2. Learning from the Test: Dos and Don’ts for Using Multiple-Choice Tests

doi:10.7936/K7Z60KZK
Elizabeth J. Marsh and Allison D. Cantor

Duke University

Correspondence:
Elizabeth J. Marsh
Psychology & Neuroscience
Duke University
417 Chapel Drive
Durham NC 27708-0086
(919) 660-5796
emarsh@psych.duke.edu
http://marshlab.psych.duke.edu/

Abstract

Multiple-choice tests are ubiquitous in the classroom; while typically used for assessment, our focus in this chapter is on how such tests can also serve as learning opportunities for students. We review evidence from cognitive psychology that multiple-choice tests can change what students know, helping them to remember forgotten information, boosting retention of recently learned information, and even promoting new learning. However, the educator needs to exercise care when using multiple-choice tests, because by definition multiple-choice questions pair correct answers with plausible but incorrect lures. That is, multiple-choice testing can also yield a negative testing effect, whereby prior exposure to multiple-choice questions boosts the likelihood that students will use multiple-choice lures to answer later general-knowledge questions. We evaluate a number of solutions to this problem, with the goal of maximizing the benefits and minimizing any costs of multiple-choice testing. While a number of possible solutions involve changes to test construction (e.g., changes to the plausibility and number of lures), the best solution turns out to be a simple one: educators should make sure to tell students the correct answers after they complete the multiple-choice test. We conclude with a discussion of future directions for research, with an emphasis on the need for additional studies in classroom settings.
Introduction

Scantron® answer sheets are a common sight in classrooms across America, albeit sometimes an unwelcome one. The drill is a familiar one: bring one’s number two pencil and fill in the bubbles corresponding to one’s answers to a series of multiple-choice questions. Such multiple-choice tests are commonly used to evaluate student knowledge, with educators choosing them over other types of tests because multiple-choice questions are readily available (e.g., through test banks) and easy to grade. Furthermore, both teachers and students perceive the grading of multiple-choice tests as relatively objective. For these and other reasons, multiple-choice tests are unlikely to disappear from our classrooms. Consequently, much research has been focused on how to construct multiple-choice tests to best measure what students do versus don’t know, without introducing bias, and how to create questions that go beyond fact knowledge to tap higher-level cognitive skills like applying and reasoning from knowledge (e.g., Frederiksen, 1984). While these issues are important ones, our focus in this chapter is different. Instead of focusing on multiple-choice tests as assessment devices, our focus is on multiple-choice testing as a learning tool. That is, while we are all familiar with the use of tests to determine how much students have learned from a text, lecture, or other resource, less familiar is the perspective that tests not only measure learning but also—importantly—change learning.

Known as the testing effect in the psychology literature, numerous empirical studies have demonstrated that the act of retrieving something from memory is a powerful way to promote its later retention (see Roediger & Butler, 2011, for a review). This phenomenon is best explained with an example, so we begin with a description of how testing improved performance in a web-based university course on Brain and Behavior (McDaniel, Anderson, Derbish, & Morrisette, 2007). As would be expected in such a course, students were assigned reading from the course textbook each week. The key manipulation involved whether or not students practiced the course material after (allegedly) reading it, and if so what form that practice took. Of interest was how well students learned facts like “All preganglionic axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter.” For facts in the short-answer practice condition, students were quizzed with questions like “All preganglionic axons, whether sympathetic or parasympathetic, release __________ as a neurotransmitter” and had to fill in the blanks. For facts in the multiple-choice practice condition, students were quizzed with questions like “All preganglionic axons, whether sympathetic or parasympathetic, release __________ as a neurotransmitter” and selected a response from four options (e.g., “acetylcholine,” “epinephrine,” “norepinephrine,” and “adenosine”). In both quiz conditions, students received feedback about their answers. Additional facts were assigned to a re-exposure control condition involving re-reading the facts or to a no-exposure control where the facts were not practiced (neither quizzed nor re-read).
Our focus is on how these practice conditions affected performance on two multiple-choice unit tests, each of which covered three weeks' worth of material. Figure 1 depicts performance collapsed across the two unit tests. Overall, there was no benefit on unit test performance from re-reading the facts, as compared to the no-exposure control. However, there was a benefit of prior quizzing, with students later remembering more facts previously tested on multiple-choice or short-answer quizzes, as compared to the no-exposure control. This testing effect cannot be solely attributed to re-exposure of material, since the re-reading control did not yield a benefit over the no-exposure control. Briefly, testing benefits memory because it is a good match to what students later will be asked to do: retrieve information from memory (no teacher asks students to simply read on exams). Returning to our earlier example, answering the question “All preganglionic axons, whether sympathetic or parasympathetic, release ________ as a neurotransmitter” requires the student to retrieve from memory the meanings of the terms as well as the answer. In the multiple-choice version of the question, the learner still needs to retrieve the meanings of the terms in the question prompt as well as those in the response alternatives. The student might also reason her way to the correct answer to the multiple-choice question, retrieving related knowledge that allows her to reject some of the lures. Consistent with these ideas, many studies have shown that learners benefit from answering both short-answer (Butler & Roediger, 2007; Kang, McDermott, & Roediger, 2007) and multiple-choice questions, retaining more information on later tests (e.g., Roediger & Marsh, 2005; Butler, Karpicke, & Roediger, 2007).

