they quizzed themselves "to figure out how well I have learned the information I'm studying" (Kornell & Bjork, 2007, p. 222). Only 18% of respondents recognized that testing actually facilitated further learning.

In another survey on study habits, Karpicke, Butler, and Roediger (2009) asked college students to list their most commonly used study habits (rather than asking directly if they used testing, as in the Kornell and Bjork (2007) survey). When the question was framed in this open-ended manner, only 11% of students listed retrieval practice as a study technique they used, suggesting that students may be generally unaware of the direct or indirect benefits of testing. On a forced response question, students had to choose between studying and testing in a hypothetical situation of preparing for a test. Only 18% of students chose to self-test and more than half of those explained that they chose to self-test to identify what they did or did not know to guide further study. Thus these two points are in broad agreement with the Kornell and Bjork (2007) findings.

In further surveys, McCabe (2011) found that college students' knowledge of effective study strategies is quite poor without specific instruction. She provided students with educational scenarios and asked them to select study strategies that would be effective. She based her strategies on findings from cognitive psychology studies, including such principles as dual coding and retrieval practice. McCabe found that students were generally unaware of the effectiveness of the strategies. If this is the case with college students, one can only assume that high school students and others in lower grades would, at best, show the same outcome.

Testing one's memory allows one to evaluate whether the information is really learned and accessible. Karpicke et al. (2009) suggested that one of the reasons students reread materials rather than testing themselves is that rereading leads to increased feelings of fluency of the material—it

seems so familiar as they reread it they assume they must know it. Also, in contrast to self-testing, restudying is easy. In short, students may lack metacognitive awareness of the direct benefits of testing, while at the same time understand that self-testing can be useful as a guide to future studying. Testing helps students learn because it helps them understand what facts they might not know, so they can allocate future study time accordingly.

4. Benefit 3: Testing Causes Students to Learn More from the Next Study Episode

Another benefit of retrieval practice is it can enhance learning during future study sessions. That is, when students take a test and then restudy material, they learn more from the presentation than they would if they restudied without taking a test. This outcome is called test-potentiated learning (Izawa, 1966). The benefits of test potentiation are distinctly different from the direct benefits of testing per se, although in many practical situations (e.g., receiving feedback after tests) the two are mixed together.

Izawa (1966) was perhaps the first researcher to study the test potentiation effect and has contributed much to our understanding of test potentiation. Her initial forays into the area emerged after asking questions about whether learning could occur during a test. She proposed three specific hypotheses. First, neither learning nor forgetting occurred on tests. Second, learning and forgetting (as well as learning of incorrect information) could occur on test trials. Finally, although learning and forgetting might not occur on a test session, taking a test might influence the amount of learning during a future study session. Izawa studied how different patterns of study, test, and neutral trials affected later performance.

Across many experiments (e.g., Izawa, 1966, 1968, 1970), Izawa concluded that neither forgetting nor learning occurred on test trials, but taking a test could improve the amount of material learned on a subsequent study session. While this conclusion may appear to contradict the basic finding of the testing effect, the contradiction is resolved by examining how learning and forgetting are defined in Izawa's basic paradigm. Izawa's conclusion was that no learning or forgetting occurred during a test trial, but she made no assumptions about how learning or forgetting would be affected *after* the test trial; the testing effect can be interpreted as a slowing of forgetting after the test.

Other researchers have continued to explore test potentiation in different contexts. Pyc and Rawson (2010) showed that subjects formed

more effective mediators (mnemonic devices that link a cue to a target) when they were tested before a study session compared to when they were not. Karpicke and Roediger (2007) found that subjects learned more from a single study session after being tested three times relative to completing one test prior to study. Similarly, Karpicke (2009) showed a test potentiation effect by comparing three different patterns of study and test on how students learned foreign language vocabulary. One condition was the standard cycle alternating between study and test trials; during a study trial, subjects saw both a Swahili word and its English translation, and on a test trial, they saw the Swahili word and were asked to recall the English word, without any corrective feedback. The standard cycle consisted of three alternative study–test trials, or STSTST. Another group studied three times before the first test and had one intervening study session before the final test (SSSTST). Finally, a third group had five study sessions before the final test period (SSSSST).

Figure 3 shows the results of the experiment. Clearly, alternating study and test trials caused subjects to recall more word pairs on the final test than for others who spent equivalent time studying. This outcome can be

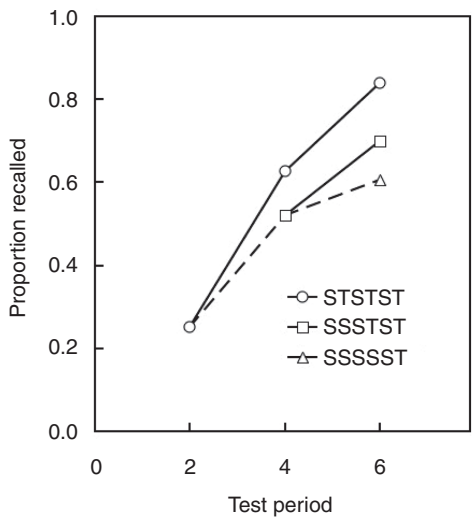

Figure 3 The potentiating effects of testing on learning. Subjects alternated study and test (STSTST), studied with only one intervening test (SSSTST), or studied with no intervening tests (SSSSST). The dashed line connects performance on the first test across conditions to show the effect of repeated studying on recall. The solid lines connect performance within each condition. The results show that inserting test trials leads to greater learning by the final test.
Adapted from Experiment 1 of Karpicke (2009).

interpreted as the test potentiating later learning, because tests enabled learning from the later study episode.

Other researchers, however, have had difficulty obtaining test potentiation effects when they are examined in more complex designs that discount the fundamental testing effect (McDermott & Arnold, 2010). For example, all the experiments described in the previous paragraph could be interpreted as exemplifying direct effects of testing because the two effects are mixed together in those designs (e.g., the design of Karpicke and Roediger (2007) and others described above). Thus, the major difficulty in examining test potentiation is separating its effects (enhanced learning from restudying) from other factors related to testing (such as the direct effect of testing on improving recall). However, McDermott and Arnold (2010) have succeeded in replicating Izawa's work showing test-potentiated learning under certain conditions, so both the direct effect and the indirect effect of test-potentiated learning are secure findings.

In many standard studies on testing, feedback is provided after the test and this condition is compared to a condition in which no test is given (but students study the material). The test plus feedback condition usually greatly outpaces the restudy-only condition, even when timing parameters are equated (i.e., subjects are exposed to material for the same amount of time). The benefit of testing probably arises both from the direct effect of testing and from the indirect effect of testing potentiating future learning (from feedback), but further research is needed to establish this point and determine the relative contributions of the testing effect and the test potentiation effect in these circumstances.

