Conceptual Foundations

for Designing Continuing Certification Assessments for Physicians

SCOTT H. FRAUNDORF PHD
University of Pittsburgh

ZACHARY CADDICK MS
University of Pittsburgh

BENJAMIN M. ROTTMAN PHD
University of Pittsburgh

TIMOTHY NOKES-MALACH PHD
University of Pittsburgh

DAVID B. SWANSON PHD
American Board of Medical Specialties

ANDREW BAZEMORE MD MPH
American Board of Family Medicine

THOMAS O'NEILL PHD
American Board of Family Medicine

REBECCA S. LIPNER PH.D
American Board of Internal Medicine

Contents

Executive Summary 5

1. Cognitive Skills Need to be Kept Current 5
2. Self-Assessment is not enough 5
3. Testing Enhances Learning and Retention 6
4. Goals and Consequences Motivate 6

Learning Model 6

Lessons Learned 7

Conclusions and Limitations 8

Chapter 1: Introduction 9

Shifts in Design of Continuing Certification Assessments 9

Goals of the Research 10

Overview of the White Paper 10

References 11

Chapter 2: Cognitive Skills Need to be Kept Current 12

Key Points 12

Overview 12

Acquiring Medical Expertise 12

Maintaining Expertise 16

Aging and its Relationship to Memory and Learning 19

Keeping up with Changing Standards of Care 23

Chapter Summary 24

Future Directions 25

References 29

Chapter 3: Self-Assessment is not enough 36

Key Points 36

Overview 36

Monitoring Accuracy Has Two Components 37

Metacognitive Monitoring Can Be Reasonably Accurate 38

People Can Accurately Control Reporting 40

Metacognitive Monitoring Is Subject to Systematic Biases 41

Poor Performers Overestimate Their Performance 44

Other Factors Can Influence What Learners Choose to Study 45

Theoretical Mechanisms 46
Explicit Instruction Does Not Remove Self-Assessment Biases 47
Chapter Summary 48
Future Directions 49
References 52
Chapter 4: Testing Enhances Learning and Retention 60
Key Points 60
Overview 60
Benefits to Spaced Learning in Medicine 62
Moderators of Testing Effect 64
Feedback 70
Training People to Use Retrieval Practice 73
Cognitive Mechanisms Underlying the Testing Effect 75
Chapter Summary 76
Future Directions 78
References 80
Chapter 5: Goals and Consequences Motivate 92
Key Points 92
Overview 92
Expectancies 93
Perceived Benefits/Value: What is the value of the test to the learner? 96
Benefits of Pursuing Mastery and Achievement Goals 100
Growth Mindsets Benefits Motivation and Learning 102
External Incentives Can Undermine Learning 103
Perceived Costs of Testing 104
Test Anxiety 105
Chapter Summary 107
Future Directions 108
References 111
Chapter 6: Synthesis and Summary 122
Assessment and Learning Model 122
Cognitive Skills Must Be Kept Current 122
Self-Assessment is not enough 123

Testing Enhances Learning and Retention 124
Goals and Consequences Motivate 124
A Cross-Cutting Theme - Feedback on Performance 125
Recommendations 126
Proposed Studies 129
The Role of Longitudinal Assessment in Comparison to Other Life-Long Learning Mechanisms 130
Features of Learning Opportunities 131
Six Lifelong Learning Opportunities 132
Translatability of Basic Research to Lifelong Learning in Medicine 137
Conclusion 138
References 139

Executive Summary

ABMS member boards are shifting assessment formats from a single point-in-time, high-stakes assessment to more frequent, learner-engaged assessments. This change may be characterized as a shift towards a blend of *assessment for learning* and *assessment of learning*. Given this shift in the certification paradigm, we were asked to explore and define a theoretical framework for the continuing assessment of physicians' clinical knowledge based upon foundational science from multiple disciplines. Our review revealed four central themes that underpin the need for continuing medical certification and support the shift towards longitudinal assessment, specifically that: 1) cognitive skills need to be kept current, 2) self-assessment is not enough, 3) testing enhances learning and retention, and 4) goals and consequences motivate. We briefly summarize each theme and present a learning model.

1. Cognitive Skills Need to be Kept Current

Over the course of training, physicians develop significant knowledge and expertise. However, existing literature clearly supports the need for individuals to engage in appropriate training and study to keep these cognitive skills current. Equally, it reinforces the idea that, in the absence of such training and study, cognitive skills will decline over time. Specifically, it supports the idea that decay of cognitive skills can manifest itself in the form of lower quality care as physicians get further from residency training. Evidence also suggests that knowledge or skills gained may not remain accessible to physicians for a number of reasons. These include an absence of study, and also the presence of other similar knowledge or skills that compete in what is brought to mind through a process called interference. Cognitive decline can also interfere with a task fundamental to clinical care, the ability to retrieve knowledge fast enough to be useful. Clinicians also face standards of care that continuously evolve over time and it is a considerable challenge to keep up with these changes. As such, evidence supports the idea that longitudinal assessment has the potential to serve not only an evaluation mechanism but also to promote learning and retention serving to maintain existing levels of knowledge and increasing awareness of new standards of care.

2. Self-Assessment is not enough

Physician learners have asked why cognitive skills cannot be kept current through self-assessment and self-directed remediation in areas of weakness. Although the existing literature shows that learners have some ability to self-assess strengths and weaknesses, considerable systematic biases in self-appraisal exist. Learners across disciplines tend to overestimate how much they will remember after learning. Of greater concern, is that it is the relatively poor performers in a domain that are particularly unaware of their poor performance or capacity. Furthermore, individuals frequently employ ineffective learning strategies, prioritizing activities that are more enjoyable (e.g., interesting subject matter or easier topics) over more pertinent learning activities directed at areas of weakness. Evidence from both basic and applied science on the difficulty of accurately measuring one's own knowledge and cognitive skills suggest that physicians should be at least partly guided in their choices of learning

activities and affirms the potential peril to public stakeholders if learners retain absolute control over the subject matter studied to maintain cognitive skills.