![Figure 1](image.png)

**Figure 1.** Unit exam performance for practiced items (which were read, tested in multiple-choice format, or tested on a short-answer test), as compared to no-exposure control items. Adapted from McDaniel et al. (2007).

A second result in Figure 1 highlights a puzzle in the literature. McDaniel et al. (2007) found that the benefit of prior quizzing was not as great for facts previously tested in multiple-choice format than in short-answer format. This result is consistent with the traditional view
of multiple-choice tests as requiring less effort from the student, which in turn would translate into less benefit to the learner (as much research has shown that difficult practice pays off in the long run; the concept of desirable difficulties; Schmidt & Bjork, 1992). Consistent with this perspective, a number of studies have shown better long-term retention following short-answer testing than multiple-choice testing (e.g., Butler & Roediger, 2007; Kang et al., 2007; McDaniel et al., 2007). However, more recent work has shown the opposite pattern, with better retention following multiple-choice testing than short-answer testing (e.g., Little, Bjork, Bjork, & Angello, 2012; McDermott, Agarwal, D’Antonio, Roediger, & McDaniel, 2014). What is safe to say is that testing benefits the learner; less clear is whether it matters as much what format the questions take, so long as they require retrieval practice.

**What might be different about multiple-choice tests?**

As already alluded to, multiple-choice tests may sometimes involve less extensive retrieval practice. The structure of multiple-choice tests implies additional differences in learning may result, as compared to using other types of tests. Consider a question like the following: *The maintenance of a constant internal salt concentration by brine shrimp is called: phagocytosis, spermatogenesis, parthenogenesis, homeostasis, or peristalsis.* The provision of response alternatives raises the possibility of several scenarios unlikely with short-answer questions.

**Possible Benefits.**

First, the provision of multiple-choice alternatives suggests some unique benefits to the learner. There is the possibility of learning new information from the test: for example, consider the possibility that the learner does not know the answer to the above question and would not be able to generate the correct answer “homeostasis” in response to the prompt but might be able to reason her way to the answer based on knowledge of the other response alternatives (e.g., she knows what spermatogenesis is and thus can eliminate that response option).

