CHAPTER 38

Test-enhanced learning

Douglas P. Larsen and Andrew C. Butler

The active recall of a fact from within is, as a rule, better than its impressions without.

Edward Thorndike

Introduction

A chief resident faced a difficult educational task. The residents in the programme did not feel that they adequately understood, recognized, or knew how to treat rare metabolic diseases. The chief resident collaborated with a faculty member who had special expertise in this area to create two 1-hour didactic conferences that reviewed the diseases. The faculty member provided an extensive overview and many case examples. After the conferences, the residents felt that they had learned the material, and they were grateful that their educational need had been met. No further formal exposure to the material occurred.

This example illustrates a fundamental misconception about learning that is ubiquitous in medical education—the assumption that performance during learning (or immediately afterwards) will be maintained. Both objective assessments (e.g. tests) and subjective judgements (e.g. feelings of mastery) during learning are often poor predictors of long-term retention because they reflect the accessibility of knowledge at a given moment rather than how well that knowledge has been stored in memory (Bjork and Bjork 1992). This misconception undermines the critical educational objective of helping clinicians to acquire and retain the large body of medical knowledge that they will need to apply in the future.

With this misconception in mind, educators must consider how likely it is that the residents in the real-life scenario above remembered the material that was taught in the conferences. Based on what we know from cognitive science, the answer is probably little or nothing at all, which is troubling because much of medical education occurs through similar methods and settings. For example, residents at most teaching hospitals in the United States spend at least 8 hours a week in formal didactic conferences. In fact, in the United States, the Accreditation Council on Graduate Medical Education (ACGME) mandates these conferences (ACGME 2011, p. 7). Yet studies have shown no difference in knowledge between clinicians who attended such conferences and peers who did not (Cacamese et al. 2004; FitzGerald and Wenger 2003; Picciano et al. 2003; Winter et al. 2007). The same problem exists at all levels of medical education. Students spend countless hours in classrooms prior to their clinical years and then spend several hours a week in didactic sessions during their clinical rotations. Practising physicians are required by regulatory agencies to spend a certain number of hours each year at continuing medical education conferences.

The challenge for medical education is to develop and implement learning methods that produce long-term retention of knowledge that can be flexibly recalled and applied in the future. This chapter reviews one such method, called *test-enhanced learning* (Roediger and Karpicke 2006a, Larsen et al. 2008, Roediger and Butler 2011) Test-enhanced learning is based on the finding that retrieving information from memory produces superior long-term retention, commonly referred to as the *testing effect*. Although practising retrieval of information is often implemented as a test, it can take many forms and is not limited to traditional paper or electronic tests. A large body of research in cognitive science and related fields has shown the testing effect to be a robust and replicable finding. In fact, the evidence is so strong that the Institute of Education Sciences from the Department of Education in the US has recommended using retrieval practice to promote retention at all levels of education (Pashler et al. 2007).

The goals of this chapter are to introduce the idea of test-enhanced learning, review the main findings in the literature, and provide some guidance as to how test-enhanced learning might be implemented in medical education.

Overview of test-enhanced learning

In education, tests are typically synonymous with assessment—most educators and students consider testing a tool for assessing student learning and providing feedback to guide future activities (Black and William 1998). As conceptualized within test-enhanced learning, testing has a different purpose: to directly increase retention and understanding by the act of taking the test. The memory retrieval that occurs while taking a test is often thought to be a neutral event—similar to measuring someone's weight. Much like stepping on scale does not change a person's weight, memory retrieval during a test is assumed to sample one's knowledge but leave it unchanged. Research in cognitive science indicates that this assumption is false; rather, the act of retrieving information from memory actually changes memory (Bjork 1975), leading to superior retention over time and better understanding (Roediger

and Butler 2011; Roediger and Karpicke 2006a). Although it is difficult to divorce testing from assessment, the mnemonic benefits of retrieval practice suggest that testing is a powerful learning tool.

A brief history of testing effect research

The idea that practising memory retrieval promotes long-term retention dates back many centuries. Consider the following statement: 'Exercise in repeatedly recalling a thing strengthens the memory'. Although this quotation sounds as though it could be part of this chapter, it is actually from Aristotle's classic treatise on memory—*De Memoria et Reminiscentia* (Hammond 1902, p. 202). The first empirical demonstration of the mnemonic benefits of testing in a controlled experiment occurred just over a hundred years ago (Abbott 1909). Over the next 30 years, educational psychologists became interested in applying this phenomenon to the classroom (Gates 1917; Jones 1923–1924; Spitzer 1939). However, interest dwindled in the second half of the 20th century and testing effect research became sporadic despite the publication of many important studies (Carrier and Pashler 1992; Glover 1989; Tulving 1967; Wheeler and Roediger 1992). More recently, there has been a resurgence of interest in the phenomenon (Roediger and Butler 2011; Roediger and Karpicke 2006a).

Robustness and replicability

The findings in recent studies have firmly established that the phenomenon is robust and replicable. One powerful example comes from a recent study by Karpicke and Roediger (2008, pp. 966–968) in which they examined various methods for learning foreign vocabulary with flash cards. They gave undergraduate students Swahili–English word pairs (e.g. *mashua–boat*) to learn through repeatedly studying and testing the pairs until each pair had been successfully recalled once. After a pair had been recalled, it was assigned to one of four types of additional practice: (1) repeated studying and testing, (2) repeated studying only, (3) repeated testing only, and (4) no further activity. One week later, the students were given a final cued recall test in which they had to recall the English translations when prompted with the Swahili words. The results were striking; repeated testing only produced a much higher level of correct recall relative to repeated studying only and no further activity. Interestingly, additional study had little or no effect on retention—repeated studying and testing did not improve correct recall relative to repeated testing only, and the repeated study only was marginally better than no further activity.

The rapid accumulation of studies has allowed researchers to quantify the effect of retrieval practice (Bangert-Drowns, Kulik and Kulik 1991; Phelps 2012; Rawson and Dunlosky 2011). Phelps (2012, pp. 21–43) recently conducted a meta-analysis that included several hundred studies conducted over the past 100 years. He found that the mean effect size related to testing was either moderate ($d = 0.55$) or large ($d = 0.88$), depending on how the effect size was calculated. When testing was more frequent and post-test feedback was provided, the effect of testing on achievement was even larger.

Most studies on the testing effect have sought to examine the benefits of retrieval practice by comparing it to a restudy control condition (Butler and Roediger 2007; Carrier and Pashler 1992; Glover 1989). Restudy is an ideal comparison activity because it usually involves reprocessing all of the to-be-learned material (whereas testing involves reprocessing only what can be recalled) and it is common in education (Karpicke et al. 2009). However, one possible criticism is that restudy could be considered a more passive task than testing. With this criticism in mind, several recent studies have compared retrieval practice to other more active learning strategies, such as note-taking, concept-mapping, self-explanation, and various mnemonic techniques (Fritz et al. 2007a; Karpicke and Blunt 2011; Larsen et al. 2013; McDaniel et al. 2009). All of these studies have found benefits of testing relative to these other learning strategies.