5. Benefit 4: Testing Produces Better Organization of Knowledge

Another indirect benefit of retrieval practice is that it can improve the conceptual organization of practiced materials, especially on tests that are relatively open-ended (such as free recall in the lab or essay tests in the classroom). Gates (1917) postulated that one of the reasons retrieval practice leads to increased performance is that retrieval (or recitation, as he called it) causes students to organize information more than does reading. He suggested that as students actively recall material, they are more likely to notice important details and weave them into a cohesive structure.

Masson and McDaniel (1981) showed that an additional testing session after study resulted in higher performance on delayed recall and recognition tests and, more important, that the additional test yielded higher organization on the final recall test. Their primary measure of

organization was the adjusted ratio of clustering (ARC), which is a measure of how often words from the same category are recalled together in free recall with an adjustment for the overall level of recall. Scores range from −1 to 1, with 1 representing perfect organization or clustering and 0 representing chance clustering (Roenker, Thompson, & Brown, 1971). Masson and McDaniel's results suggested that the test resulted in improved organization and higher recall on final tests.

More recently, other research (Zaromb, 2010; Congleton & Rajaram, 2010) has explored the relationship between testing and organization. Experiments reported by Zaromb and Roediger (2010), for example, showed that retrieval practice during testing improves both the organization of materials and their recall. In fact, the increased organization from previous retrievals may provide an underlying mechanism of the testing effect, at least in free recall.

In one experiment (Zaromb & Roediger, 2010, Experiment 2), subjects studied categorized word lists in one of several learning conditions (although we are considering only two groups here). One group studied the list of words twice with different encoding instructions; in the first cycle, subjects made pleasantness ratings and in the second cycle, they were given intentional learning instructions. A second group of subjects learned a list of items by making pleasantness ratings, and then they immediately attempted a final free recall of the list (with no feedback). Both groups returned to the lab after a 24-h delay and took both a free recall test and a cued recall test. Table 2 shows the results. In the free recall test, subjects who had taken an intermediate test showed increased performance as measured by total number of words recalled.

The same outcome occurred when total words were decomposed into the number of categories recalled (Rc; subjects are given credit for recalling a category if one item is recalled from that category) and the number of words recalled per category (Rw/c). Most important, the tested group showed greater ARC score compared to the group that studied twice. A similar pattern of results in recall was obtained for the cued recall test where subjects where provided with the category labels as retrieval cues. Zaromb and Roediger also showed that testing improves subjective organization, or recall of items in a more consistent order (Tulving, 1962).

In sum, testing can increase both category clustering and subjective organization of materials compared to restudying, and this may be one of the underlying mechanisms driving the testing effect, at least in free recall and other open-ended kinds of tests (e.g., essay tests). Further research is needed to generalize this result to educational contexts, but extrapolating from the current work, the prediction would be that testing improves organization of knowledge.

Table 2 Mean Proportion of Words Recalled, Number of Categories Recalled (Rc), Number of Words Per Category Recalled (Rw/c), ARC Scores on Delayed Free, and Cued Recall Tests

Measure		Free Recall		Cued Recall	
		S_pS_i	S_pT	S_pS_i	S_pT
Recall	*Prop.*	.21	.45	.37	.61
	CI	(.06)	(.06)	(.06)	(.05)
Rc	*M*	8.19	12.56	15.69	17.25
	CI	(1.32)	(.74)	(1.09)	(.67)
Rw/c	*M*	2.16	3.17	2.09	3.17
	CI	(.35)	(.28)	(.26)	(.27)
ARC	*M*	.60	.85		
	CI	(.17)	(.04)		

Note: Values in parentheses are 95% confidence intervals (CI). Subjects made pleasantness ratings on the first trial and had intentional learning instructions on the second trial (S_pS_i) or made pleasantness ratings on the first trial followed by a recall test on the second trial (S_pT). Adapted from Experiment 2 of Zaromb and Roediger (2010).

6. Benefit 5: Testing Improves Transfer of Knowledge to New Contexts

One criticism of retrieval practice or testing research is that students may be learning little factoids in a rote, verbatim way. Critics complain that testing is the old "kill and drill" procedure of education from 100 years ago that produces "inert knowledge" that cannot be transferred to new situations. However, proponents of testing argue that retrieval practice induces readily accessible information that can be flexibly used to solve new problems. This issue leads to the crucial question of whether knowledge acquired via retrieval practice (relative to other techniques) can be applied to new settings.

Recent research shows that the mnemonic benefits of taking a test are not limited to the specific questions or facts that were tested; retrieval practice also improves transfer of knowledge to new contexts. Transfer may be defined as applying knowledge learned in one situation to a new situation. Researchers often categorize transfer as being near or far; near transfer occurs if the new situation is similar to the learning situation, whereas far transfer occurs if the new situation is very different from the learning situation. Barnett and Ceci (2002) proposed a taxonomy for transfer studies, arguing that transfer might be measured on many

continuous dimensions (e.g., knowledge domain, physical context, temporal context, etc.).

The topic of transfer is an old one—Ebbinghaus (1885) conducted transfer experiments—but there has been a large growth in research over the past decade. Furthermore, transfer is extremely important in education; the purpose of education is to teach students information that they will be able to apply later in school, as well as in life after their schooling is finished. However, transfer of knowledge can be difficult to obtain (e.g., Gick & Holyoak, 1980). Far transfer is very difficult to obtain, yet is arguably the most important type of education to apply to settings encountered later in life (Barnett & Ceci, 2002). In fact, Detterman (1993) maintained that experiments investigating transfer are insignificant unless they are able to obtain far transfer on a number of dimensions. Given the important role of transfer in education and the difficulty in promoting its occurrence, the finding that testing can improve transfer is an important one.

Some evidence suggests that repeated testing can facilitate transfer better than restudying. For example, Carpenter, Pashler, and Vul (2006) showed that testing with word–word paired associates (denoted by A–B here) improved performance on a later test relative to additional study opportunities. When given A, subjects could recall B more often when they had previously been tested relative to only studying the pairs. More important, Carpenter et al. also tested subjects' recall for the A member of the pair when they were given B, so they were tested on the member of the pair that was not directly retrieved during initial testing. Recall was improved for these A items when learning had occurred via testing relative to repeated studying. Repeatedly testing with one member of the pair transferred to higher performance in recalling the other member of the pair. This could be considered a case of near transfer.

Similar benefits of testing have been shown with more complex materials, even in learning concepts. Jacoby, Wahlheim, and Coane (2010) showed that testing can improve classification of novel exemplars when students learn categories of birds. Students learned to classify birds by repeatedly studying or repeatedly testing examples of various classes of birds. During a study trial, students were presented with a picture of a bird and the name of the bird family to which it belonged (e.g., warbler presented with a picture of this type of bird). During a test trial, students were presented with only a picture of a bird and asked to name the family to which the bird belonged (like warbler), and then they received feedback (the correct name of the category). Students who were repeatedly tested were better able to classify new birds than those who repeatedly studied them, showing that testing helped subjects better apply their knowledge to new exemplars. In two other generally similar examples of transfer, testing

improved transfer relative to restudying using multimedia materials (Johnson & Mayer, 2009) and with elementary school children learning about maps (Rohrer, Taylor, & Sholar, 2010).