A related situation likely occurs when the learner does in fact know something and yet is unable to call it to mind. Most students are familiar with the tip-of-the-tongue state and the feeling that they know something but cannot produce it at that moment in time (this experience has often been compared to being on the brink of a sneeze). Cognitive psychologists use the term *marginal knowledge* to refer more generally to knowledge that is stored in memory but is momentarily inaccessible, regardless of whether or not the learner is in a tip-of-the-tongue state (Berger, Hall, & Bahrick, 1999). While cognitive psychologists may often focus on the acquisition of new knowledge, it remains a major goal of education to retain learned information and prevent forgetting (in other words, to prevent knowledge from becoming marginal). Returning to our focus on multiple-choice testing, answering
multiple-choice questions is a powerful way to stabilize access to marginal knowledge (Cantor, Eslick, Marsh, Bjork, & Bjork, in press). This work draws on elegant work by Berger and colleagues (1999), who showed that a simple five-second re-exposure to information was enough to stabilize access to marginal knowledge after nine days. In other words, given that a student failed to answer a question like “What is the last name of the person who proposed the Theory of Relativity?” a five-second exposure to “Einstein” was sufficient to help students retrieve that fact nine days later. How did the researchers know that they were stabilizing access to previously stored knowledge, as opposed to teaching students new facts? To tease these possibilities apart, they examined the effects of giving feedback in response to failures to retrieve true facts (like the example just given) versus false facts (e.g., a failure to answer “What was the last name of the person who proposed the theory of maladaptability?”). The key logic is that students can only have marginal knowledge for true facts, whereas any reproduction of false facts must represent new learning. After nine days, retention of the false facts was at zero; the feedback only helped retention of true facts for which marginal knowledge was possible. The key point here is that similar effects are found when students answer multiple-choice questions, even if they never receive feedback on their answers. This benefit is driven by the student’s ability to correctly select the right multiple-choice option, which serves to expose them to the correct answer. Multiple-choice testing was just as powerful as studying in fact, helping students to produce facts one week later that they had failed to produce on an initial short-answer test (Cantor et al., in press).

Possible Costs.

Thus far we have outlined two benefits that may be unique to multiple-choice tests, but multiple-choice tests may also pose unique hazards to the student. Aside from the fact that multiple-choice tests may sometimes provide less extensive retrieval practice than short-answer tests, the concern is that by definition a multiple-choice question exposes the learner to incorrect answers. In almost no other situation would an educator expose students to wrong information, but multiple-choice testing requires the student to read and consider plausible wrong information. In other words, returning to our example, what happens when the student decides that “phagocytosis” is the process by which a brine shrimp maintains a constant internal salt concentration—does he or she leave the testing situation with a false belief? This chapter focuses on evaluating this possible negative consequence of multiple-choice testing, both because we think it is a relatively unfamiliar concern to educators (but one, as discussed below, that may be avoided with a few precautions) and because the likely persistence of multiple-choice testing in our educational system means it is important to understand how to minimize any problems.
Unfortunately, several experiments have demonstrated that prior multiple-choice testing increases the likelihood that students will incorrectly answer later questions with multiple-choice lures; we refer to this effect as a negative testing effect. Consider a sample experiment, where students answered multiple-choice questions in a laboratory setting; of interest were the effects of answering (retired) SAT II questions on later performance (Marsh, Agarwal, & Roediger, 2009). On an initial multiple-choice test, Duke University undergraduates answered biology, chemistry, U. S. History, and World History questions. Each multiple-choice question was paired with the correct answer and four distractors, and included a “don’t know” option. Students were instructed to treat the experiment as if they were actually taking an SAT II test and were informed they would receive one point for correct answers, lose one-fourth of a point for incorrect answers, and lose no points for skipping questions. After completing the initial multiple-choice test, subjects completed a filler task for about five minutes (to ensure no answers were held in short-term memory) and then took a final short-answer test. In this test, questions were not paired with answer choices; answering a question required subjects to produce a response. The questions included the forty items from the initial test as well as forty new questions that served as a baseline measure of performance without prior testing.

The left side of Figure 2a shows the benefits of prior multiple-choice testing: students answered more questions correctly on the final test if the questions had appeared on the initial multiple-choice test (a positive testing effect). This effect was very robust—students answered many more questions correctly after testing. For example, students answered 26 percent more short-answer questions correctly when the questions had appeared on the initial multiple-choice test than when they had not.
Figure 2. Production of correct answers and multiple-choice lures on the short-answer final test as a function of whether the questions appeared on the initial multiple-choice test. Panel A shows the data from undergraduates and Panel B shows the data from high school juniors. Data from Marsh et al. (2009).