A host of studies have also demonstrated the durability of testing effects. Although many studies have used relatively short retention intervals ranging from minutes to a few days, other studies have examined retention over much longer intervals. These studies have found reliable benefits of testing after periods ranging from several weeks (Butler and Roediger 2007; Kromann et al. 2009; Rawson and Dunlosky 2011) to more than 6 months (Carpenter et al. 2009; Larsen et al. 2009; 2012, 2013; McDaniel et al. 2011). In fact, one small study demonstrated that the benefits of retrieval can last up to 5 years (Bahrick et al. 1993).

Generalizability

The generalizability of test-enhanced learning is also well established with respect to several important variables: learners, materials, and performance measures. Concerning learners, the testing effect has been obtained in many different demographics that have a wide variety of characteristics and abilities. The mnemonic benefits of retrieval practice have been demonstrated across the age spectrum from young children (Fritz et al. 2007b) to older adults (Tse et al. 2010). The testing effect has also been observed with medical students (Kromann et al. 2009; Larsen et al. 2012; Rees 1986) and medical residents (Larsen et al., 2009). In terms of differences in student knowledge and ability, testing seems to benefit learners regardless of their level of prior knowledge (Carroll et al. 2007) or their memory ability and intelligence (Brewer and Unsworth 2012); however, there is some indication that the magnitude of the testing effect may be reduced in individuals with greater prior knowledge, memory ability, and/or intelligence.

The testing effect also generalizes across many types of materials. Traditional laboratory studies of retrieval practice have often used simple materials, such as word pairs (Karpicke and Roediger 2008) or general knowledge facts (Butler et al. 2008). However, the benefits of testing have been shown to extend to a variety of more complex materials. For example, studies have found testing effects using texts (Kang et al. 2007), lectures (Butler and Roediger 2007), multimedia presentations (Johnson and Mayer 2009), and maps (Carpenter and Pashler, 2007). The benefits of testing also seem to extend to inductive function learning (Kang et al. 2011), identifying bird species (Jacoby et al. 2010), and various skills like resuscitation (Kromann et al. 2009). In addition, it is important to note that the phenomenon seems to transcend knowledge domains, having been observed with materials from a variety of disciplines such as history (Carpenter et al. 2009), science (McDaniel et al. 2007), and medicine (Larsen et al. 2009).

Finally, one other variable that is critical to generalizability is the outcome measure used to assess the benefits of retrieval practice. Most testing effect studies have used a final assessment that is an exact repetition of the same test that was given during learning. In recent years, researchers have begun to explore whether the effects of testing extend beyond the retention of information to the understanding and use of that information. Transfer of knowledge involves applying previously learned information to a new context

(Barnett and Ceci 2002), an important outcome for educational purposes. Many testing effect studies have shown that practising retrieval improves transfer of knowledge (Butler 2010; Johnson and Mayer 2009; Karpicke and Blunt 2011; Larsen et al. 2012; McDaniel et al. 2009). Overall, these studies suggest that testing can improve both the retention and understanding of material, enabling the application of knowledge to a variety of contexts.

Theoretical mechanisms

When discussing theoretical explanations for the mnemonic benefits of retrieval practice, it is important to distinguish between the direct and indirect effects of testing. Direct effects of testing refer to the improved retention and understanding that result from the act of successfully retrieving information from memory (i.e. the focus of this chapter). In contrast, the indirect effects of testing refer to a host of other ways in which testing can influence learning. For example, testing can help students to assess what they know and do not know, providing valuable feedback that they can use to guide future study. In addition, frequent testing can motivate students to study and attend class (Fitch et al. 1951; Mawhinney et al. 1971), helping them to avoid putting off studying until the last minute (Michael 1991).

One of the first formal hypotheses put forth to explain the testing effect focused on differences in the amount of exposure to the material. In many early testing effect studies, the experimental group would study material and then take a test, while the control group would simply study the material, and then both groups would take a final test to measure retention. Based on this comparison, some researchers pointed out that the testing group received two exposures to the material (they were re-exposed to the material that they retrieved on the test), and this difference may be driving the effect (Thompson et al. 1978; Slamecka and Katsaiti 1988). However, many subsequent studies have disproven this idea by showing that testing still produces a benefit when the control group has the opportunity to re-study the material and total exposure to the material is matched (Carrier and Pashler 1992; Glover 1989; Karpicke and Roediger 2008).

Other theories that have attempted to explain the testing effect have focused on how the act of retrieval affects memory, and these theories can be categorized into two groups. One group of theories revolves around the idea that the mnemonic benefits of testing result from the reprocessing of the material that occurs during retrieval (Carpenter 2009; Pyc and Rawson 2009). When a memory is retrieved, the memory trace is elaborated and new retrieval routes are created, making it more likely it will be successfully retrieved again in the future. The amount of effort that is involved in retrieving information is considered to be an index of the amount of reprocessing that occurs; this notion of retrieval effort helps to explain why production tests (e.g. short answer tests), which require more effort, tend to produce better retention than recognition tests (e.g. multiple choice questions), which require less effort (Butler and Roediger 2007; Kang et al. 2007).

A second group of theories centres on the relationship between initial learning and the final test, invoking a principle called transfer-appropriate processing (Morris et al. 1977; Roediger et al. 2002). Transfer-appropriate processing posits that memory performance is enhanced when the cognitive processes that are engaged during learning match the processes that are required during retrieval. With respect to the testing effect, this principle applies because the cognitive processes engaged while taking an initial test provide a better match for the final retention test relative to the processes engaged while restudying the material (the traditional control condition). When considering how integral memory retrieval is to the application of knowledge outside the classroom, the principle of transfer-appropriate processing suggests that students should be engaging in activities during learning that provide retrieval practice.

Overall, there is ample evidence to support both groups of theories, and it is important to note that they are not mutually exclusive. Further development of theory is ongoing and many researchers are now concentrating on gaining a better understanding of the underlying mechanisms that produce the mnemonic benefits of retrieval practice. In the future, we expect that these psychological theories will be enriched by new evidence and ideas from cognitive neuroscience that specify possible brain mechanisms (Roediger and Butler 2011).

Finally, it is important to briefly touch on a few of the theories that have been put forth to explain some of the indirect effects of testing. A full review of these theories is beyond the scope of this chapter. However, one important category of theories about the indirect effects of testing focuses on metacognition. Agrawal et al. (2012, pp. 326–335) have argued that the act of taking a test and reviewing feedback stimulates self-monitoring to identify areas of unexpected results. This rehearsal influences subsequent studying behaviour, thereby facilitating further learning and long-term retention (Kulhavy and Stock 1989). Similarly, Pyc and Rawson (2010, p. 335; 2012, pp. 737–746) have suggested that testing enables learners to discover whether the strategy that they used to encode the to-be-remembered information was effective—an idea that they refer to as the mediator effectiveness hypothesis. When learners take a test and fail to retrieve a piece of information, they can subsequently use a different strategy to encode the information. Finally, Butler et al. (2008, pp. 918–928) have shown that receiving feedback after a test can help to improve metacognition by making learners better able to distinguish between test responses that are correct and incorrect. They argue that feedback is important for low-confidence correct responses because it helps the learner to correct a metacognitive error (i.e. thinking that a response is incorrect when it is actually correct).