3. Testing Enhances Learning and Retention

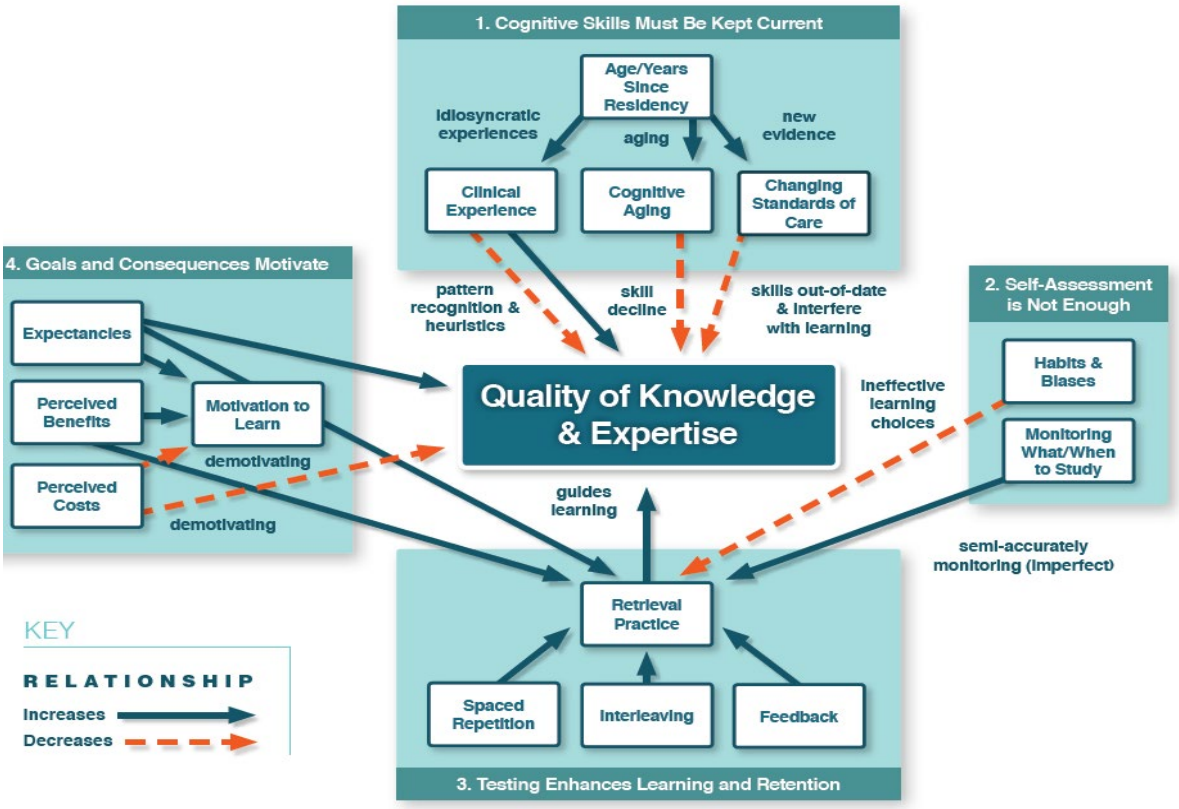
Meta–analysis studies provide robust evidence that testing is a powerful device for learning and retaining cognitive skills. Meta-analysis also provides robust evidence that taking a test serves as a strong learning experience in its own right because it allows learners to practice retrieval from memory. That testing aids learning has been demonstrated across numerous domains, including medicine. Testing is especially helpful when information must be retained and retrievable for long periods of time, as is the case for physicians. Evidence suggests that the effects of testing are boosted by spacing tests over time, and that testing benefits not only rote memorization, but also complex thinking, such as diagnostic classification and clinical reasoning. Additionally, the learning benefits of testing can transfer to related material that was not studied, demonstrating value beyond the narrowly-learned subject matter that is directly assessed. Feedback, specifically providing and explaining correct answers, makes learning more effective and, when applicable, should be provided to learners. Contrary to some lay concerns, making an error on a test has not been shown to be harmful to long-term retention, as long as feedback on why the answer is wrong is provided.

4. Goals and Consequences Motivate

Assessment can also serve as an important motivator. Physicians will be more motivated to study and practice their skills when there are clear *consequences* (i.e., benefit) for doing so and clear *costs* for not doing so. Physicians learn and retain more when they expect to be tested, thus testing should be difficult enough to engender deeper and more effective learning. Emphasizing how maintenance of medical expertise aligns with physicians' values (e.g., their best intentions in treating patients), and offering some degree of autonomy/control over learning content may foster intrinsic motivation within physicians. Research supports the tendency for intrinsically motivated individuals to work harder and persist longer in the face of difficulty, adopt better learning strategies, and procrastinate less. It additionally verifies that people learn and perform better when motivated by their own values and interests rather than strictly external rewards. There is reason to believe that an external framework offered by a tailored longitudinal assessment program may aid in these aspects, given evidence suggesting that physicians are highly motivated individuals.

Learning Model

The learning model presents a synthesis of topics influencing quality of knowledge and expertise and is presented below. We detail its components in the full white paper.



Lessons Learned

We identified a number of important lessons and recommendations in this critical review of existing literature, including the following:

- The unique clinical experiences and tendencies of physicians can lead to bias in self-assessment and learning strategies. Tailored longitudinal assessment is needed to fill in potential gaps in experience and account for idiosyncratic experiences.
- Clear and recurring feedback is critical, paired with ongoing assessment of confidence to grow self-awareness of knowledge gaps. Self-reported confidence ratings should be collected *before* feedback is provided, as ratings gathered *after* feedback may be biased by the feedback.
- To maximize deep learning, concepts should be spaced both within and across assessments (i.e., having all questions on a single topic in a row or within the same quarter) in a longitudinal paradigm.
- Framing longitudinal assessment in terms of its learning benefits and growth may serve to foster a natural intrinsic motivation present in physicians, to reduce anxiety and to dampen negative perceptions of how “high-stakes” a test is.

Conclusions and Limitations

There is strong evidence from multiple areas, including medicine, that testing and retesting is an efficient mechanism for promoting learning and retention. Boards should consider this evidence in designing their longitudinal assessment programs.