However, a negative testing effect was also observed, as shown in the right side of Figure 2a. Final short-answer questions were more likely to be answered with the multiple-choice lures if the questions had appeared on the initial test. Students were more likely to answer “What is the term that describes how brine shrimp maintain a constant internal salt concentration?” with one of the multiple-choice lures (e.g., phagocytosis) if the question had had appeared on the initial multiple-choice test than if it had not.

An examination of Figure 2a reveals an important point: the positive testing effect was much larger than the negative testing effect. Overall, even though undergraduates picked up some multiple-choice lures from the initial test, multiple-choice testing benefited them much more than it impaired performance.

However, Figure 2b reveals a very different conclusion. These data come from a very similar experiment, using the same materials and very similar methods, with the crucial difference that the experiment was conducted with juniors in an Illinois public high school. The high school students again showed both positive and negative testing effects: having answered the questions on the initial multiple-choice test increased both correct and multiple-choice lure answers on the final short-answer test. However, compared to the undergraduates, the high school students showed a smaller benefit from prior testing and a larger cost. Therefore,
while multiple-choice testing was overall beneficial for the undergraduates, it was much less helpful for these high school students.

To understand these group differences, we need to consider how the students performed on the initial multiple-choice test. Critically, the undergraduates answered 55 percent of the initial multiple-choice questions correctly and endorsed multiple-choice lures 22 percent of the time (they chose “I don’t know” for the remaining 23 percent of questions). In contrast, the high school students only answered 34 percent of initial questions correctly, endorsing multiple-choice lures for a whopping 56 percent of questions (choosing “I don’t know” for only 9 percent of questions). This difference is crucial because it turns out that across studies, students only reproduce multiple-choice lures that they selected when answering the multiple-choice questions (Roediger & Marsh, 2005; Butler, Marsh, Goode, & Roediger, 2006). Merely reading multiple-choice lures does not appear to have any consequences, given that students reject them. The problem is these high school students had less knowledge and also appear to have less metacognitive knowledge about what they do versus do not know (and thus take less advantage of the “don’t know” option). Returning to the ideas discussed earlier, the problem for the high school sample was that the multiple-choice test represented an undesirable difficulty for them; it is not useful to make practice so hard that students cannot succeed. It is important to note that the results shown in Figure 2 are likely more about ability to succeed on the multiple-choice test than about undergraduates vs. high school students per se. A study comparing students from an elite high school to students from a junior college might even find the reverse pattern, if the elite high school students scored better on the multiple-choice test.

The negative testing effect has been observed in numerous studies (e.g., Roediger & Marsh, 2005; Odegaard & Koen, 2007). It holds even when only the highest confidence responses are observed or when students are warned against guessing on the final test (Fazio, Agarwal, Marsh, & Roediger, 2010; Roediger & Marsh, 2005; Roediger, Agarwal, Kang, & Marsh, 2010). There is some evidence that students are not simply memorizing a response, but rather are picking up a false belief (Marsh, Roediger, Bjork, & Bjork, 2007). That is, after students answer multiple-choice questions like “What biological term describes fish slowly adjusting to water temperature in a new tank?” they are more likely to use a multiple-choice lure to answer a question about a new application of the same concept (e.g., Animals that thicken their fur during winter are exhibiting what scientific phenomena? Correct Answer: acclimation). Overall, the negative testing effect appears to be real and the challenge becomes minimizing it while keeping the benefits of multiple-choice testing.

**Evaluating Solutions**
Because of the ubiquity of multiple-choice testing as well as the fact that (in some circumstances) it can benefit the learner, it is important to evaluate possible solutions for eliminating the negative testing effect. We have already alluded to one solution: test difficulty needs to be properly calibrated, so that it is difficult enough for students to have to retrieve information but not so difficult that students endorse a large number of multiple-choice lures (as was the case with the high school students whose data is depicted in Figure 2b). This solution may not be feasible, however—instructors may not have a sense of how difficult a test will be until they administer it. Furthermore, given that tests typically serve an assessment function, an instructor is unlikely to want to limit the range of performance.