Implementing test-enhanced learning in medical education

The following section discusses some of the factors that influence the efficacy of retrieval practice, while also offering practical recommendations for using test-enhanced learning in the classroom and clinic (fig. 38.1).

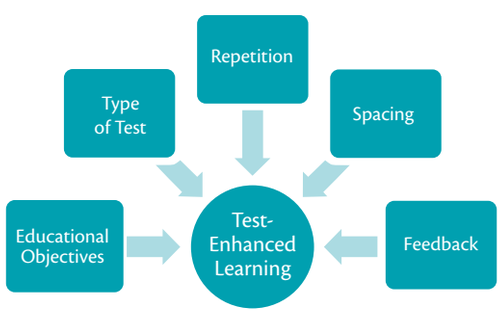

Figure 38.1 Factors that influence test-enhanced learning.

Aligning retrieval practice with educational objectives

When considering the implementation of test-enhanced learning, it is important to recognize that the principle of retrieval practice can be broadly applied to a variety of activities beyond simple written tests. For example, retrieval practice occurs when students answer questions orally, attempt to diagnose a patient, or perform a surgical technique. The key aspect of the activity is that the information, procedure, or skill is retrieved from memory. Although educators have a wide range of possible activities from which to choose, the form of retrieval practice must be tailored to the educational objectives in order for learning to be optimized. As with all good educational planning, educators must ask themselves: 'What do I want my students *to know*? What do I want them *to be able to do*?' These questions help to identify the type of learning that is needed in a given situation. Learning may focus on declarative facts, concepts (grouping and categorization of facts), principles (rules that determine how facts are applied), problem-solving (principles that lead to the solution of novel situations), or psychomotor tasks (Smith and Ragan 2005, pp. 78–82). Once the type of learning is identified, the educator must match the form of retrieval practice to the desired type of learning.

Learning facts is often derided as a 'lower' form of learning. This position neglects the reality that much of the practice of medicine is based on the knowledge of facts. For example, physicians must learn the characteristics of diseases, drug dosing and side effects, and what constitutes 'normal' and 'abnormal' on a clinical test. Much of the research on test-enhanced learning has focused on learning and retaining factual knowledge. For instance, Larsen et al. (2009, pp. 1174–1181) investigated the effects of three short-answer written tests at 2-week intervals after a didactic conference covering the diagnosis and treatment of two different neurological conditions. They found that repeated testing led to better retention of these facts after 6 months when compared to repeated studying. Similarly, Turner et al. (2011, pp. 731–737) demonstrated that after a life support course, four unannounced oral tests given over the telephone significantly improved the retention of factual knowledge at 2 months compared to a control group that only received a single oral test. The oral tests were given without feedback so that the groups only differed in their amounts of retrieval practice. In practical terms, educators must have clear idea of which facts are foundational and applicable to a clinical learning objective and then make sure that learners have opportunities to retrieve these facts from memory.

Concept learning is another critical aspect of medical education. One of the main cognitive tasks involved in making a medical diagnosis is to correctly categorize the symptoms and signs of a patient's illness. This process of diagnosis is based on the similarities and differences that the patient possesses with different disease categories based on 'illness scripts' that have developed as the practitioner's mental representation of a distinct disease (Schmidt and Rikers 2007). If learners are to distinguish between similar diseases or identify how similar symptoms may indicate divergent diagnoses, then they must have the opportunity to practise sorting these concepts and identifying how to apply them to a given case (Smith and Ragan 2005, p. 178).

One paradigm for studying the effects of testing on concept learning in the cognitive psychology literature uses bird species identification (Jacoby et al. 2010). In studies in this field, testing was shown to improve the recognition of both studied and novel exemplars of bird species, and the classification of these exemplars as well. Testing is thought to improve learning in these cases by enabling students to practise the identification of key details that distinguished birds from each other or characteristics that allowed them to be grouped together. These findings are important because they demonstrate that testing can improve the ability of subjects to identify the relationships between key elements of knowledge—a key characteristic of 'deeper' levels of learning (Marton and Saljo 1976).

Jacoby et al. (2010, pp. 1441–1451) also demonstrated the effects of testing on subjects' awareness of their levels of knowledge (i.e. metacognition). They found an increase in the ability of subjects to determine how well they had learned the material and to predict which categorization tasks would be more difficult than others (known as classification judgement learning). These findings closely complement the work described earlier by Agrawal et al. (2012, pp. 326–335), regarding the effects of testing on enhancing self-monitoring and also the mediator shift-hypothesis developed by Pyc and Rawson (2010, pp. 335; 2012, pp. 737–746). Improved ability to predict classification difficulty has particularly important implications for both education and clinical practice. If learners are able to predict which concepts are difficult to classify, they can direct further study towards those topics. In clinical medicine, practitioners would be more aware of when a particular diagnosis can be difficult, and therefore would take more care in order to avoid errors.

In terms of practical application in medical education, concepts could be tested through clinical case scenarios. Learners must be exposed to a sufficient number of cases to be able to learn to make the distinctions between similarities and differences. Learners should see examples that fit in the category and counter-examples that do not fit (Smith and Ragan 2005, pp. 176–178). Too often in case-based learning only a single case is presented, which is unlikely to allow learners to develop a clear set of rules to be applied to future cases. Learners need repeated retrieval attempts to form and verify mental rules regarding the relationships between the pieces of information that they have learned.

Testing that allows learners to practise application can facilitate the application of knowledge. Although much less research has been directed at how testing can be used to achieve these educational objectives, some studies have begun to investigate whether retrieval practice can facilitate transfer of learning (Butler 2010; Johnson and Mayer 2009; Karpicke and Blunt 2011; Larsen et al. 2012; McDaniel et al. 2009). For example, Butler (2010, pp. 1118–1133) demonstrated that repeated retrieval practice on facts and concepts improved the ability to apply knowledge to novel situations. After studying scientific texts, learners were asked questions that required them to retrieve facts and concepts from the texts. Performance on a final application test 1 week later demonstrated improved transfer of learning for facts and concepts that were tested compared to facts and concepts that were repeatedly studied. Importantly, an additional experiment showed that learners who engaged in repeated testing were better able to apply concepts that they had learned to novel situations in an unrelated knowledge domain.

In addition to the purely cognitive domains of learning, testing has been shown to produce increased retention in the area of

psychomotor skills. One of the first studies to examine test-enhanced learning in the medical education literature investigated the effects of testing on cardiopulmonary resuscitation. Kromann et al. (2009, pp. 21–27) found that a single test at the end of a cardiac resuscitation course improved retention by almost 10% at 2 weeks compared to students who had received the traditional training. Follow-up at 6 months continued to show an effect on retention with a clinically relevant effect size ($d = 0.40$) (Kromann et al. 2010). Another example with real-life application is the work done by Wayne et al. (2006a, pp. 251–256). The researchers demonstrated that simulation-based repeated retrieval practice by internal medicine residents led to mastery of advanced cardiac life support protocols. This mastery level was maintained without significant decrement for at least 14 months (Wayne et al. 2006b). Importantly, real-life performance in cardiac resuscitation was superior for residents trained with simulation-based deliberate practice compared to non-simulation trained residents (Wayne et al. 2008).