Based on literature and research spanning a range of disciplines related to the development and maintenance of expertise in physicians, we identified a need for physicians to keep cognitive skills current in the contexts of changing standards of care and of skill decline over time. Self-assessing areas of weaknesses, although helpful, is unlikely to be sufficient given the presence of several systematic biases in self-assessment. Periodic tests are likely to be extremely valuable in complementing self-assessment because the act of taking a test itself enhances learning and retention of cognitive skills. Testing can also serve as a motivator by providing goals and consequences. We also identified a cross-cutting finding that feedback loops in the medical system are often incomplete, which can lead to poor learning and maintenance of expertise and can also lead to poor metacognitive awareness of limitations in knowledge. Longitudinal assessment holds the potential to serve as one way to close the feedback loop and promote learning and retention.

The basic science of learning provides general answers--or at least important considerations--for many of the questions raised in this report. At the same time, because much of the research has been conducted in settings outside of medicine, this review of the literature does not guarantee that the findings will translate seamless for medical expertise. Still, the learning sciences can provide useful methods for answering many of the remaining open questions. The boards of medical specialties should consider teaming with experts in cognitive, social, and motivational psychology--and the learning sciences more broadly--to design and conduct applied studies that can simultaneously optimize learning in longitudinal assessment and address questions of interest to basic science.

Chapter 1: Introduction

Noting the need to ensure that physicians maintained cognitive skills across the span of a professional career, the 24 Member Boards of the American Board of Medical Specialties (ABMS) pivoted from lifelong to time-limited certificates requiring physicians who were initially certified by the Board, known as *Diplomates,* to take and pass an assessment every six to ten years to maintain their certification.[1] Historically, these assessments took the form of point-in-time multiple-choice question assessments taken by Diplomates at secure testing centers, much like the assessments used for initial certification. These are best viewed as retrospective "assessments of learning" (i.e., summative assessments), designed to determine if the current knowledge base of a Diplomate remains at or above a level commensurate with certification in the associated specialty or subspecialty.

Another shift in the certification paradigm started in 2014 when the American Board of Anesthesiology began pilot work on their "MOCA Minute" program (Sun et al, 2016). In contrast to traditional point-in-time assessments, MOCA Minute was designed as a proactive "assessment for learning" (i.e., a formative assessment) in which Diplomates completed a series of questions taken longitudinally over the course of the year. Participation was intended to assist Diplomates in keeping up with changes in medicine, promoting, learning, retaining, and applying knowledge in patient care. This approach draws on advances in cognitive psychology (Birnbaum et al, 2013, Brown et al, 2014; Cepeda et al, 2006; Dempster, 1988; Karpicke & Roediger, 2008; Roediger & Butler, 2011), including spaced learning and repetition, and immediate performance feedback settings, as well as upon recent advances in internet-based testing such as inclusion of hyperlinks to learning resources.

Physician certification programs have evolved in content, approach, and also in name over recent decades. Initial certification programs first required recertification before moving into a new paradigm of Maintenance of Certification (MOC), emphasizing continuous professional development at the start of the new millennium. Over the past decade, all 24 ABMS member boards agreed to develop and migrate towards continuing certification (CC) programs signaling that training and acquisition of medical practice knowledge and skill begins in medical school, is enhanced during residency, and maintained throughout a specialist's career.

Shifts in Design of Continuing Certification Assessments

Shortly after the introduction of MOCA Minute, some other ABMS Member Boards began planning more frequent, lower-stakes assessments as part of their assessment programs. As of mid-2020, all 24 Boards have announced programs that blend both "assessment of learning" and "assessment for learning." In common across the programs is an emphasis on provision of specific immediate feedback on performance, timely identification of areas of strength and weakness (assessment for learning), use of aggregated performance over time to make summative decisions (assessment of learning) regarding continuing certification (Price et al, 2018), and increased relevance of the assessments to Diplomates' practice. At the same time, the programs are diverse in the frequency of

[1] Most Boards had additional requirements for maintenance of certification, including possession of an active, unrestricted medical license, acquisition of a specified number of continuing medical education credits, and engagement in quality improvement projects.

the summative assessment, the participation requirement, the number of questions included, the time allotted per question, the use of spaced repetition, and the format of the aggregate feedback provided.

Goals of the Research

Because of the diversity of the longitudinal assessment programs across the specialty boards, the American Board of Internal Medicine along with the American Board of Family Medicine and the ABMS decided in 2020 to support the development of a "white paper" that reviewed the foundational research in cognitive and learning sciences and medical education underlying longitudinal assessment, synthesized the findings into recommendations for best practices, and identified key research gaps to be addressed. Cognitive psychology is the subfield of psychology (and also a part of the interdisciplinary field of cognitive science) that studies mental processes such as learning, memory, thinking, and problem solving. A team of cognitive psychologists from the University of Pittsburgh were commissioned to do this work so as to present an unbiased view of the state of the research.

This white paper is intended to provide a theoretical framework for continuing assessment of physicians' clinical knowledge. The framework presents a model of the foundational science and addresses some practical implications for the form that assessment and learning should take through a professional's career, the frequency with which Diplomates should engage with continuing assessment, the potential of spaced repetition in the design of the assessment, the most appropriate ways to motivate learning, and the key areas of research that are important for helping the Board's community to determine whether the longitudinal programs were, in fact, improving cognitive skills and, in turn, patient care.

Overview of the White Paper

We review the foundational science behind longitudinal assessment and arrive at four critical themes: 1) cognitive skills must be kept current, otherwise they will decline over time, 2) self-assessment is not enough to reliably and effectively assess one's own competencies and to guide one's own learning, 3) testing enhances learning and retention of cognitive skills and knowledge, and 4) the role of motivation for learning in relation to assessments. These themes have been divided into individual chapters. In our review, we prioritized empirical findings from basic cognitive science and complementary medical evidence, where it was found. Additionally, we identified gaps in knowledge and propose a number of follow up studies that have relevance for longitudinal assessment of medical knowledge. In the summary chapter we synthesize evidence across the four themes and present a learning model. The final chapter also presents recommendations for longitudinal assessment programs based on our review.

Not at all empirical evidence is equal. To properly situate the strength of the evidence and claims made throughout this paper, we have attached evidence levels (EL) to in-text citations for empirical claims. The evidence levels range from 1 to 6; 1 being the strongest evidence (quantitative meta-analyses), 2 (narrative review), 3 (Evidence from multiple original experiments) 4 (single experiment), 5 (correlational or quasi-experimental studies) and 6 being the weakest evidence (opinion papers).