In a series of experiments, Butler (2010) recently demonstrated that repeated testing not only increases retention of facts and concepts learned from prose passages, but also increases transfer of knowledge to new contexts (relative to repeated studying). In Experiments 1 and 2, repeatedly testing with questions in one knowledge domain (e.g., information about bats) promoted retention in answering the same questions as well as new questions within the same knowledge domain. Better performance on new questions provided evidence of near transfer. More impressively, in Experiment 3 Butler showed that repeated testing improved far transfer—that is, transfer to new questions in different knowledge domains (again, relative to repeated restudying). In this experiment, subjects studied prose passages on various topics (e.g., bats; the respiratory system). Subjects then restudied some of the passages three times and took three tests on other passages. After each question during the repeated tests, subjects were presented with the question and the correct answer for feedback. One week later subjects completed the final transfer test. On the final test, subjects were required to transfer what they learned during the initial learning session to new inferential questions in different knowledge domains (e.g., from echolocation in bats to similar processes used in sonar on submarines).

Figure 4 depicts the results from the final transfer test. This experiment showed that repeated testing led to improved transfer to new questions in a new domain relative to restudying the material. Butler (2010) also showed through conditional analyses that retrieving the information during the initial test was important in producing transfer to a new domain. Subjects were more likely to correctly answer a transfer question when they had answered the corresponding question during initial testing. According to Butler, retrieval of information may be a critical mechanism producing greater transfer of that information later.

Practicing retrieval has been shown time and again to produce enhanced memory later for the tested material. One criticism in educational circles has been that testing appears to produce enhanced memory for the facts tested, but that such "kill and drill" procedures may produce "inert" or "encapsulated" learning that will not transfer to new settings. However, the experiments reviewed here show that testing does produce transfer, even far transfer (Butler, 2010). Along with the other evidence reviewed, it appears that retrieval practice produces knowledge that can be flexibly transferred, which overcomes this criticism.

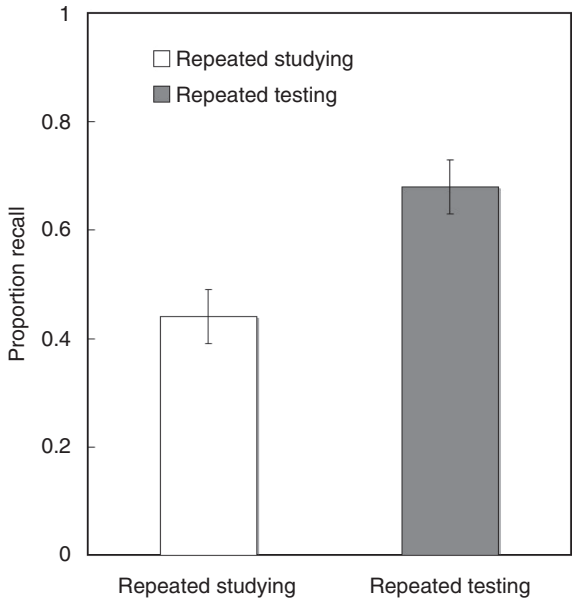

Figure 4 Performance on the final transfer test containing inferential questions from different knowledge domains 1 week after initial learning. Error bars represent the standard error of the mean. During initial learning, subjects repeatedly studied the prose passages or were repeatedly tested on the prose passages.
Adapted from Experiment 3 of Butler (2010).

7. Benefit 6: Testing can Facilitate Retrieval of Material That was not Tested

One potential limiting factor of implementing testing in a classroom setting is choosing which material to test. It is unrealistic for an instructor to test students on everything. Fortunately, research on testing suggests that retrieval practice does not simply enhance retention of the individual items retrieved during the initial test: taking a test can also produce retrieval-induced facilitation—a phenomenon that shows testing also improves retention of nontested but related material.

Chan, McDermott, and Roediger (2006) were the first to coin the term retrieval-induced facilitation, providing evidence for the effect in three experiments. Students studied a prose passage and then completed two initial short answer tests, restudied the passage twice, or did nothing (the control condition). Those in the initial testing group answered questions related to a subset of information from the passage. More important, another subset from the passage was not tested during the initial test, but

this material was related to the questions that had been answered on the initial test. In the restudy condition, students read the answers but did not receive a test. After 24 h, all the students returned to complete a final test covering the entire passage. Results of the final test revealed that retention of the nontested information was superior when students had taken a test relative to conditions in which they restudied the material or in which they had no further exposure after study. Chan et al. concluded that testing not only improves retention for information covered within a test, but also improves retention for nontested information, at least when that information is related to the tested information.

In contrast, other researchers have found that retrieving some information may actually lead to forgetting of other information, a finding termed retrieval-induced forgetting (e.g., Anderson, Bjork, & Bjork, 1994). In a typical retrieval-induced forgetting experiment, subjects first study words in categories and then take an initial test. For some categories, half of the items are repeatedly retrieved during the initial test; for other categories, none of the items are retrieved during the initial test. The general finding is that the unpracticed items from the categories cued for retrieval practice are impaired on a later retention test relative to items from the nontested categories.

Retrieval-induced facilitation and retrieval-induced forgetting are obviously contradictory findings. Consequently, Chan (2009) sought to differentiate between conditions causing facilitation and conditions causing forgetting in these paradigms. In two experiments, he demonstrated the importance of integration of the materials and the delay of the test for the retrieval-induced facilitation and retrieval-induced forgetting effects. In his first experiment, subjects studied two prose passages; each passage was presented one sentence at a time on the computer. During study, some subjects were given the sentences in a coherent order and were told to integrate the information (the high integration condition). For another group of subjects, the sentences within each paragraph were scrambled to disrupt integration of information during study (the low integration condition). Similar to the Chan et al. (2006) experiments, an initial test occurred immediately after studying one of the passages, and subjects completed the same test twice in a row. Subjects completed the final test covering material from both the passages 20 min or 24 h after the completion of the initial learning phase.