As some of our examples demonstrate, retrieval practice in medical education can have a direct impact on the care that patients receive. Educators should plan for and carry out retrieval practice based on specific educational objectives. Figure 38.2 shows some examples.

Type of test

Educators who implement test-enhanced learning must use the test format that will have the greatest impact. Tests can be separated into two categories: production tests and recognition tests (fig. 38.3). Production tests, such as short answer and essay tests, involve generating a response from memory. In contrast, recognition tests, such as multiple-choice and true–false tests, involve selecting a response from information that is provided. Both types of test have been shown to improve retention (McDaniel et al. 2011). However, production tests generally produce better long-term retention than recognition tests (Glover 1989; Kang et al. 2007). Butler and Roediger (2007, pp. 604–618) gave students either a short-answer test or a multiple-choice test after watching a videotaped lecture. When retention was measured on a final test 1 month later, the initial short-answer test produced better performance than the initial multiple-choice test.

The superior retention that results from production tests can be explained by the idea of retrieval effort. That is, production tests tend to require considerable mental effort to generate the information, whereas recognition tests involve simply selecting the correct information. One form of production test that requires substantial effort is the free recall test. Free recall tests require learners to generate information with relatively few or no cues. For example, a free recall test might be given to students asking them to name all of the nerves in the human body. One reason that free recall tests are particularly effective is that learners have to retrieve the organization of the information as well as the individual items. As a result, free recall tests can induce learners to create better organizational structures of knowledge (Zaromb and Roediger, 2010).

Although studies in the cognitive psychology laboratory often use relatively simple free recall tests (e.g. essays), free recall can be implemented in many more complex ways that directly correspond to processing that educators target in medical education. One such example is the use of mechanical simulation in psychomotor skills learning (Kromann et al. 2009; 2010; Wayne et al. 2006b; 2008). During manikin simulation, learners must retrieve and apply their knowledge with few or no explicit cues. Of course, the various symptoms and signs manifested by the manikin clearly provide some implicit cues, but retrieval is largely self-directed because the learner must remember both the information (e.g. the specific steps to take during a medical intervention) and the organization of that information (e.g. the correct ordering of the steps).

Another way of implementing free recall in medical education is through standardized patient encounters (fig. 38.4). In a study by Larsen et al. (2012), students learned the necessary information to diagnose and treat patients with three neurological conditions in a

Knowledge of declarative facts
Concepts used to categorize groups of facts
Transfer of learning to new contexts
Principles used in problem-solving
Interactions with patients in live encounters
Procedural skills learned on simulation models with anticipation of real-life application

Figure 38.2 Examples of educational objectives that retrieval practice can help to achieve.

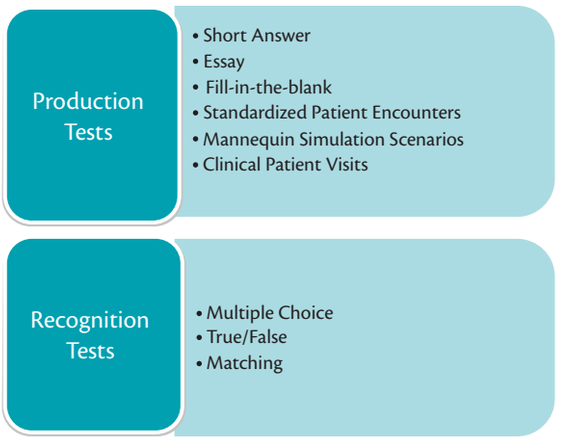

Figure 38.3 Examples of production and recognition tests.

Figure 38.4 Patient encounters (both real and simulated) should be considered retrieval practice opportunities that should be incorporated into test-enhanced learning.

teaching session. Next, they performed one of three learning activities for each of the three topics:

- take a written short-answer test
- see a standardized patient, or
- study a review sheet.

Each activity covered identical information. Assignment of a topic to activity was randomized. Students performed the activity assigned to each topic four times at one-week intervals. Six months after initial learning students took a final test that consisted of seeing a standardized patient for each of the three topics. One week later they completed a final written short-answer test on all three topics. On average for the standardized patient final test, the students who learned their particular topic through seeing standardized patients performed significantly better (59%) than students who had learned the same topic through written testing (49%) or studying a review sheet (43%). Interestingly, on the final written test, students who learned the topic through standardized patients and through written tests performed equivalently (both 61% retention on average) and better than students who learned the topic through studying a review sheet (48%). One possible explanation for the difference in the pattern of results across the two types of test is that the written test provided more cues for the students relative to the standardized patient test. That is, learning by seeing a standardized patient was essentially like taking a free recall test—students had to retrieve both the information itself and the organization for the information. In contrast, learning by taking a written test led students to be more dependent on the cues provided by the test. Thus, when students had not had practice retrieving the organization of the material, they had more difficulty in the final standardized patient test relative to the final written test in which cues were provided. This finding highlights the need to make sure that the type of retrieval that is practised during learning is a good match for the way in which the information will need to be retrieved and used in the future.

Patient encounters (simulated or real) should be planned for as retrieval practice opportunities. In some cases, the structured practice and feedback afforded by simulated patients may produce superior results, even compared to actual patient encounters. For example, Safdieh et al. (2011, p. 5634) examined performance of second year students on a neurological exam by comparing a group that received their school's standard curriculum of small groups and lectures and a group that received the standard curriculum plus a single standardized patient session dedicated to practising the neurological exam. The final outcome measure was based on an OSCE in which students performed a neurological exam with a standardized patient. Students who practised the exam with a standardized patient demonstrated superior performance compared to students who had received the standard curriculum—a durable effect that persisted over 2 years. Interestingly, the intervention group even outperformed students in the control group who had completed their neurology clerkship, during which they would have had repeated opportunities to practice their neurological exam on actual patients.

The differential effectiveness of instructor-generated practice relative to student-generated practice is also suggested by findings of Larsen et al. (2012). In this study, 71% of students reported studying the review sheets by self-quizzing. However, despite these efforts, the study group performed worse than both the written testing and standardized patient groups. While the differences between instructor-generated and student-generated testing needs to be more thoroughly explored in future studies, there are several reasons to suspect that student-generated testing may not be as beneficial as instructor-generated testing. First, student-generated testing often occurs immediately after the student is exposed to the material. For example, students might read a passage, cover it up, and then try to recall what was just read. Or they might listen to a lecture on the physical exam and then practise it right away in a small group. Recall is relatively easy when attempted immediately after study—the resulting high level of performance can inflate students' judgements of learning and generate an illusion of competence (Bjork 1994). However, the level of performance immediately after learning is a poor indicator of future retention.