References

Birnbaum, M. S., Kornell, N., Bjork, E. L., & Bjork, R. A. (2013). Why interleaving enhances inductive learning: The roles of discrimination and retrieval. *Memory & Cognition*, *41*(3), 392-402.

Brown, P. C., Roediger III, H. L., & McDaniel, M. A. (2014). Make it stick: The science of successful learning. Cambridge, MA, Harvard University Press.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354-380.

Dempster, F. N. (1988). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist*, *43*(8), 627-634.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966-968.

Price D., Swanson D. B., Irons, M., Hawkins, R. E. (2018). Longitudinal assessments in continuing specialty certification and lifelong learning. *Medical Teacher*, *40*(9), 917-919.

Roediger, H.L., 3rd, Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Sun, H., Zhou, Y., Culley, D. J., Lien, C. A., Harman, A. E., Warner, D. O. (2016). Association between participation in an intensive longitudinal assessment program and performance on a cognitive examination in the Maintenance of Certification in Anesthesiology Program®. *Journal of the American Society of Anesthesiologists*, *125*(5), 1046-1055.

Chapter 2: Cognitive Skills Need to be Kept Current

Key Points

- Utilizing experience in medical decision-making can enable very fast pattern recognition, which frequently leads to accurate diagnoses.
- However, in certain instances, relying too heavily on experience can lead to irrational decisions biased by idiosyncratic experiences and poor feedback mechanisms.
- There is converging evidence that physicians further out from training tend to perform worse on tests of medical knowledge and provide poorer patient care. The extent to which this is due to cognitive aspects of aging versus failing to keep up with changing standards of care versus specialization of a physician's practice and loss of broader skills and knowledge is unclear.
- Though there are many barriers to following guidelines and current standards of care, one that is within the control of the physician is knowledge and awareness of new standards.

Overview

A primary goal of continuing certification programs is to ensure that physicians who hold themselves out to the public as being board-certified maintain at least a certain minimum level of expertise. In this chapter, we break expertise down into four processes. The first is initially acquiring medical expertise. To this end, in the first section we discuss psychological theories of the acquisition of higher-order cognitive medical sills, and in particular, how diagnosis is learned. The second process is maintaining medical expertise. The third process is aging and its relationship to memory and learning. And the fourth process is keeping up with changing standards of care. The goal of this chapter is to evaluate these four processes related to keeping skills current, and how a longitudinal assessment could influence each process.

Throughout this chapter, as well as this entire report, we focus on medical decision making and expertise from the traditional information-processing perspective of cognition that undergirds cognitive psychology. We acknowledge that in reality medical decision making is much more complex in that it is situated in a complex environment with other physicians and healthcare professionals and in the larger context of medical systems (see the 2020 special issue of the journal *Diagnosis*, Volume 7, Issue 3, for many articles on this perspective). However, because continuing certification program assessments test a physician's cognitive abilities independently, not their performance within a specific health system, we focus on the individual physician's cognitive skills.

Acquiring Medical Expertise

Expertise is marked by the acquisition of large amounts of knowledge, which affects how information is organized, represented, and processed. General aptitude measures struggle to predict expert performance, suggesting that expertise is not just reserved for the highly intelligent (Moneta-Koehler, Brown, Petrie, Evans, Chalkley, 2017, EL: 5). Rather, experiences play an important role in the development of expertise. For instance, the amount of *deliberate practice* activities that an individual has completed that are designed to improve targeted aspects of performance predicts their level of

expertise (Ericsson, Krampe, Tesch-romer, 1993, EL: 5). It is likely the *quality* of deliberate practice, rather than quantity, is what is necessary to develop expertise; the mere number of deliberate practice hours on their own do not adequately explain expert performance (Macnamara, Hambrick, & Oswald, 2014, EL: 1). Ample and accurate feedback is also crucial (Kahneman & Klein, 2009, EL: 2). In the process of developing expertise, individuals learn to categorize information based on abstract principles, whereas novices categorize based on superficial details (Chi, Feltovich, & Glaser, 1981, EL: 3).

Dual-Process Theories and Advantages of Experience in Medical Decision Making

A common theme in the cognitive psychology literature is the existence of two distinct systems for information processing (for overview see Evans, 2008, EL: 2). (The term "systems" in this literature refers to collections of cognitive strategies and habits, not necessarily to neural or anatomical distinctions.) Most dual-processing theories hold that System 1 is fast, unconscious, evolutionarily old, associative, and universal. In contrast, System 2 is slow, conscious, evolutionarily new, and rule based (Evans, 2008, EL: 2). An important difference between System 1 and System 2 processing is that System 2 is under the control and guidance of the individual whereas System 1 occurs automatically. Although it may seem intuitive that the conscious and controlled System 2 is superior to System 1, this is not always the case. System 1 can, at times, produce highly accurate decisions, efficiently and from little information (Marewski & Gigerenzer, 2012).

Within medicine, the dual-process theory is widely accepted as the dominant paradigm for understanding medical decision making generally, and especially for diagnosis (Croskerry, 2009a, EL: 2; 2009b, EL: 2; Croskerry et al., 2013, EL: 2; Norman and Eva, 2010, EL: 2; Pelaccia et al., 2011, EL: 2). For example, the dual-process theory provides the theoretical backbone of the Institute of Medicine's 2015 report on Improving Diagnosis in Healthcare (National Academies of Sciences, Engineering, & Medicine, 2015, Chapter 2, EL: 2). One commonly discussed type of non-analytic System 1 decision making in medicine is the very-fast pattern recognition process. Pattern recognition allows a physician to very quickly think of myocardial infarction upon hearing chest pain on exertion in an older unfit man, or to think of hyperthyroidism when seeing a patient who is skinny, tremulous, perspiring a lot, with bulging eyes and a swollen neck. Pattern recognition involves classifying a current patient as similar to a prior patient (called an *exemplar*) or similar to an abstracted pattern of multiple prior patients with the same disease (called a *prototype*). Though most often discussed in terms of diagnosis, this pattern recognition process is also relevant to other decisions, such as deciding whether or not to order additional diagnostic testing, choosing a treatment, or deciding whether or not to refer to a specialist. Pattern recognition is believed to rely on the same cognitive processes that people use every day for categorizing animals as dogs versus cats or that arborists use to identify different species of trees (Cohen & Lefebvre, 2017). Another aspect of non-analytic System 1 decision making in medicine is the use of *heuristics*--mental shortcuts that allows them to reach decisions efficiently. In fact, pattern recognition through categorization can be viewed as one type of heuristic (e.g., Nilsson, Juslin, & Olsson, 2008), though heuristics are broader than just pattern recognition (e.g., National Academies of Sciences, Engineering, & Medicine, 2015, Chapter 2; Whelehan, Conlon, and Ridgway, 2020). For example, one heuristic called *representativeness* leads physicians to judge the probability that a patient has a given disease based on the *sensitivity* of the diagnostic information (probability of a positive test given that