Figure 5 depicts performance on the final test. Results reveal both a retrieval-induced facilitation effect (see the fourth pair of bars in Figure 5) and a retrieval-induced forgetting effect (see the first pair of bars in Figure 5) within the same experiment. This outcome demonstrates the importance of both integration of materials and delay of the final test for these effects. When subjects were instructed to integrate the information during study (i.e., the high integration condition) and the test was delayed

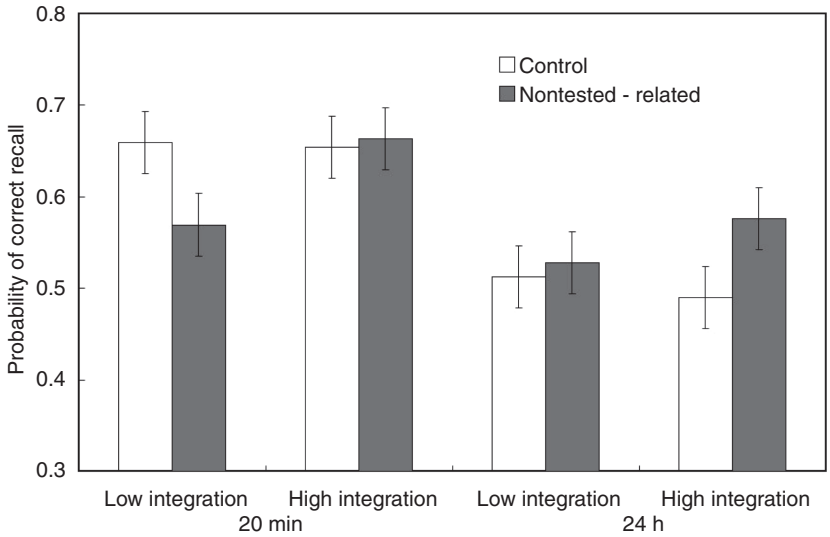

Figure 5 **Performance on the final test for questions drawn from the passage that was not tested initially (control items) and questions drawn from the tested passage but were not present on the initial test (nontested related items). During the initial learning session, subjects studied two passages either in a coherent order with integration instructions or in a randomized order (low integration). Subjects completed an initial test for one of the passages. The final test was completed 20 min or 24 h after the initial learning session. Error bars represent standard errors.**
Adapted from Experiment 1 of Chan (2009).

24 h, a retrieval-induced facilitation effect was found—subjects' performance was enhanced for the nontested items from the passage that was tested relative to the control items. However, when the ability to integrate during study was disrupted (i.e., the low integration condition) and the final test was only 20 min after the initial learning phase, a retrieval-induced inhibition effect was found—subjects' performance was reduced for the nontested items relative to control items. Despite the fact that contradictory results from retrieval-induced facilitation and retrieval-induced forgetting literatures emerge, it seems that these two effects do occur under different sets of conditions. The other two conditions in the experiment of Chan (2009) produced intermediate results.

Evidence from the retrieval-induced facilitation literature provides additional support for the use of testing to enhance learning and memory in educational settings. Notably, it seems that when conditions are more similar to those in educational settings, retrieval-induced facilitation occurs (see Cranney, Ahn, McKinnon, Morris, and Watts (2009) for further evidence of retrieval-induced facilitation in classroom settings). In addition, these effects seem to be durable—Chan (2010) increased the

length of the retention interval, showing that the benefits of retrieval-induced facilitation can last up to 7 days. The experiments reviewed here show that testing can be used in classroom settings to enhance retention of both the tested material and the related but untested material. Retrieval-induced forgetting does not seem to occur on tests delayed a day or more (MacLeod & Macrae, 2001).

8. Benefit 7: Testing Improves Metacognitive Monitoring

Another benefit of testing is improvement of metacognitive accuracy relative to restudying (e.g., Roediger & Karpicke, 2006a; Shaughnessy & Zechmeister, 1992). This point is related to the second one discussed—testing informs students as to what they know and what they do not know. However, in this case, the focus is on students' accurate predictions of their future performance. Testing permits students to have better calibration of their knowledge. If students only study material repeatedly, they may think that their familiarity with the material means that they know it and can retrieve it when needed. However, such familiarity can be misleading. These points have direct implications for educational settings—the better students are at differentiating what they do know and what they do not know well, the better they will be at acquiring new and more difficult material and studying efficiently (Thomas & McDaniel, 2007; Kornell & Son, 2009). Therefore, instead of simply restudying, teachers can administer quizzes and students can self-test to determine what material they know well and what material they do not know well.

Students' ability to accurately predict what they know and do not know is an important skill in education, but unfortunately students often make inaccurate predictions. When students reread material repeatedly, they are often overconfident in how well they know the material. Taking a test, however, can lead to students becoming less confident, a finding known as the underconfidence-with-practice effect (Koriat, Scheffer, & Ma'ayan, 2002; see also Finn & Metcalfe, 2007, 2008). Testing can help compensate for the tendency to be overly confident, which results in a more accurate assessment of learning.

In the first section on the direct effects of testing, we described an experiment by Roediger and Karpicke (2006a), showing that testing produces greater long-term benefits relative to studying. In particular, studying a passage once and taking three tests improved retention a week later relative to studying the passage three times and taking one test or studying the passage four times (see Figure 2, right-hand side). At the end of the first session in the same study, the authors had students judge how well they would do when they were tested in a week (a metacognitive

judgment). After learning the passages in their respective conditions (SSSS, SSST, and STTT), subjects completed a questionnaire about the learning phase. They were asked to predict how well they thought they would remember the passage in 1 week, and predictions were made on a scale ranging from 1 (*not very well*) to 7 (*very well*). Even though testing produced greater long-term benefits relative to repeated studying after 1 week, the subjects in the repeated study condition (SSSS) were more confident that they would remember the content of the passage relative to those in the tested groups (SSST and STTT). Thus, repeatedly studying inflated students' predictions about their performance, causing them to be overconfident (see also Karpicke & Roediger, 2008). Put another way, testing reduced students' confidence even while aiding their performance. Interestingly, however, students' predictions do line up with their performance on a test given a few minutes after the learning session (see the left-hand side of Figure 2, where the SSSS condition was best). Thus, when students try to make a long-term prediction (how will I do a week from now?), they may base their judgments on their current retrieval fluency (what Bjork and Bjork (1992) called retrieval strength). They cannot accurately assess the quality that will lead to success a week later (storage strength, in the Bjorks' terms).

Testing is a powerful way to improve retention, but when students are given control over their own learning, they do not often choose to test themselves or do not test themselves very frequently (Karpicke, 2009; Kornell & Bjork, 2007). During paired-associate learning, when students are given the opportunity to drop, restudy, or retest on items they have correctly retrieved, they often choose to drop items despite benefits that would accrue if they continued to test themselves. When given control early in the learning phase, students often choose to study pairs instead of testing themselves on them and receiving feedback. These decisions seem to be guided by their inflated judgments of learning, but they lead to poor learning strategies (Karpicke, 2009; Metcalfe & Finn, 2008).

Students seem to lack a good theory about what study strategies are effective. As noted in a previous section, surveys have shown that university students do not realize the direct benefits of retrieval practice as a study strategy. Future research is needed to determine if students can be educated on this aspect. For example, if students experience the benefits of retrieval practice on learning in one context, will they then adopt this strategy for learning in a different context? While we must await the answer to this question, we can say that testing does cause students to become less overconfident in the judgments of learning (even to the point of underconfidence, as in the underconfidence-with-practice effect). Because tests generally improve metacognition, educators should encourage their students to self-test during learning and while studying.