Another potential problem is that when the principle of retrieval effort is considered, immediate retrieval from working or short-term memory is much easier than retrieval from long-term memory (Bjork 1994). Thus, the more difficult recall engendered by an instructor-generated test would be expected to produce greater retention. Another point that may influence the efficacy of student-generated versus instructor-generated testing is the fact that learners will often stop quizzing themselves once they have successfully recalled an item (Kornell and Bjork 2008). Karpicke and Roediger (2007, pp. 151–162) demonstrated that repeated retrieval after a successful initial recall event produces much better retention. Repeated instructor-generated tests may force a learner to continue to practise retrieval after they would have stopped on their own.

Overall, a review of the findings regarding the type of test indicates that once educators have identified their educational objectives, they must think broadly about the types of retrieval practice that will best help them achieve these objectives. Tests should be designed to require the generation of information rather than the recognition of information. In addition, tests should be designed to approximate the settings in which the learning will be applied in the future. Mechanical simulation and simulated patient encounters appear to provide increased retention that surpasses written testing when considering eventual clinical application. However, it is important to note that it may be the type of test (free recall versus cued recall) that is driving the superiority of simulation testing. Although self-testing is better for students than simply studying material by re-reading, instructors must recognize that incorporating retrieval practice opportunities as part of the formal curriculum may produce better results than solely relying on self-testing.

Repetition

Educators must also think about the frequency with which tests are given to students. Although a single test is better than no test, repeated testing will produce greater long-term retention. Several of the examples discussed above show how single tests (especially in the psychomotor domain) have lasting effects. Kromann et al. (2010, pp. 395–401) demonstrated that a single test with cardiac resuscitation led to improved performance with a clinically relevant effect size even after 6 months. The study by Safdieh et al. (2011, p. 5634) showed that a single test with standardized patients produced superior neurological exam performance approximately two years later. Clearly, a single test is effective.

Nevertheless, a multitude of studies have found that even higher levels of retention are possible with repeated retrieval practice. For

example, retention on final recall generally improves as the number of successful retrievals of the information increases (Karpicke and Roediger, 2007; Wheeler and Roediger 1992). Item analyses in the study by Larsen et al. (2012) have also found that a higher number of successful retrieval events was associated with a greater likelihood of retention 6 months after initial learning. When considered with regard to the principle of repetition, the study by Karpicke and Roediger (2008, pp. 966–968) described in the first part of the chapter is particularly instructive. Note that in this experiment students learned all word pairs sufficiently well to be recalled at least once—essentially a single test. As the results clearly show, additional study after successful retrieval produced no benefit, yet repeated testing after the first successful retrieval generated superior retention.

Repeated retrieval practice is embedded within the concept of deliberate practice—the idea that deliberate effort to improve performance in a specific domain is critical to becoming an expert in that domain. Deliberate practice has emerged from the simulation literature as a key component of successful simulation (Issenberg et al. 2005). Indeed, the word *practice* implies repeated effort. Deliberate practice includes well-defined learning objectives, which leads to repeated practice with clear outcome measures (McGaghie et al. 2011), and it forms an iterative process of feedback and monitoring that leads to further practice until mastery is reached. In the Best Evidence Medical Education (BEME) review of simulation-based education (Issenberg et al. 2005), deliberate practice was found to be a key element that leads to improvement in patient care. In a meta-analysis of simulation-based medical education compared with traditional curricula, McGaghie et al. (2011, pp. 706–711) demonstrated a combined effect size with a correlation coefficient of 0.71 in favour of improved skill learning through deliberate practice using simulation compared to traditional curricula. Despite the clear benefits of deliberate practice, it is not universally applied. A survey of all anesthesia residents in Canada found that while 94% of residencies used high-fidelity manikin simulation in training, 81% of residents reported not utilizing repeated practice of the simulation scenarios (Price et al. 2010).

Unfortunately, repetition (let alone repeated retrieval practice) is rarely planned into medical education curricula. The importance of repetition becomes apparent when one considers the trajectory of forgetting that naturally occurs once information is learned. Ebbinghaus, the 19th century psychologist, was the first to describe the forgetting curve in which large amounts of forgetting occur quickly, followed by a more slow and steady decline in retention (Ebbinghaus 1967/1885). Ebbinghaus' finding has been confirmed by countless studies over the years, and it is illustrated by the study by Larsen et al. (2009, pp. 959–966), who investigated the effects of repeated tests at 2 week intervals on long-term retention of information that resident physicians learned in a didactic conference. In this study, retention dropped an average of 24% after two weeks between initial learning and a follow-up test with feedback. A third test 2 weeks later showed no further decline—rather, there was a slight increase in performance. Six months after initial learning, performance on the final test declined only slightly compared to the third test. Thus, these results indicate that residents did initially forget some of the information, but testing helped to mitigate forgetting.

Overall, the findings reviewed in this section indicate that repeated testing helps to promote even better long-term retention than a single test. Repeated retrieval practice coupled with feedback maintains initial learning while also fostering further learning, thus resulting in even higher levels of performance (Karpicke and Roediger 2007; Larsen et al 2012a, b). Repetition also allows the learner to take full advantage of feedback, and to practise to correct errors.

Spacing

The principle of repetition is linked with the concept of spacing or distributing practice over time. An extensive body of literature has demonstrated that spaced practice improves retention of information and motor skills compared to massed practice (Cepeda et al. 2006; Dempster 1989). Spacing is beneficial when implemented within a single learning session (Pyc and Rawson 2009), across multiple sessions compared to a single session (Rohrer and Taylor 2006), and using longer intervals relative to shorter intervals between sessions (Carpenter et al. 2009). Unfortunately, there is no easy recommendation regarding the optimal interval between practice attempts because it seems to depend on the interval over which the information must be retained. Recent meta-analyses indicate that a spacing interval that is 10–20% of the retention interval maximizes retention (Cepeda et al., 2006; 2008).

An important study by Cepeda et al. (2008, pp. 1095–1102) demonstrates the importance of adequate spacing during learning. Subjects were trained to recall 32 disparate trivia facts. Next, they were randomized to receive a second learning session at intervals of from 0 to 105 days. In the second learning session, they were asked to retrieve the facts two times, each with feedback. Subjects were randomized to a final recall test at intervals of 7, 35, 70, and 350 days after the second learning session. The results showed an interaction between the practice interval (i.e. the delay between initial and second learning sessions) and the retention interval (i.e. the delay between the last practice session and final recall). The effect of spacing formed an asymmetric U function. For all intervals, final test performance (i.e. retention) initially improved as the practice interval increased. However, this benefit began to slowly decrease after a point, which was different for each retention interval. Thus, the point of maximal retention that occurred right before the effect began to decline marks the point of optimal spacing (the top of the upside-down U).

Cepeda et al. (2008, pp. 1095–1102) found that a ratio of 10–20% between the practice interval and the retention interval maximized retention. For retention of 7, 35, 70, and 350 days, the optimal spacing was found to be 1, 11, 21, and 21 days, respectively. Although it is unclear if these exact numbers would be equally applicable to all types of education, the principles demonstrated in the study have important practical applications. If educators want learners to retain information for long periods of time (months to years), then they must space practice over weeks and months.