the patient has the disease) rather than its *positive predictive value* (probability that a patient has a disease given a positive test; Casscells, Schoenberger, & Graboys, 1978; Eddy, 1982; Rottman, 2017). This heuristic is statistically appropriate in situations in which the prevalence of two diseases are roughly equivalent, but leads to a phenomenon called *base-rate neglect* when considering one disease that is very common and another that is very rare.

Thus, heuristics are neither inherently bad nor inherently good. Heuristics can help physicians make fast decisions which can be critical in situations with time pressure. And simple rule-based heuristics have even been shown to sometimes outperform formal statistical regression analysis (Marewski & Gigerenzer, 2012, EL: 2). On the other hand, sometimes heuristics are applied in the wrong context or are overly simple, such as the base rate neglect example, which can lead to suboptimal decisions.

In contrast to the non-analytical decision making of System 1, System 2 is understood as the analytical, hypothetico-deductive reasoning process. For example, a physician might diagnose a patient by systematically running one test, ruling out one diagnosis, and then following it up with a different test relevant to a different potential diagnosis--a process that involves a series of carefully thought-out decisions with logical reasoning. It is believed that System 1 and System 2 interact with each other, though exactly how this interaction occurs is debated (Croskerry, 2009b, EL: 2).

An important way to distinguish these two processes is the role of experience with prior patients. Experience with prior patients is at the core of System 1 (non-analytic) decision-making, which operates by recognizing the pattern in the current case as similar to prior cases. The same is not true for analytical decision-making, for which clinicians are encouraged to rely on rules, evidence, guidelines, and knowledge of pathophysiological and pharmacology rather than their own idiosyncratic past experience. (We do not claim that prior experiences with individual patients cannot be part of the analytical model of decision making and slow deliberative thought, only that experience is not central to this system the way it is for System 1.) For this reason, the role of prior individual experiences is less central in System 2. The strength of utilizing experience is that it allows for very fast pattern recognition, which frequently leads to accurate diagnoses. For example, one study (Norman et al., 1989, EL: 5) examined the accuracy of dermatologists reading slides. They found that not only did the dermatologists demonstrate high levels of accuracy, they in fact answered significantly faster on slides they got correct, showing how diagnosis can often be both extremely fast and accurate. A number of studies with primary care and emergency medicine physicians have found similar results: clinicians think of a few potential diagnoses within seconds to minutes and are usually right (Barrows et al., 1982, EL: 5; Elstein, Shulman, and Sprafka, 1978, EL: 5; Gruppen, Woolliscroft, & Wolf, 1988, EL: 5; Pelaccia, et al., 2014, EL: 5). This has led to the provocative question by Norman, Young, and Brooks (2007): “How can it be that experts with minimal information are able to advance tentative hypotheses about the diagnoses, seemingly effortlessly, and apparently without conscious awareness of the retrieval process? ... Where do the hypotheses come from?” The answer, according to this line of research, is that with enough experience clinicians may pattern-match a target case to a large set of prior cases to quickly come up with a diagnosis.

Shortcomings of Experience

However, there are also shortcomings to relying heavily on experience. One is that, even though pattern recognition and relying on the diagnosis and treatment decisions made for past patients can often work out well, it can also lead to biases. For example, in one experiment, family medicine residents were given a set of cases to practice interpreting ECGs. In the initial set of cases, a brief clinical scenario accompanied each case, along with the correct diagnoses. In the latter set of test cases, participants had to identify the correct diagnosis. For some of the test cases, the accompanying clinical scenario involved irrelevant features, such as the patient's job, that matched features from the initial cases. When an irrelevant feature matched a prior case, the residents were more likely to give the same diagnosis as the prior case, which turned out to be wrong (Hatala, Norman, & Brooks, 1999, EL: 4; Brooks et al., 1991, EL: 4; Young, Brooks, & Norman, 2011, EL: 4). A similar study relying on prior experience for use in real-world medical decision making is demonstrated by Choudhry et al. (2006, EL: 5). This study investigated how often physicians prescribed warfarin for patients with atrial fibrillation in order to prevent a stroke and, in particular, what happened to prescribing habits when one of the physician's patients on warfarin experienced a severe bleeding event that was likely a side effect of the warfarin. Compared to the 90 days before the event, the physicians were about 20% less likely to prescribe warfarin for patients with atrial fibrillation in the 90 days after the event, and this pattern of prescribing continued for at least a year. In sum, this study demonstrates that physicians' decisions can be strongly impacted by (recent) experiences with other patients, leading in this case to under prescribing of a medication that, despite the risks, is a standard of practice.

Together these findings suggest that prior patient experiences can have important consequences for subsequent decisions. Though experience is often believed to be beneficial, in that it can facilitate very fast and often accurate diagnoses and other decisions, sometimes prior experiences inadvertently lead to errors. An individual physician's experiences are necessarily idiosyncratic in a variety of ways. First, the cases that an individual physician sees are idiosyncratic. If a physician works in a specialized clinic, they may see certain types of patients in that clinic even though they still need to be able to diagnose and treat a broader set of patients that they see less frequently. This means that physicians are systematically missing out on experience with certain types of patients. For example, in one study, residents' beliefs about the prevalence of a disease was correlated with their probability of providing it as a potential diagnosis (Rottman, Prochaska, & Deaño, 2016, EL: 5). In general, this tendency is good and makes sense from the rational Bayesian perspective of diagnosis in which general "prior" beliefs about the likelihood of diseases in the population are sequentially updated with knowledge of the signs and symptoms and diagnostic tests of the specific patient to form a "posterior" probability of each disease on the differential (Ledley & Lusted, 1959; Pauker & Kassirer, 1980). However, to the extent that prevalent beliefs are distorted by particular experience, some potential diagnoses could be overlooked. Second, the appearance of patients with rare diseases or rare side effects from treatments (e.g., the bleeding events discussed above) is governed by chance. This means that physicians may be influenced by the vicissitudes of daily practice. For all these reasons, it is important to receive corrective feedback and not to overly rely on one's own experiences.