9. Benefit 8: Testing Prevents Interference from Prior Material when Learning New Material

Another indirect benefit of testing is that tests create a release from proactive interference. Proactive interference occurs when sets of materials are learned in succession; the previous material learned influences the retention of new materials in a negative manner. Thus, proactive interference refers to the poorer retention of material learned later, caused by prior learning (Underwood, 1957; see Crowder (1976) for a review). Elongated study sessions may therefore cause a buildup of proactive interference. However, research has shown that when tests are inserted between study episodes, they cause a release from proactive interference and enable new learning to be more successful.

Szpunar, McDermott, and Roediger (2007) reported evidence of a release from proactive interference caused by testing in a paradigm in which subjects learned five lists of words. During learning, each list was separated from the next list by an immediate test or a short break of equivalent length. The group that took tests between each list performed better on a final test relative to the group that took short breaks. In addition, the tested group was able to recall a greater proportion of studied words from the most recent list relative to the no-test control group. Thus, taking tests after learning each list protected the subjects from proactive interference during learning.

In a later experiment, Szpunar, McDermott, and Roediger (2008) directly tested the idea that testing protects against the buildup of proactive interference. In two experiments, subjects studied five lists composed of words that were interrelated across lists or words that were unrelated to one another. (The interrelated words belonged to the same categories across lists, for example, several different types of birds or furniture in each list). Between each list, subjects completed math problems for 2 min or completed math problems for 1 min followed by a 1-min free recall test over the list learned most recently. Both groups were tested on the fifth list after its presentation. In addition, a cumulative final test was given to all subjects. For the final test, subjects were instructed to recall as many words from each of the studied lists as possible.

Figure 6 shows the mean number of words recalled from list 5 on the initial test and the final test. The top panel of the figure shows the results from the experiment with interrelated word lists, while the bottom panel shows the results from the experiment with unrelated word lists. For both interrelated and unrelated materials, taking intervening tests during learning protected against proactive interference. Relative to the nontested group, subjects tested after each list produced more correct words from the list 5 and produced fewer intrusions, thus showing that the tests protected

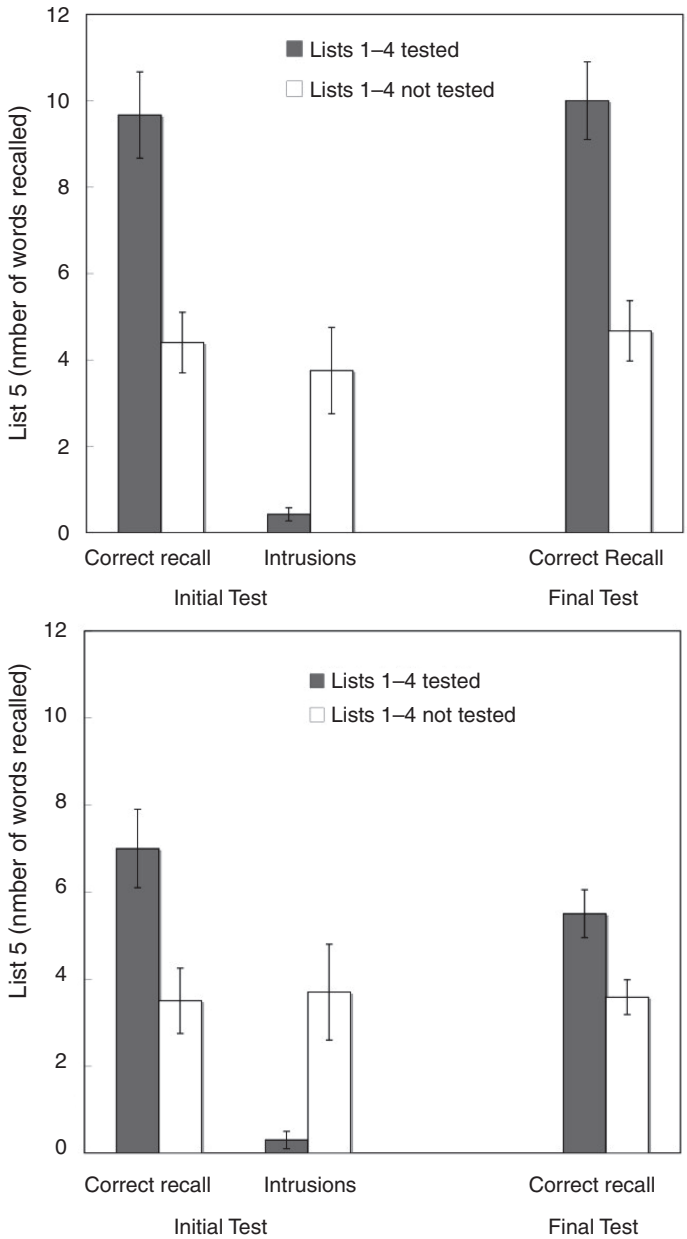

Figure 6 Mean number of words recalled from list 5 on the initial test and the final test when both interrelated lists (top panel) and unrelated words lists were used (bottom panel). Error bars represent standard errors of the mean (estimated from Figures 1 and 2 of Szpunar et al. (2008)). Subjects learned five successive lists of words and between each list some subjects completed a free recall test while other subjects completed a filler task (math problems). All subjects were tested after list 5 and were given a final cumulative free recall test.

subjects from the buildup of proactive interference. In additional experiments, Szpunar et al. (2008) ruled out the hypothesis that the release from proactive interference caused by testing is due to re-exposure to the material because a comparison condition having subjects restudy the lists (rather than receiving tests) did not protect against the buildup of proactive interference.

The results from these and other experiments provide compelling evidence that testing protects subjects from the negative effects of proactive interference, at least when they are required to learn lists of words in succession. While testing causes a release from proactive interference in experimental settings, it is not yet clear whether these results have implications for classroom practice. Bridging experiments using nonfiction prose materials and the like is the next step needed. However, we are optimistic that these results will eventually provide lessons for classroom practice and for self-testing as a study strategy. The next two sections discuss the indirect benefits testing produces within the classroom.

10. Benefit 9: Testing Provides Feedback to Instructors

So far our discussion on the benefits of testing has focused on how testing can have an impact on the learning and memory of students in the classroom. However, classroom testing can do more than help students learn: testing can provide teachers with valuable feedback about what students do and do not know, and teachers in turn can encourage students to change their study behavior. Although these points may seem obvious, they are often overlooked benefits of using frequent testing in the classroom.

Tests and quizzes in the classroom are perhaps one of the most important ways in which teachers can formally assess the knowledge of their students, but of course homework can be used for this purpose, too. Testing is typically seen as an evaluation of what students have learned, and indeed this is true. Conscientious teachers will pay attention to how students perform on tests and use that knowledge to inform their teaching in the future. If many students fail a particular topic on the test, it may be a sign to spend more time covering that material next time or use a different approach to teaching the materials. Teachers can also learn how individual students perform and what the students' respective strengths and weaknesses are. In turn, teachers can use that information to guide further instruction.