**Passive Solutions.**

The educator will be relieved to hear that the easiest solution is a very passive one—simply wait for time to pass. Research has found that both the positive and negative testing effects are largest immediately after testing and that both effects diminish over time (Fazio et al., 2010). However, the catch is that time will also diminish the positive effects of testing, so what are really needed are solutions that keep the positives but lose the negatives of multiple-choice testing.

**Active Solutions: Changing the Test.**

We begin our evaluation of active strategies with a discussion of solutions that involve changes in test construction, some of which may be more feasible to the educator than others. One issue involves the nature of the multiple-choice lures, which of course can vary considerably—the same question could be very easy or very hard depending on the answer choices. Little and colleagues (2012) argued that multiple-choice questions can be designed to promote even greater learning than short-answer questions if questions are designed to have plausible lures that require students to consider both why each correct answer is the right choice and why each lure is incorrect. In other words, Little et al. argued that good multiple-choice questions require a lot of retrieval practice, as opposed to simple recognition of the correct answer. While intriguing, we believe this research is still in the early stages and is not yet ready for classroom implementation. One issue is that there is no “roadmap” for how to make these types of items; how does an educator know (other than relying on intuition) that an item requires evaluation of both the correct answer and the lures? Furthermore, this benefit depends upon students having the requisite knowledge stored in memory. In the example given earlier, if the student knows what spermatogenesis is she can reject it. However, the concern is that plausible distractors may increase the likelihood that students select multiple-choice lures as their answers, which in turn would increase the likelihood of yielding a negative testing effect. In the Little et al. study, performance on the initial multiple-
choice test was quite high, but one can easily imagine a situation where the lures are so plausible that performance on the multiple-choice test plummets, reducing the positive and increasing the negative testing effect (yielding performance like that depicted in Figure 2b). Clearly, however, understanding how to design effective multiple-choice questions is an important direction for future research.

Because instructors may not have the time or ability to write new multiple-choice distractors, we must consider whether simpler modifications might help. We will evaluate whether there are any benefits from changing the number of lures (as opposed to their content), including a “none of the above” option, or allowing students to skip questions. All of these options have some appeal because they would be easy for instructors to implement and potentially may reduce the likelihood that multiple-choice lures will be endorsed.

We begin by considering whether it matters how many lures are included for each multiple-choice question. That is, should a multiple-choice prompt be paired with 2, 3, 4, or 5 answer choices (or some other magic number)? On the one hand, with fewer answer choices, the student will be less likely to endorse a lure (with consequent reductions in the negative testing effect). But on the other hand, an increased number of lures might be conceptualized as a desirable difficulty, which should boost learning. Both of these positions have support in the literature: sometimes increasing the number of lures boosts memory, and sometimes it hurts memory (see Butler et al., 2006, for a review). The key determinant is how well students do on the initial multiple-choice test and hits upon a familiar theme: difficulties are desirable if and only if the learner can overcome them. That is, memory benefits when students face extra multiple-choice lures and still select the correct answer. But when an increased number of multiple-choice lures is accompanied with an increase in errors, memory suffers. Because there is no clear general answer to the question “How many alternatives is best?” we recommend sticking with what students are used to (normally 4 or 5 response options).

A similar issue arises when we evaluate whether or not to recommend including a “none-of-the-above” choice. The answer depends on whether that choice is a correct or incorrect answer (Odegard & Koen, 2007). When none-of-the-above is incorrect (meaning that the correct answer is present in the list of response options), a positive testing effect emerges. However, it is problematic when “none-of-the-above” is the correct answer (meaning that the correct answer is not present in the list of response options). In this situation, no benefits of initial testing are observed on a later short-answer test, presumably because students are less likely to engage in retrieval practice of the correct answer when it was not in the list of response options. Furthermore, students are more likely to later answer short-answer questions with multiple-choice lures, presumably because they are more likely to endorse lures when the correct answer was not present. Because it is hard to envision a situation
where "none-of-the-above" is never right, it is not clear that this is the best modification to make to a multiple-choice exam.