The effects of spacing have been demonstrated in medical education literature also. Using online multiple choice questions covering core topics in urology delivered by email to urology residents, Kerfoot (2009, pp. 2671–2673) demonstrated that spaced learning improved retention over a two-year period compared to massed learning. In another study, Schmidmaier et al. (2011, pp. 1101–1110) investigated a test-enhanced learning paradigm with students using four back-to-back cycles of short answer testing using electronic flashcards covering topics in clinical nephrology. Students who used the repeated testing performed significantly better on a cued-recall test 1 week after initial learning, compared to students who had simply restudied the material. However, there was no difference between the groups at 6 months. These findings stand in contrast to other medical

education studies that have found significant differences between testing and control groups at intervals of 2 to 6 months (Larsen et al. 2009; 2012; 2013; Turner et al. 2011). The major difference between these studies, which showed a long-term improvement of retention, and the study Schmidmaier et al. (2011, pp. 1101–1110) conducted was the interval over which tests were spaced during learning. The studies that showed effects over long retrieval intervals used intervals of 1–2 weeks, whereas Schmidmaier et al. (2011, pp. 1101–1110) used much shorter intervals. The differential outcomes of these studies illustrates the findings of Cepeda et al. (2008, pp. 1095–1102)—longer spacing intervals during learning are critical to promoting retention over longer retention intervals.

The study by Larsen et al. (2009, pp. 1174–1181) used 2-week testing intervals and saw a dramatic drop in retention within the first 2 weeks. Subsequent studies by the same group (Larsen et al. 2012a, 2013) used 1-week testing intervals and increased the number of testing events from three to four. The more recent studies did not show the dramatic drop in performance during initial learning that was found with the two-week interval. By the end of the initial learning phase in these two newer studies, performance was greater than or equal to performance immediately after learning. The end result was better long-term retention on the final test (albeit comparing across studies with several other differences).

Although these studies did not directly compare different testing intervals, they still illustrate an important practical point. Optimal spacing is a delicate balance—the next test should be delayed long enough to make it effortful and promote retention, but not so long that the information will be forgotten. Different types of memories (i.e. procedural skills versus factual knowledge) may be forgotten at different rates, and therefore the optimal practice interval is likely to differ depending on what students need to learn.

Feedback

For repetition and spacing to have maximal impact on learning from tests, feedback must be provided. Feedback is critical because it helps the learner to close the gap between actual and desired learning (Bangert-Drowns et al. 1991; Hattie and Timperley 2007). Providing feedback after a test enables students to correct memory errors (Butler and Roediger 2008) and maintain correct responses (Butler et al. 2008). Although testing improves retention even without feedback (Glover, 1989; Roediger and Karpicke 2006b; Karpicke and Roediger 2008), feedback can enhance the benefits of testing, especially when learners fail to retrieve the correct response (Kang et al. 2007).

Before discussing the benefits of feedback, it is important to stress that testing increases retention even without feedback. Many studies have found that testing without feedback enhances retention in laboratory settings (Butler and Roediger 2008; Karpicke and Roediger 2010; Roediger and Karpicke 2006b) as well as in real-life medical education settings (Turner et al. 2011). The fact that testing without feedback improves retention is evidence that retrieval has a direct effect on memory, thereby improving retention even in the absence of further studying or exposure to information.

Nevertheless, providing feedback after a test can further improve retention relative to testing without feedback (Butler et al. 2007; Butler and Roediger 2008). For example, Karpicke and Roediger (2010, pp. 116–124) showed when repeated testing was coupled with feedback, the level of retention rose by 25% or more compared to testing without feedback. This finding is an example of an indirect effect of testing—improved learning through the studying of feedback materials. Agrawal et al. (2012, pp. 326–335) have demonstrated that testing provides important opportunities for monitoring learning because students realize the limits of their knowledge when they confront test questions. Feedback allows learners to then build on those realizations and focus their learning on correcting errors. The act of attempting retrieval before restudying information may be important to effective learning from feedback. Kornell et al. (2009, pp. 989–998) showed that if learners attempted and failed to answer a difficult question before studying the answer to the question, they remembered more than if they studied the question and answer together. Students are conscious of the learning process from tests. In the study by Larsen et al. (2012), when students were asked how testing affected their learning, they reported that testing allowed them to verify their levels of knowledge, correct mistakes, and work on improved performance.

The timing of feedback may also be an important factor in determining retention. Although many educators and researchers assume that feedback must be given immediately in order to be effective (Mory 2004), recent studies have shown that delaying feedback may be more beneficial (Butler et al. 2007; Butler and Roediger 2008; Metcalfe et al. 2009). However, one critical assumption in recommending delayed feedback is that all of the feedback is fully processed. Often, students are not motivated to go over feedback when it is given after a delay; if full processing of the feedback cannot be guaranteed, then it may be better to give immediate feedback.

The forms of feedback can be as varied as the forms of testing. In addition to the traditional forms of formal testing with formal answers (whether electronic or paper), educators must think about simulation and clinical practice as well. Feedback and debriefing have long been considered important elements of learning from simulation (Rudolph et al. 2008). In a clinical setting, testing may take the form of oral questions given by a supervising clinician or may take the form of a patient encounter. In all of these setting educators should consider what types of feedback are provided. For patient encounters in particular, educators must consider whether clinical supervision is provided in a way that provides meaningful feedback—either from direct observation of clinical activities by the supervising physician or from thorough discussion and follow-up. Feedback amplifies the direct effects of testing and makes tests even more powerful learning interventions.

Conclusions

- Test-enhanced learning represents a powerful learning tool that could be utilized to improve medical education.
- Retrieval practice can take many forms, ranging from written tests to actual patient encounters.
- The form of testing used should be closely aligned with educational objectives.
- Production tests (e.g. short-answer, free recall or simulation) tend to promote better long-term retention than recognition tests (e.g. multiple choice tests).
- Use repeated retrieval practice spaced out over time whenever possible, with intervals that are close enough to prevent forgetting but long enough to require some effort to recall.
- Provide feedback after each test to facilitate learning and improve metacognition.