Teachers often drastically overestimate what they believe their students to know (Kelly, 1999) and testing provides one way to improve a teacher's estimation of their students' knowledge. The problem of "the curse of knowledge" permeates education. That is, instructors (especially those

just beginning) can fail to realize the state of knowledge of their students and pitch their presentations at too high a level. (Most readers can think of their first calculus or statistics course in this regard.) The general idea is that once we know something and understand it well, it is hard to imagine what it was like not to know it. For example, Newton (1990) conducted a study in which students sat across from each other separated by a screen. Each was given a list of 25 common tunes that most Americans know (Happy Birthday to You, the Star Spangled Banner, etc.). One student (the sender) was picked to tap out the tune with his or her knuckles on the table and give an estimate of the likelihood that the other student could name the tune. The other student (the receiver) tried to decipher the tune and name it. This is a classic situation similar to a teacher and student where one person knows the information (tune, in this case) and is trying to communicate it to the other person who does not know it. When the senders judged how well they did in communicating the tune to the other student, they thought they succeeded about 50% of the time. However, the students on the receiving end of the taps could recognize the tune only 3% of the time! When the sender was tapping out Happy Birthday, she was hearing all that music in her mind's ear and tapping in time to it. What the receiver heard, however, was a series of erratic taps. This tale is an allegory of an expert in a subject matter trying to teach it to a novice, especially the first time. Again, it is hard to know what it is like not to know something you know well.

One hopeful new technology may help overcome the instructor's curse of knowledge. The introduction of student response (clicker) systems that permit teachers to quiz students' understanding during lectures may provide assessment on the fly. Teachers can give 2–3-item quizzes in the middle of a lecture to assess understanding of a difficult point; if many students fail to answer correctly, the instructor can go back and try to present the information in a different way. As smart phones increase in use and become more standardized, they may be adapted in classrooms for the same purpose. These new technologies represent a relatively new approach that provides immediate feedback to both students and instructors about students' understanding.

A more formal approach that utilizes testing to understand the current state of individual students is referred to as formative assessment (Black & Wiliam, 1998a, 1998b; for a brief review of formative assessment from a cognitive psychology perspective, see Roediger and Karpicke (2006b)). Formative assessment not only helps teachers better understand what their students know, but also aims to improve the metacognitive judgments of the students' own knowledge. Students will be better able to assess their current knowledge state and their goal knowledge state, as well as understand what steps they need to take to close that gap if they are given proper

feedback. Black and Wiliam (1998a) reviewed studies of formative assessment, and one of their major conclusions was that implementing formative assessment programs generally improved performance in the classroom. However, they also concluded that formative assessment programs themselves, as implemented, typically need improvement. One important point is that effective formative assessment programs do not simply add more tests and have teachers pay attention to students' scores, but rather implementing good formative assessment practices typically requires an overhaul of classroom pedagogy geared toward maximizing interactions between the teacher and students. In these interactions, students should have ample opportunity to show understanding, and teachers in turn should provide explicit personalized feedback about how students can improve.

11. Benefit 10: Frequent Testing Encourages Students to Study

Probably the most influential indirect benefit of testing is the one described in general terms at the beginning of the chapter: Having frequent quizzes, tests, or assignments motivates students to study. Every professor and every student knows that many students procrastinate and often do not study until the night before a test. Often university courses include only a midterm and a final exam, and it is no surprise that the episodes of studying occur primarily just before tests. Mawhinney, Bostow, Laws, Blumenfeld, and Hopkins (1971) documented this point in controlled circumstances, with tests given daily, weekly, or every three weeks. Studying was most copious and evenly spaced with daily testing. With less frequent testing, study behavior occurred only before the tests (see also Michael, 1991). In addition, in their survey of student behaviors described previously, Kornell and Bjork (2007) found that 59% of students, when choosing what to study, chose topics that were due soon or already overdue. More frequent testing across the semester would encourage students to study more and would space their studying over several weeks.

One specific example of how retrieval practice can provide benefits aside from direct mnemonic benefits can be found in Lyle and Crawford (2011; see also Leeming, 2002). The senior author taught two sections of an introductory statistics course and in one session gave students a short two- to six-question quiz at the end of every lecture. The quizzes covered only materials from the current day's lecture and the emphasis was on the quizzes as being for retrieval practice rather than assessment. As such, the quizzes played only a minor role in determining students' final grades.

This conception of daily quizzes alleviates some of students' typical concerns and stresses on testing. In a different section, the students were given the same lectures and main exams, but they did not receive the daily quizzes. In comparing the two groups, the class that had the daily quizzes earned better grades at the end of the semester on the exams than did the group without daily quizzes. More important for present purposes, however, were students' perceptions of how quizzes affected them academically. A year-end survey indicated that students felt that the quizzes (a) gave them a chance to practice questions that would be similar to exam questions, (b) helped identify important topics in the course, (c) caused students to come to lectures more often, (d) caused students to pay more attention, and (e) allowed students to better understand what they had learned during each lecture. Clearly, students had a positive attitude toward the daily quizzes.

As mentioned earlier, self-testing can help students identify what information they do or do not know, which in turn can lead to decisions about how to allocate study time. The relationship between what a student initially learns, their metacognitive judgments of what they think they know, and how they choose to study have a complex relationship with actual test performance. One model of study time allocation is called the discrepancy reduction framework (Dunlosky & Hertzog, 1998). The idea is that students have a goal state of knowledge that they wish to attain and they allocate their study opportunities to reduce the discrepancy between their current knowledge state and that they hope to achieve. Simply put, if students already know some topic reasonably well, they will not study it; if they are quite ignorant of another topic they need to know, they will devote their study efforts to that topic. In short, students will be most likely to study first the most difficult information facing them. Indeed, Nelson, Dunlosky, Graf, and Narens (1994) showed that judgments of learning for studied items were negatively correlated with additional study time; that is, items that subjects thought they knew well were not selected for further study and items that were judged most difficult received the most study time.

However, one criticism of the discrepancy reduction model for study time allocation is the assumption that students will have unlimited time to study. When a time constraint is introduced, the choices students make about what items to study change significantly. Often students tend to study not the most difficult material, but material in the medium range of difficulty, material just out of their current reach. Metcalfe (2002) and Kornell and Metcalfe (2006) developed the region of proximal learning framework to account for these new results. Essentially, their model suggests that students will try and learn the most difficult items that they will be able to learn in the time frame. If time is limited, then students will often not study the most difficult items, since they will not be able to learn them before time is up. Kornell and Metcalfe (2006) provided results

supporting the region of proximal learning framework and also showed that student learning was more effective when students chose what to study than when the items were assigned by the experimenter. This outcome suggests that, at least at the level of selecting individual pieces of knowledge to study, students know how to make study choices that will ultimately benefit their own future test performance.

Yet in other ways, students are not good at choosing what, when, how, and how long to study. Nelson and Leonesio (1988) showed that even if subjects are given unlimited time to study, they often continue to study even when the efforts result in no additional gain in performance (an effect they called "labor in vain"). Similarly, Karpicke (2009) showed that if students chose to drop materials from study after an initial recall (which they often did), they would perform worse compared to a repeated retrieval condition.