Another problem with relying on one's experience is that experience provides an imperfect feedback system. Feedback is vital for developing expertise, both generally and within medicine in

particular (e.g., Ericsson, 2015, EL: 2; Kahneman & Klein, 2009, EL: 2). However, the current medical system is imperfect in providing feedback for multiple reasons (National Academies of Sciences, Engineering, & Medicine, 2015, EL: 6; Schiff, 2008, EL: 6). First, an error in diagnosis or treatment may never be discovered, in which case feedback is never received. Second, because of the complex nature of modern medicine and the fact that an individual patient often has contact with many physicians, an individual physician often never knows the outcomes of patients that they encountered, resulting in a lack of both negative and positive feedback. Lack of feedback is believed to contribute to overconfidence (Kahneman & Klein, 2009, EL: 2). For these reasons, both the Institute of Medicine's *Best Care at Lower Cost* (McGinnis, Stuckhardt, Saunders, & Smith, 2013) report and its 2015 *Improving Diagnosis in Healthcare* (National Academies of Sciences, Engineering, & Medicine, 2015) report have called for healthcare organizations to create systems of feedback at many levels.

In summary, accumulating experience with many patients is believed to be a critical aspect of how physicians become experts. However, due to working in specialized practices, chance encounters with certain patients but not others, and imperfect feedback systems, physicians will not always experience (and re-experience) certain types of patients. For this reason, one opportunity for a longitudinal continuing certification program is to provide physicians with a well-rounded set of vignettes with feedback to supplement their real-world experiences.

Recommendation 2-1: Physicians have idiosyncratic experiences, which can lead to biased judgments. Therefore, longitudinal assessment can attempt to fill in potential gaps in experience through rich clinical vignettes in a continuous framework.

Maintaining Expertise

Cognitive Skills Decline over Time

In the absence of study, learned information and procedures are forgotten over time. Decades of research suggest that, across content domains and types of tasks, forgetting tends to follow a negatively accelerated *power law* function such that a great deal of material is forgotten initially, but the remaining material is forgotten more slowly (Ebbinghaus, 1885, EL: 4; Rubin & Wenzel, 1996, EL: 2; Wickelgren, 1974, EL: 4; Wickens, 1998, EL: 6; Wixted, 2004, EL: 3; Wixted & Carpenter, 2007, EL: 3; see *Figure 2-1*). That is, people forget much of what they see and hear in the initial minutes and hours afterwards, somewhat more in the following days and weeks, and comparatively little of what remains in the months and years ahead. This power-law function may reflect the rate at which people are likely to stop re-encountering those topics (Anderson & Schooler, 1991, EL: 5).

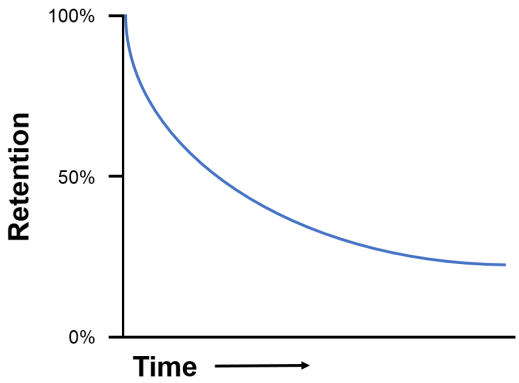


Figure 2-1. The forgetting curve

In terms of long-term retention of medical expertise, the power law suggests that some knowledge will be retained relatively well over long periods of time, but it is almost inevitable that other material will be quickly forgotten after being encountered in training if it is not deliberately practiced. The question of "how often does important information need to be practiced?" is likely to vary across individuals and be influenced by a number of other variables. Nevertheless, the rapid decline in retention after any given study episode suggests that it is more beneficial to distribute study over time so as to alleviate these losses.

Decay or Interference Both Reflect Failures to Retrieve Information

Why do people forget? One intuitive hypothesis might be that we simply run out of mental "storage space" and prior knowledge is forced out to make room for the new. It is clear that there are sharp restrictions on how much can be held in *working memory*, or what we are currently thinking about (although the specification and cause of those limits remain debated; e.g., Cowan, 2010, EL: 2; Miller, 1956, EL: 2). However, it is not clear that there are practical limits on the total capacity of long-term memory. As long as material is sufficiently distinct and meaningful, laboratory studies have demonstrated that people can readily learn and remember hundreds or thousands of pictures or sentences even after only seconds of exposure to each (e.g., Shepard, 1967, EL: 4; Standing, 1973, EL: 4; Standing, Conezio, & Haber, 1970, EL: 4). Indeed, the cortex of the human brain contains approximately 150 trillion (1.5×10^{14}) synapses (Drachman, 2005, EL: 3), which is orders of magnitude more than what the average individual knows: 40,000 (4×10^4) words (Brysbaert, Stevens, Mandera & Keuleers, 2016, EL: 3), 750 (7.5×10^2) people (Zheng, Salganik, Gleman, 2006, EL: 4), or, more generally, approximately 1 billion (10^9) bits of information (Landauer, 1986, EL: 2). Consequently, contemporary accounts of memory generally do not emphasize a total capacity limit as the primary cause of forgetting (e.g., Drachman, 2005; Landauer, 1986).

Why, then, do people forget? First, one intuitive hypothesis is that skills and knowledge are simply lost to *decay*; that is, memories fade and are eventually lost simply due to the passing of time. Nevertheless, not all memories are fated to decay; people can remember the names and locations of buildings on their college campus (Bahrick, 1983, EL: 5) or the names and faces of their high school classmates (Bahrick, Bahrick, & Wittlinger, 1975, EL: 5) even after decades of disuse.