The final simple test modification we will evaluate involves giving students the option to skip questions. In the studies depicted in Figure 2, students were given the option to skip; students lost no points for skipping questions but lost one-fourth of a point for each multiple-choice lure endorsed (Marsh et al., 2009). In another experiment, only one-half of the undergraduates were given the option to skip questions whereas other students were required to answer every question (inevitably increasing errors). On a later test, the negative testing effect was smaller in the group given the option to skip, although this reduction, while statistically significant, was relatively small (a drop of 3 percent). In addition, it is not clear how general any benefit would be since it depends upon people's ability to effectively determine what they do versus do not know. If people don't know when to say "don't know," they will not show any benefits; for example, it is unlikely that the option to skip helped the high school students whose data are depicted in Figure 2b, since they rarely took advantage of it.

**Active Solutions: Providing Feedback.**

We conclude this section with a discussion of the best-known strategy for avoiding the negative testing effect, namely the provision of feedback. By feedback, we mean giving the student information about the correctness of their responses (this information can take several forms, as described below). Providing feedback does not require the instructor to modify his or her test and sometimes can be provided automatically (e.g., it is an option that can be selected on most Scantron® machines). Figure 3 shows the power of feedback in eliminating the negative testing effect, drawing on the relevant conditions from an experiment by Butler and Roediger (2008). Students read a series of passages, and some took a multiple-choice test on the content they had just studied. Tested subjects received no feedback, immediate feedback, or delayed feedback on their responses. Immediate feedback took the form of the correct answer and was delivered immediately after each response, whereas delayed feedback also consisted of the correct answer but was delivered at the end of the test. A week later subjects returned and took a final short-answer test on the material; the data presented here come from a condition where subjects were required to answer all questions on that final test (as is often the norm in education, where students are typically better off guessing than not responding). Students used more multiple-choice lures to answer final questions in the Tested – No Feedback condition than in the not tested control (the negative testing effect), but this effect was eliminated when feedback was presented immediately or after a delay. Figure 3 only shows the production of multiple-choice lures as answers on the final test, but it should be noted that feedback also boosted correct answers on
the final test. Similar results have been obtained with learners as young as seven years of age (Marsh, Fazio, & Goswick, 2012). Overall, feedback is a powerful way for educators to both increase the size of the positive testing effect and decrease the size of the negative testing effect.

![Bar chart showing proportion of false answers](image)

*Figure 3. Proportion of final short-answer questions answered with multiple-choice lures, as a function of initial learning condition. Data from Butler and Roediger (2008).*

When administering feedback, there are a few key decisions an educator must make. We have already alluded to one such decision, namely the choice of when to provide the feedback. In the Butler and Roediger (2008) study just described, immediate and delayed feedback were similarly effective at reducing the negative testing effect (see also Butler *et al.*, 2007). Delaying the feedback may, however, boost retention of correct responses (e.g., Butler & Roediger, 2008; Butler *et al.*, 2007). Benefits of delayed feedback are likely due to the fact that it provides another spaced presentation of the to-be-learned material; spacing is known to promote superior long-term retention of material relative to massed presentation (See Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006 for a review). Thus on balance, delaying the feedback may be advised—with the important caveat that the educator must ensure that students process the feedback carefully (which may be less likely if feedback is delayed).

A second decision concerns the content of the feedback, which could range from indicating whether an answer is right or wrong (verification feedback) to providing the correct answer (answer feedback) or an explanation of the answer (elaborative feedback). Our example earlier (Figure 3) involved answer feedback; of particular relevance for present purposes is the distinction between verification and answer feedback since many Scantron® machines provide instructors with both of these options. While much research suggests that verification feedback is not particularly effective for correcting errors (e.g., Pashler, Cepeda, Wixted, & Rohrer, 2005), most of this work has examined correcting errors made on short-answer tests.
Verification feedback provides more information to the learner after a multiple-choice selection than after a response to an open-ended question. Labeling an open-ended response as incorrect does nothing to winnow down the possible alternatives, whereas labeling a multiple-choice selection as incorrect constrains the possible choices. These ideas were tested in a study where students answered multiple-choice questions such as “What is the capital of Belize?” (Correct answer: Belmopan) and received no feedback for one-third of responses, verification feedback for another third, and answer feedback for the rest of their responses (Marsh, Lozito, Umanath, Bjork, & Bjork, 2012). Of interest was students’ ability to correct errors made on the multiple-choice test on a later general-knowledge test. Both verification and answer feedback increased the retention of correct responses, as compared to no feedback. However, when it came to the correction of errors, both types of feedback yielded improvement, but answer feedback was much more beneficial than verification feedback. In short, verification feedback may be more useful for multiple-choice tests than for short-answer tests, but the greatest benefits are still obtained with answer feedback.