In conclusion, frequent testing encourages students to study and also permits them to comprehend the gaps in their knowledge (our second benefit). Thus, testing permits students some accuracy in choosing what to study in some circumstances, but in other situations they may make poor choices (Karpicke, 2009; Kornell & Bjork, 2007). Students often choose to stop studying before they have mastered material and will often choose to mass their study immediately before a test rather than spacing it out. Integrating more tests across the course of the semester will encourage students to study more consistently throughout the semester, which will increase performance.

12. Possible Negative Consequences of Testing

We have reviewed 10 benefits that we believe testing confers on learning and memory, directly or indirectly. Yet our message is slow to permeate the educational establishment. Critics have raised a number of objections to any emphasis on testing in the schools (whether achievement testing or giving frequent classroom tests). The arguments against testing range from philosophical to empirical. Some of the latter criticisms are valid, and we have already briefly considered some of the issues in the chapter. Here, we cover this ground rather rapidly because we have touched on these issues in earlier parts of this chapter or in previous writings (see Roediger & Karpicke, 2006b).

First, quizzing in class may take time away from other critical classroom activities, such as lectures, discussion, and demonstrations. Is that a problem? This point is true to an extent, but how does one know (in absence of proper studies) whether these activities are better than retrieval via quizzing? For example, Karpicke and Blunt (2011) showed

that retrieval practice produced better retention later than did concept mapping, a widely used study technique. We expect that when other such studies are conducted, they may show that some quizzing is as beneficial as, or more beneficial than, an equal amount of time spent on lecturing (just as testing is better than restudying). In addition, as discussed above, having classroom quizzes may keep motivation up and provide the indirect benefit of having students study more. At any rate, we do not think this criticism holds water, but future research may change our opinion.

Second, critics sometimes argue that retrieval practice through testing produces "rote" learning of a superficial sort, as if the student can parrot back the information but not really understand it or know it in a deep fashion. Learning is said to become "inert" or "encapsulated" in little factoid bubbles. Perhaps this criticism is justified in some cases, but we think that good programs of quizzing with feedback usually prevent this problem. We reviewed evidence previously showing that retrieval (via testing) can lead to deep knowledge that can be used flexibly and transferred to other contexts (e.g., Butler, 2010). Again, the burden is on the critics to show that testing leads to problems rather than simply asserting that these problems might exist. The next two criticisms are based on data and must be taken more seriously.

Third, many studies have documented a phenomenon variously called output interference (Tulving & Arbuckle, 1966), the inhibitory effects of recall (Roediger, 1974, 1978), or retrieval-induced forgetting (Anderson et al., 1994). The basic phenomenon is that while the act of retrieval may boost recall of the retrieved information (the testing effect), it can actually harm recall of nontested information. We discussed this point in Section 7. Thus, in educational settings, the fear is that if students repeatedly retrieve some information, they may actually cause themselves to forget other information.

There is now a vast literature on these topics (see Bäuml (2008) for a review). Although the various phenomena encapsulated under the rubric of retrieval-induced forgetting are highly reliable, as we discussed in Section 7, the implications for educational practice may not be great. For one thing, the phenomenon is often short lived, so if a delay is interposed between retrieval practice and testing, the inhibition dissipates or even evaporates altogether (MacLeod & Macrae, 2001). In addition, most experiments on retrieval-induced forgetting have used word lists. As noted in Section 7, when well-integrated materials such as prose passages are used, the inhibition effect can disappear (Anderson & McCulloch, 1999) or even reverse altogether, leading to retrieval-induced facilitation (Chan et al., 2006). As discussed previously, Chan and his collaborators (see also Chan, 2009, 2010) showed that testing can sometimes enhance recall of material related to the tested material. Thus, although much research remains to be done, the various phenomena showing that testing

of some material can have negative effects on retrieval of other material may not have strong implications for the kinds of material learned in educational settings.

A fourth issue of concern about testing is that the construction of some tests themselves can lead to acquisition of erroneous knowledge. Although educators would never consider knowingly providing erroneous information during lectures or in assigned readings, they do it all the time when they give certain types of tests. In true/false tests, students are given a set of statements and asked to judge which are true and which are false. Of course, false items are often tricky, incorporating some true and some false elements. Thus, students are forced to consider erroneous information and perhaps they will even judge some false statements as true. Similarly, in the more commonly used multiple-choice test, students are given a stem and then four choices to complete the stem. Three of the choices supply incorrect information, so students have to ponder these erroneous statements. Unfortunately, a well-known principle in cognitive psychology is the "mere truth effect," the fact that repeatedly exposed statements gain credibility and are judged more likely to be true regardless of their truth value (Hasher, Goldstein, & Toppino, 1977; see also Bacon, 1979; Begg, Armour, & Kerr, 1985). Thus, because (as we have repeatedly seen in the course of this chapter) students learn from tests, the danger exists that students who are exposed to wrong information on tests will learn that information. Remmers and Remmers (1926) raised the specter of such difficulties long ago and termed possible negative effects of testing the negative suggestibility effect. Ironically, their own research did not show much to worry about, but more recent studies have shown that negative suggestibility is real, at least on true/false and multiple-choice tests.

Toppino and Brochin (1989) had students take true/false tests. On a later occasion, they then asked the students to judge the truth of objectively false statements they had seen before mixed in with new (equivalent) statements they had not seen before. Sure enough, students judged the previously read statements as truer than the new statements. Toppino and Luipersbeck (1993) extended this finding to multiple-choice tests. The wrong choices on the multiple-choice tests were later judged to be truer than other distracter items (see also Brown, Schilling, & Hockensmith, 1999).

Roediger and Marsh (2005) had students take multiple-choice tests using a design in which both positive and negative effects of testing could be measured on later cued recall test. Are negative suggestibility effects so great that they will overcome the positive effects of testing? Without going into the details of the experiment, Roediger and Marsh found both positive and negative effects of taking a multiple-choice test on a later cued recall test. When students got an answer right on the multiple-choice test, their performance was boosted on a later cued recall test for

the information. However, when they answered erroneously, the negative suggestibility effect occurred: students tended to supply the wrong answer on the cued recall test later at levels much greater than that in the control condition (see also Fazio, Agarwal, Marsh, & Roediger, 2010; Marsh, Roediger, Bjork, & Bjork, 2007). However, the positive effects of testing outweighed the negative suggestibility effect in these studies. Interestingly, the same pattern of results occurs on the widely used Scholastic Assessment Test (the SAT; Marsh, Agarwal, & Roediger, 2009), and in one study in that series in which students did very badly on the initial multiple-choice form of the SAT, the negative effects outweighed the positive effects on the final test given later.

Although these negative suggestibility that effects on multiple-choice tests are quite real, they can be overcome simply by providing feedback on the tests (Butler & Roediger, 2008). Feedback increases the testing effect for items answered correctly and overcomes the negative suggestibility effect for items given erroneous answers (see also Butler, Karpicke, & Roediger, 2007, 2008; Pashler, Cepeda, Wixted, & Rohrer, 2005).