Second, cognitive psychologists often emphasize *interference* from other, similar information as a major cause of retrieval failures. Many things commonly forgotten in daily life are those that compete with many other similar memories. For example, it is often difficult to remember where I left my phone this morning because I have many other competing memories of other places where I left my phone at different times. One experimental demonstration of interference is the *fan effect* (Anderson, 1974, EL: 3; Anderson & Reder, 1999, EL: 4): Learning multiple overlapping associations makes any individual association harder to retrieve (e.g., learning *the lawyer is in the cave* and *the lawyer is on the beach* is harder than learning *the lawyer is in the cave* and *the fireman is on the beach*). The process of interference can happen both *proactively*, when old knowledge makes it harder to learn competing new knowledge (Watkins & Watkins, 1975, EL: 2), and *retroactively*, when new knowledge, once acquired, interferes with retrieving old knowledge (Postman & Underwood, 1973, EL: 2). Indeed, interference-based failures to retrieve *some* information may be an inevitable consequence of remembering *other*, competing information (*retrieval-induced forgetting*; Anderson, Bjork, & Bjork, 1994, EL: 2; Roediger, 1978, EL: 2). For instance, imagine the process of diagnosing a patient with chest pain. Retrieving *myocardial infarction* in response to the cue *chest pain* reinforces the likelihood of thinking of *myocardial infarction* for future cases, and it also correspondingly weakens the likelihood of thinking of--or at least considering--*aortic dissection* as an alternate choice. Thus, less common concepts and information are particularly vulnerable to interference (Anderson, Bjork, & Bjork, 1994, EL: 3).

Ongoing research continues to probe the relative contribution of decay and interference to forgetting (Sadeh, Ozubko, Winocur, & Moscovitch, 2014: EL 2; Wixted, 2004: EL 2; Wixted, 2005: EL 2). For our purposes, however, these findings indicate at least two types of physicians' skills and knowledge that are particularly likely to be forgotten: (a) those that may have decayed because of substantial time lapse (as discussed in the previous section) and (b) those that are subject to interference because they are similar to other skills and knowledge, especially more frequently used ones.

Recommendation 2-2: Common concepts (e.g., common diagnoses, common treatment plans) may interfere with more rare concepts (e.g., rare diagnoses, less frequently used treatment plans). Due to interference, one goal for longitudinal assessment could be to test knowledge of the rare but important concepts that are especially easy to confuse with more common concepts.

Temporarily Inaccessible Knowledge Can Often Be Recovered or Relearned

Although physicians (and people in general) may sometimes be unable to bring to mind the desired correct knowledge or skills, that does not necessarily mean the learning is lost forever. Knowledge that is forgotten at one point in time can sometimes spontaneously be retrieved later (a phenomenon known as *hypermnesia*; Erdelyi & Becker, 1974, EL: 3), demonstrating that memories can be temporarily inaccessible without being permanently lost. A classic example is the *tip-of-the-tongue* phenomenon (e.g., Burke, MacKay, Worthley, & Wade, 1991, EL: 5), when one has a sense of knowing a particular word or name but being unable to retrieve it--only to spontaneously recover it later on. Thus,

failure to retrieve an idea at any point in time is not necessarily--or even likely--to be reflective of permanent loss.

The fact that inaccessible knowledge and skills are not fully lost also leads to *savings* in that previously-encountered knowledge can be re-learned more quickly than when it was initially acquired (Ebbinghaus, 1885, EL: 4; Nelson, 1978, EL: 4). Although inaccessible knowledge can sometimes be retrieved spontaneously, it is more apt to be retrieved with appropriate *retrieval cues* (e.g., Tullis & Benjamin, 2015, EL: 2; Tullis & Fraundorf, 2017, EL: 3), characteristics of the environment that help to "jog" one's memory (although certain cues can be unhelpful if they disrupt a planned retrieval strategy; Basden & Basden, 1995, EL: 3; Roediger, 1978, EL: 3). In general, human memory is partially context-dependent, such that memories more readily come to mind when the environment matches how they were initially learned or acquired (Bjork & Richardson-Klavehn, 1989, EL: 2). Thus, a change in perspective can bring to mind memories that had been inaccessible in previous moments (Anderson & Pichert, 1978, EL: 3; Fraundorf & Benjamin, 2016, EL: 3). Therefore, unused skills can be brought to mind or re-learned more quickly (i.e., savings), especially with the right set of external cues. More frequent longitudinal assessment, rather than point-in-time assessment, could serve as a cue to keep this knowledge accessible and/or facilitate re-learning.

Aging and its Relationship to Memory and Learning

Beyond the time that has elapsed since medical training, another source of skill decline may be aging. These two factors are highly confounded among physicians insofar as most physicians enter medical school at roughly similar ages. We initially cover the basic science of aging which has elucidated how learning and memory change across the lifespan using both cross-sectional and longitudinal designs, and then address aging more specifically as it relates to physicians.

Age Affects Some Cognitive Skills More Than Others

A clear conclusion from the basic science of memory aging is that age differentially affects different types of knowledge. Beginning in early adulthood (e.g., the 20s), performance steadily declines with age on tasks that require *fluid intelligence*; that is, those that involve novel learning or reasoning (Horn & Cattell, 1966, EL: 5; Horn & Cattell, 1967, EL: 5). This decline may be driven at least in part by declines in more fundamental aspects of cognition: The speed of even very basic cognitive processing (e.g., as measured by the speed of identifying whether two strings of letters are the same or different) declines with age, as does the ability to temporarily hold information in *working memory* (Park, Lautenschalger, Hedden, Davidson, A. Smith, & P. Smith, 2002, EL: 5; Salthouse, 1991, EL: 5; Salthouse, 1996, EL: 2; Salthouse, 2004, EL: 2; Salthouse, 2005, EL: 5; Salthouse & Babcock, 1991, EL: 5; Stine-Morrow, Soederberg Miller, Gagne, & Hertzog, 2008, EL: 5). For instance, declines in basic speed may drive age differences in more complex tasks insofar as cognitive skills may break down if people cannot retrieve or compute relevant information sufficiently quickly to be useful for the task at hand (Hertzog, Dixon, Hultsch, & MacDonald, 2003, EL: 5; Salthouse, 1991, EL: 5; Salthouse, 1996, EL: 2; Salthouse & Babcock, 1991, EL: 5; Salthouse, 2005, EL: 5; Stine-Morrow et al., 2008, EL: 5). Decline in processing speed is relevant to many areas of medicine as physicians often see high volumes of patients in a given

day, need to address multiple problems per visit, and, in some settings, need to switch quickly between patients. It is not enough simply to have acquired the relevant cognitive skills; physicians need to be able to bring to mind--or know where to look up--the relevant knowledge in time to be practically useful.