References

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

Accreditation Council for Graduate Medical Education (2011) *Common Program Requirements*. Chicago, IL: [Online]http://www.acgme.org/acWebsite/dutyHours/dh_dutyhoursCommonPR07012007.pdf Accessed 19 April 2012

Agrawal, S., Norman, G.R., and Eva, K.W. (2012) Influences on medical students' self-regulated learning after test completion. *Med Educ.* 46: 326–335

Bahrick, H.P., Bahrick, L.E., Bahrick, A.S., and Bahrick, P.E. (1993) Maintenance of foreign language vocabulary and the spacing effect. *Psychol Sci.* 4: 316–321

Bangert-Drowns, R.L., Kulik, J.A., and Kulik, C.C. (1991) Effects of frequent classroom testing. *J Educ Res.* 85: 89–99

Bangert-Drowns, R.L., Kulik, C.C., Kulik, J.A., and Morgan, M. (1991) The instructional effect of feedback in test-like events. *Rev Educ Res.* 61: 213–238

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

Bjork, R.A. (1975) Retrieval as a memory modifier. In: R. Solso (ed.) *Information Processing and Cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Lawrence Erlbaum Associates

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

Bjork, R.A., and Bjork, E.L. (1992) A new theory of disuse and an old theory of stimulus fluctuation. In: A. Healy, S. Kosslyn, and 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., and William, D. (1998) Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice.* 5: 7–74

Brewer, G.A., and Unsworth, N. (2012) Individual differences in the effects of retrieval from long-term memory. *J Memory Language.* 66: 407–415

Butler, A.C. (2010) Repeated testing produces superior transfer of learning relative to repeated studying. *J Exp Psychol: Learn Memory Cogn.* 36: 1118–1133

Butler, A.C., and Roediger, H.L., III (2007) Testing improves long-term retention in a simulated classroom setting. *Eur J Cogn Psychol*. 19: 514–527

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

Butler, A.C., Karpicke, J.D., and Roediger, H.L., III (2007) The effect of type and timing of feedback on learning from multiple-choice tests. *J Exp Psychol Appl.* 13: 273–281

Butler, A.C., Karpicke, J.D., and Roediger, H.L., III (2008) Correcting a metacognitive error: Feedback enhances retention of low confidence correct responses. *J Exp Psychol Learn Memory Cogn.* 34: 918–928

Cacamese, S.M., Eubank, K.J., Hebert, R.S., and Wright, S.M. (2004) Conference attendance and performance on the in-training examination in internal medicine. *Med Teach.* 26: 640–644

Carpenter, S.K. (2009) Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *J Exp Psychol: Learn Memory Cogn.* 35: 1563–1569

Carpenter, S.K. and Pashler, H. (2007) Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bull Rev.* 14: 474–478

Carpenter, S.K., Pashler, H., and Cepeda, N.J. (2009) Using tests to enhance 8th grade students' retention of U. S. history facts. *Appl Cogn Psychol.* 23: 760–771

Carrier, M., and Pashler, H. (1992) The influence of retrieval on retention. *Memory Cogn.* 20: 632–642

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., and Perfect, T. (2007) Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *Eur J Cogn Psychol.* 19: 580–606

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

Cepeda, N.J., Vul, E., Rohrer, D., Wixted, J.T., and Pashler, H. (2008) Spacing effect in learning: A temporal ridgeline of optimal retention. *Psychol Sci.* 19: 1095–1102

Dempster, F.N. (1989) Spacing effects and their implications for theory and practice. *Educ Psychol Rev.* 1: 309–330

Ebbinghaus, H. (1967) *Memory: A Contribution to Experimental Psychology* (H. A. Ruger and C. E. Bussenius, Trans.) New York: Dover (Original work published 1885)

Fitch, M.L., Drucker, A.J., and Norton, J.A. (1951) Frequent testing as a motivating factor in large lecture courses. *J Educ Psychol.* 42: 1–20

FitzGerald, J.D., and Wenger, N.S. (2003) Didactic teaching conferences for IM residents: who attends, and is attendance related to medical certifying examination scores? *Acad Med.* 78: 84–89

Fritz, C.O., Morris, P.E., Acton, M., Voelkel, A.R., and Etkind, R. (2007a) Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Appl Cogn Psychol.* 21: 499–526

Fritz, C.O., Morris, P.E., Nolan, D., and Singleton, J. (2007b) Expanding retrieval practice: An effective aid to preschool children's learning. *Q J Exp Psychol*, 60, 991–1004.

Gates, A.I. (1917) Recitation as a factor in memorizing. *Arch Psychol.* 6(40): 1–104

Glover, J.A. (1989) The 'testing' phenomenon: Not gone but nearly forgotten. *J Educ Psychol.* 81: 392–399

Hammond, W.A. (1902) *Aristotle's Psychology: A Treatise on the Principle of Life: (De Anima and Parva Naturalia)* Macmillan: New York

Hattie, J. and Timperley, H. (2007) The power of feedback. *Rev of Educ Res.* 77: 81–112

Issenberg, S.B., McGaghie, W.C., Petrusa, E.R., Lee, G.D., and Scalese, R.J. (2005) Features and uses of high-fidelity medical simulations that lead to effective learning: A BEME systematic review. *Med Teach.* 27: 10–28

Jacoby, L.L., Wahlheim, C.N., and Coane, J.H. (2010) Test-enhanced learning of natural concepts: effects on recognition memory, classification, and metacognition. *J Exp Psychol Learn Memory Cogn.* 36: 1441–1451

Johnson, C.I. and Mayer, R.E. (2009) A testing effect with multimedia learning. *J Educ Psychol.* 101: 621–629

Jones, H.E. (1923–1924) The effects of examination n the performance of learning. *Arch Psychol.* 10: 1–70

Kang, S.H.K., McDaniel, M.A. and Pashler, H. (2011) Effects of testing on learning of functions. *Psychonomic Bull Rev.* 18: 998–1005

Kang, S.H.K., McDermott, K.B. and Roediger, H.L., III (2007) Test format and corrective feedback modulate the effect of testing on memory retention. *Eur J Cogn Psychol.* 19: 528–558

Karpicke, J.D., and Blunt, J.R. (2011) Retrieval practice produces more learning than elaborative studying with concept mapping. *Science.* 331: 772–775

Karpicke, J.D. and Roediger, H.L., III (2007) Repeated retrieval during learning is the key to long-term retention. *J Memory Language*. 57: 151–162

Karpicke, J. D. and Roediger, H. L., III (2008) The critical importance of retrieval for learning. *Science.* 15: 966–968

Karpicke, J. D. and Roediger, H. L. III (2010) Is expanding retrieval a superior method for learning text materials? *Memory Cogn.* 38: 116–124

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

Kerfoot, B.P. (2009) Learning benefits of on-line spaced education persist for 2 years. *J Urol.* 181: 2671–2673

Kornell, N. and Bjork, R. A. (2008) Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory.* 16: 125–136

Kornell, N., Hays, M.J., Bjork, R.A. (2009) Unsuccessful retrieval attempts enhance subsequent learning. *J Exp Psychol Learn Memory Cogn.* 35: 989–998

Kromann, C.B., Jensen, M.L., and Ringsted, C. (2009) The effects of testing on skills learning. *Med Educ.* 43: 21–27

Kromann, C.B., Jensen, M.L., and Ringsted, C. (2010) The testing effect on skills might last 6 months. *Adv Health Sci Educ.* 15: 395–401

Kulhavy, R.W., and Stock, W.A. (1989) Feedback in written instruction: The place of response certitude. *Educ Psychol Rev.* 1: 279–308

Larsen, D.P., Butler, A.C., Lawson, A.L., and Roediger, H.L., III (2012) The importance of seeing the patient: Test-enhanced learning with

standardized patients and written tests improves clinical application of knowledge. *Adv Health Sci Educ.* doi: 10.1007/s10459–012–9379–7 (published online ahead of print)