In sum, we have considered four possible negative consequences of testing. The most serious of these is the negative suggestibility effect on true/false and multiple-choice tests, but if feedback is provided after the tests, even this difficulty disappears. As long as students receive feedback on their exams, we see no major drawbacks in using tests as a learning mechanism (either from quizzes in class or self-testing as a study tool).

13. Conclusion

We have reviewed 10 reasons why increased testing in educational settings is beneficial to learning and memory, as a self-study strategy for students or as a classroom tactic. The benefits can be indirect—students study more and attend more fully if they expect a test – but we have emphasized the direct effects of testing. Retrieval practice from testing provides a potent boost to future retention. Retrieval practice provides a relatively straightforward method of enhancing learning and retention in educational settings. We end with our 10 benefits of testing in summary form:

Benefit 1: The testing effect: Retrieval aids later retention.
Benefit 2: Testing identifies gaps in knowledge.
Benefit 3: Testing causes students to learn more from the next learning episode.
Benefit 4: Testing produces better organization of knowledge.
Benefit 5: Testing improves transfer of knowledge to new contexts.
Benefit 6: Testing can facilitate retrieval of information that was not tested.

Benefit 7: Testing improves metacognitive monitoring.
Benefit 8: Testing prevents interference from prior material when learning new material.
Benefit 9: Testing provides feedback to instructors.
Benefit 10: Frequent testing encourages students to study.

Finally, testing can of course be relied on to fulfill its traditional functions: Permitting instructors to assign grades to students.

REFERENCES

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

Amlund, J. T., Kardash, C. A., & Kulhavy, R. W. (1986). Repetitive reading and recall of expository test. *Reading Research Quarterly*, *21*, 49–58.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (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.

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

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.

Bäuml, K. H. (2008). Inhibitory processes. In H. L. Roediger (Ed.), *Cognitive psychology of memory* (pp. 195–217). Vol. 2 of *Learning and Memory: A comprehensive reference*, 4 vols (J. Byrne, Ed.). Oxford: Elsevier.

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

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.

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

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

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.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1118–1133.

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

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

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

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.

Carpenter, S. K., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830.

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

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*, 153–170.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*, 49–57.

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

Congleton, A., & Rajaram, S. (2010, November). *Examining the immediate and delayed aspects of the testing effect*. Paper presented at the meeting of Psychonomic Society, St. Louis, MO.

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*, 919–940.

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.

Detterman, D. K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D. K. Detterman, and R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 1–24). Westport, CT: Ablex Publishing.

Dunlosky, J., & Hertzog, C. (1998). Training programs to improve learning in later adulthood: Helping older adults educate themselves. In D. J. Hacker, J. Dunlosky, and A. C. Graesser, (Eds.), *Metacognition in educational theory and practice* (pp. 249–275). Mahwah, NJ: Erlbaum.

Ebbinghaus, H. (1885). Über das Gedächtnis. Leipzig: Duncker & Humblot.

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.

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, *38*, 407–418.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the under-confidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 238–244.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*, 19–34.

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

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

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

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112.
Izawa, C. (1966). Reinforcement-test sequences in paired-associate learning. *Psychological Reports*, *18*, 879–919.
Izawa, C. (1968). 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.
Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1441–1451.
James, W. (1980). *The principles of psychology.* New York: Holt.
Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*, 621–629.
Jones, H. E. (1923). The effects of examination on the performance of learning. *Archives of Psychology*, *10*, 1–70.
Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*, 469–486.
Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775.
Karpicke, J. D., Butler, A. C., & Roediger, H. L. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory*, *17*, 471–479.
Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162.
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968.
Kelly, C. M. (1999). Subjective experience as a basis for "objective" judgments: Effects of past experience on judgments of difficulty. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII*, 515–536.
Koriat, A., Scheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*, 147–162.
Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224.
Kornell, N., & Metcalfe, J. (2006). Study efficacy and the region of proximal learning framework. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 222–609.
Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*, 493–501.
Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212.
Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*, 94–97.
MacLeod, M. D., & Macrae, C. (2001). Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychological Science*, *12*, 148–152.
Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, *15*, 1–11.
Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, *14*, 194–199.

Masson, M. E., & McDaniel, M. A. (1981). The role of organizational processes in long-term retention. *Journal of Experimental Psychology: Human Learning and Memory, 2*, 100–110.

Mawhinney, V. T., Bostow, D. E., Laws, D. R., Blumenfeld, G. J., & Hopkins, B. L. (1971). A comparison of students studying-behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis*, *4*, 257–264.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*, 462–476.

McDermott, K. B., & Arnold, K. M. (2010, November). *Test taking facilitates future learning.* Paper presented at the meeting of the Psychonomic Society, St. Louis, MO.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131*, 349–363.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*, 174–179.

Michael, J. (1991). A behavioral perspective on college teaching. *Behavioral Analysis*, *14*, 229–239.

Nelson, T. O., Dunlosky, J., Graf, A., & Narens, L. (1994). Utilization of metacognitive judgments in the allocation of study during multitrial learning. *Psychological Science*, *5*, 207–213.

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

Newton, L. (1990). *Overconfidence in the communication of intent: Heard and unheard melodies.* Unpublished doctoral dissertation. Stanford, CA: Stanford University.

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

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, *35*, 1917–1927.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*, 335.

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

Roediger, H. L. (1974). Inhibiting effects of recall. *Memory & Cognition*, *2*, 261–269.

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

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

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

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

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, *76*, 45–48.

Rohrer, K., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 233–239.

Shaughnessy, J. J., & Zechmeister, E. B. (1992). Memory-monitoring accuracy as influenced by the distribution of retrieval practice. *Bulletin of the Psychonomic Society, 30*, 125–128.

Slamecka, N. J., & Katsaiti, L. T. (1988). Normal forgetting of verbal lists as a function of prior testing. *Journal of Verbal Learning and Verbal Behavior*, *10*, 400–408.

Son, L. K., & Kornell, N. (2008). Research on the allocation of study time: Key studies from 1890 to the present (and beyond). In J. Dunlosky, & R. A. Bjork, (Eds.), *A handbook of memory and metamemory* (pp. 333–351). Hillsdale, NJ: Psychology Press.

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

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*, 1007–1013.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2008). Testing during study insulates against the buildup of proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1392–1399.

Thomas, A. K., & McDaniel, M. A. (2007). Metacomprehension for educationally relevant materials: Dramatic effects of encoding-retrieval interactions. *Psychonomic Bulletin & Review*, *14*, 212–218.

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

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

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

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

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

Underwood, B. J. (1957). Interference and forgetting. *Psychological Review*, *64*, 49–60.

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

Zaromb, F. M. (2010). Organizational processes contribute to the testing effect in free recall. (Unpublished doctoral dissertation). Washington University of St. Louis, Saint Louis, MO.

Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *38*, 995–1008.