**Conclusions and Future Directions**

- Our belief is that testing is a powerful tool for the educator, which can be used to promote learning and retention in students. Multiple-choice tests may be particularly appealing given the existence of test banks and the relative ease of grading them. The point of the present chapter was to point out that multiple-choice tests should be used with care, since they do have the potential to teach students wrong information. This problem may be particularly challenging when tests are hard and/or students are underprepared. Fortunately for the educator, there is a simple solution: provide the students with the correct answers after taking the test. When thinking about implementing this tool in the classroom, it is important to keep in mind that there are many different ways to do so. That is, the instructor could give low-stakes multiple-choice quizzes, use clickers to collect in-class responses, or assign multiple-choice questions for homework using a classroom management platform such as Sakai. Multiple-choice questions do not need to be embedded in high-stakes tests or exams to yield benefits to the learner; regardless, the key is to provide learners feedback on their answers.

- It is important to note that many of the experiments discussed in this chapter were conducted in the laboratory. A burgeoning area of research involves extending such laboratory findings to the classroom. Laboratory studies are certainly valuable—they demonstrate basic cognitive phenomena and allow us to more clearly determine mechanism. Additionally, many laboratory studies use the same materials as those utilized in the classroom, increasing the similarities between the lab and the classroom. However, it is crucial to examine these issues in real classrooms, where there may be
much variability in students, longer retention intervals, and limited time to master the material. To date the most impressive attempt to generalize the research is a project called the Columbia Middle School project, where 1,400 middle school students participated over a five-year period (see Agarwal, Bain, & Chamberlain, 2012, for a review). Consider a prototypical experiment: students studied science textbook chapters; the classroom teacher gave lessons; students took multiple-choice quizzes using clickers; and retention was measured on exams that were weeks or months later. Overall, there were long-term benefits of the low-stakes multiple-choice quizzing, even when the students quizzed themselves at home (Roediger, Agarwal, McDaniel, & McDermott, 2011). The results held with history materials (Roediger et al., 2011) and with science materials (McDaniel, Agarwal, Huelsner, McDermott, & Roediger, 2011), and benefits were observed as much as eight months later (McDaniel et al., 2011).

- Much of the concern about testing in the classroom is that it leads to "teaching to the test," which may promote relatively rote learning of tested concepts and may be at the expense of material that is not tested (see Frederiks, 1984, for a discussion of some of these concerns). It is a very fair concern that much of the laboratory work on testing has focused on relatively simple materials (that are easily tested) and retention rather than complex materials and higher-level cognition. However, most of the classroom studies do not use simple materials—they use real classroom concepts, quizzes, and exams. One Columbia Middle School teacher noted that one advantage of multiple-choice quizzing as a learning activity was that it did not require her to change her teaching style. And more importantly, she reflected that implementing retrieval practice helped her understand what makes students succeed in the classroom (Agarwal et al., 2012, p. 445).

- In closing, we would like to point out that the early evidence is promising: tests can promote transfer of learning and performance on inference questions (e.g., Butler, 2010; Little et al., 2012; Rohrer, Taylor, & Sholar, 2010) and can promote learning of complex material like functions and skills like CPR (e.g., Kang, McDaniel, & Pashler, 2011; Kromann, Bohnstedt, Jensen, & Ringsted, 2010; Larsen, Butler, & Roediger, 2008). Initial classroom evidence finds that multiple-choice tests also promote transfer of learning (McDaniel, Thomas, Agarwal, McDermott, & Roediger, 2013). An exciting direction for future research is further investigating the conditions where multiple-choice testing promotes not only retention but also a deeper understanding of material.

References