Larsen, D.P., Butler, A.C., and Roediger, H.L., III (2008) Test-enhanced learning in medical education. *Med Educ.* 42: 959–966

Larsen, D.P., Butler, A.C., and Roediger, H.L., III (2009) Repeated testing improves long-term retention relative to repeated study: A randomized, controlled trial. *Med Educ.* 43: 1174–1181

Larsen, D.P., Butler, A.C., and Roediger, H.L., III (2013) Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Med Educ.* in press

Marton, F. and Saljo, R. (1976) On qualitative differences in learning: I—outcome and process. *Br J Educ Psychol.* 46: 4–11

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

McDaniel, M.A., Agarwal, P.K., Huelser, B.J., McDermott, K.B., and Roediger, H.L., III (2011) Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *J Educ Psychol.* 103: 399–414

McDaniel, M.A., Anderson, J.L., Derbish, M.H., and Morrisette, N. (2007) Testing the testing effect in the classroom. *Eur J Cogn Psychol.* 19: 494–513

McDaniel, M.A., Howard, D.C., and Einstein, G.O. (2009) The read-recite-review study strategy: Effective and portable. *Psychol Sci.* 20: 516–522

McGaghie, W.C., Issenberg, S.B., Cohen, E.R., Barsuk, J.H., and Wayne, D.B. (2011) Does simulation-based medical education with deliberate practice yield better results than traditional clinical education? A meta-analytic comparative review of the evidence. *Acad Med.* 86: 706–711

Metcalfe, J., Kornell, N., and Finn, B. (2009) Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory Cogn.* 37: 1077–1087

Michael, J. (1991) A behavioral perspective on college teaching. *The Behavior Analyst.* 14: 229–239

Morris, C.D., Bransford, J.D., and Franks, J.J. (1977) Levels of processing versus transfer-appropriate processing. *J Verbal Learn Verbal Behav.* 16: 519–533

Mory, E.H. (2004) Feedback research review. In: D. Jonassen (ed.) *Handbook of Research on Educational Communications and Technology* (pp. 745–783). Mahwah, NJ: Erlbaum

Pashler, H., Bain, P., Bottge, B., et al. (2007) *Organizing instruction and study to improve student learning: A practice guide* (NCER 2007–2004) Washington, DC: National Center for Education Research, Institute of Education Sciences, US Department of Education

Picciano, A., Winter, R., Ballan, D., Bimberg, B., Jacks, M., and Laing, E. (2003) Resident acquisition of knowledge during a noontime conference series. *Fam Med.* 35: 418–422

Phelps, R.P. (2012) The effect of testing on student achievement, 1910–2010. *Int J Testing.* 12: 21–43

Price, J.W., Price, J.R., Pratt, D.D., Collins, J.B., and McDonald, J. (2010) High-fidelity simulation in anesthesiology training: a survey of Canadian anesthesiology residents' simulator experience. *Can J Anesthesiol.* 57: 134–142

Pyc, M.A., and Rawson, K.A. (2009) Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *J Memory Language.* 60: 437–447

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

Pyc, M.A., and Rawson, K.A. (2012) Why is test–restudy practice beneficial for memory? an evaluation of the mediator shift hypothesis. *J Exp Psychol Learn Memory Cogn.* 38 737–746

Rawson, K.A., and Dunlosky, J. (2011) Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *J Exp Psychol Gen.* 140: 283–302

Rees, P.J. (1986) Do medical students learn from multiple choice examinations? *Med Educ.* 20: 123–125

Roediger, H.L., III and Butler, A.C. (2011) The critical role of retrieval practice in long-term retention. *Trends Cogn Sci.* 15: 20–27

Roediger, H.L., III and Karpicke, J.D. (2006a) The power of testing memory: Basic research and implications for educational practice. *Perspectives Psychol Sci.* 1: 181–210

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

Roediger, H.L., III, Gallo, D.A., and Geraci, L. (2002) Processing approaches to cognition: The impetus from the levels of processing framework. *Memory.* 10: 319–332

Rohrer, D. and Taylor, K. (2006) The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Appl Cogn Psychol*, 20: 1209–1224

Rudolph, J.W., Simon, R., Raemer, D.B., and Eppich, W.J. (2008) Debriefing as formative assessment: Closing performance gaps in medical education. *Acad Emerg Med.* 15: 1–7

Safdieh, J.E., Lin, A L., Aizer, J., et al. (2011) Standardized patient outcomes trial (SPOT) in neurology. *Med Educ Online.* 16: 5634

Schmidmaier, R., Ebersbach, R., Schiller, M., Hege, I., Holzer, M., and Fischer, M. R. (2011) Using electronic flashcards to promote learning in medical students: Retesting versus restudying. *Med Educ.* 45: 1101–1110

Schmidt, H.G., and Rikers, R.M.J.P. (2007) How expertise develops in medicine: knowledge encapsulation and illness script formation. *Med Educ.* 41: 1133–1139

Slamecka, N.J., and Katsaiti, L.T. (1988) Normal forgetting of verbal lists as a function of prior testing. *J Exp Psychol Learn Memory Cogn.* 14: 716–727

Smith, P.L., and Ragan, T.J. (2005) *Instructional Design*. Hoboken, NJ: John Wiley and Sons, Inc.

Spitzer, H.F. (1939) Studies in retention. *J Educ Psychol.* 30: 641–656

Thompson, C.P., Wenger, S.K., and Bartling, C.A. (1978) How recall facilitates subsequent recall: A reappraisal. *J Exp Psychol Human Learn Memory.* 4: 210–221

Thorndike, E.L. (1906) The principles of teaching based on psychology. New York: A. G. Seiler

Tse, C.-S., Balota, D.A., and Roediger, H.L. III. (2010) The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging.* 25: 833–845

Tulving, E. (1967) The effects of presentation and recall of material in free-recall learning. *J Verbal Learn Verbal Behav.* 6: 175–184

Turner, N.M., Scheffer, R., Custers, E., and Cate, O.T. (2011) Use of unannounced spaced telephone testing to improve retention of knowledge after life-support courses. *Med Teach.* 33: 731–737

Wayne, D.B., Butter, J., Sidall, V.J., et al. (2006a) Mastery learning of advanced cardiac life support skills by internal medicine residents using simulation technology and deliberate practice. *J Gen Intern Med.* 21: 251–256

Wayne, D.B., Siddall, V.J., Butter, J., Fudala, M.J., Wade, L.D., and Feinglass, J., (2006b) A longitudinal study of internal medicine residents' retention of advanced cardiac life support skills. *Acad Med.* 81: S9–S12

Wayne, D.B., Didwania, A., Feinglass, J., Fudala, M.J., Barsuk, J.H., and McGaghie, W.C. (2008) Simulation-based education improves quality of care during cardiac arrest team responses at an academic teaching hospital: A case-control study. *Chest.* 133: 56–61

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

Winter, R.O., Picciano, A., Bimberg, B., et al. (2007) Resident knowledge acquisition during a block conference series. *Fam Med.* 39: 498–503

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