By contrast, fixed knowledge, often referred to as *crystallized intelligence*, is preserved or even increases with age (Horn & Cattell, 1966, EL: 5; Horn & Cattell, 1967, EL: 5; Park et al., 2002, EL: 5; Salthouse, 2004, EL: 2; Zacks & Hasher, 2006, EL: 2). Even fixed knowledge may decline at especially advanced ages (e.g., the 80s or above; Park et al., 2002, EL: 5; Salthouse, 2004, EL: 2), but physicians would likely be retired at this age. In general, then, older adults rely less on novel (fluid) episodic learning and more on existing (crystallized) knowledge about the world (Castel, 2005, EL: 4; Castel, 2007, EL: 4; Castel, McGillivray & Worden, 2013, EL: 3; Koutstaal & Schacter, 1997, EL: 3; McGillivray & Castel, 2017, EL: 3; Stine-Morrow et al., 2008, EL: 4; Zacks & Hasher, 2006, EL: 3). This has mixed implications for the retention and use of medical expertise: On the one hand, physicians' general medical knowledge might be expected to be relatively spared with age. On the other hand, older physicians might be less proficient at learning new techniques or remembering specific newly encountered cases and patients.

Further, although older adults underperform younger adults even in very basic memory tasks, age differences are larger in some types of learning and retrieval than others (Fraundorf, Hourihan, Peters, & Benjamin, 2019, EL: 1). For instance, it has been argued that older adults are especially challenged by cognitive skills that require self-initiated or controlled processing, such as deliberately committing novel information to memory (e.g., learning new standards of care) or systematically reviewing one's memory (e.g., deliberately considering each of a series of potential diagnoses). By comparison, age is less deleterious for relatively automatic or habitual uses of memory, such as applying a familiar set of actions (e.g., ordering a frequent diagnostic test) or recognizing a stimulus (e.g., a familiar set of symptoms as a particular disease) (e.g., Craik, 1986, EL: 2; Hoyer & Verhaeghen, 2006; EL: 2; Luo & Craik, 2008, EL: 2; c.f., Fraundorf et al., 2019, EL: 1). This pattern is consistent with age-related declines in the controlled, analytical System 2 but preserved or enhanced functioning of the automatic, experience-based System 1 (Eva, 2002, EL: 2 ; Eva, 2003, EL: 2). It suggests that older physicians may rely heavily on habitual, rather than new, cognitive skills and that they will likely remember patients and treatments consistent with their general experience.

Several other generalizations regarding memory in aging highlight other situations where older physicians' cognitive skills might be preserved. First, older adults perform comparatively well at remembering new information that is naturalistic (as opposed to arbitrary laboratory stimuli; Castel, 2007, EL: 4) or that allows the use of existing everyday memory strategies, such as establishing routines or leaving reminders for oneself (Bailey, Henry, Rendell, Phillips & Kliegel, 2010, EL: 4; Moscovitch, 1982, EL: 5; Rendell & Craik, 2000, EL: 3; Rendell & Thomson, 1999, EL: 3). Second, older adults are *as* or even *more* effective than younger adults at working with familiar partners to remember information as a team. These *collaborative cognition* strategies can include dividing responsibilities for remembering different kinds of information and suggesting cues to "jog" each other's memory (*collaborative cognition*; Dixon & Gould, 1996, EL: 3; Dixon, 1999, EL: 2). Third, older adults are sensitive to indicators of the *value* of to-be-retained information and perform comparatively well in remembering material that is important or that otherwise aligns with their motivational priorities. For instance, in laboratory experiments, older adults are adept at prioritizing material that is worth more "points" towards a goal (Castel, 2007, EL: 3; Castel, Benjamin, Craik, & Watkins, 2002, EL: 3; Castel, Farb, & Craik, 2007, EL: 3),

that a speaker emphasizes intentionally (Fraundorf, Watson, & Benjamin, 2012, EL: 4), that aligns with a motivational bias for positivity (Charles, Mather, & Carstensen, 2003, EL: 3; Mather & Carstensen, 2005, EL: 3; May, Rahhal, Berry, & Leighton, 2005, EL: 3), or that comes from a more trustworthy source (Rahhal, May, & Hasher, 2002, EL: 3). Indeed, even non-physician older adults better remember fictive medications with severe side effects than those with less severe side effects (Hargis & Castel, 2018, EL: 3). All three of these age-related changes would be expected to favor retention of medical skill and learning even with increasing age insofar as physicians use their medical expertise in everyday life, often work with well-established teams, and (presumably) value their medical knowledge and skills.

However, there is clear meta-analytic evidence that older adults are especially impaired in remembering the *source* or *context* of information (Fraundorf et al., 2019, EL: 1; Old & Naveh-Benjamin, 2008, EL: 1; Spencer & Raz, 1995, EL: 1). This could have deleterious consequences in medicine if older physicians confuse or misattribute the symptoms or treatments prescribed to several patients they have recently seen. In general, there is reason to be optimistic that physicians may retain much of their general medical knowledge with increasing age. However, older physicians may be vulnerable to reduced memory for specific cases or patients, and their speed of access to their knowledge may also decline.

Aging as it Relates to Physicians

The role of aging in physicians' cognitive skills has been addressed in a number of narrative reviews (e.g., Eva, 2002; 2003; Durning et al., 2010; Williams, 2006; EL: 2). Assessing the role of age in a physician's ability to provide high quality care is quite complex because there are many factors that interact with one another that it is difficult to distinguish them. First, it is possible that a physician's abilities decline with age due to memory or processing decline. Second, with the passage of time since medical school and residency, knowledge and skills may decline due to interference. Third, with the passage of time since residency, a physician's knowledge becomes out of date due to shifting standards that they did not learn in medical school or residency. Fourth, over time a physician accumulates more direct patient experience, which, as explained above, can have both positive and potentially negative impacts on performance. Fifth, some physicians specialize over time, which could lead them to lose the broader skills that have become less relevant to their specific practice.

Another reason that it is hard to understand the role of age in a physician's ability to provide high quality care has to do with the complicated nature of the dual-process theory of decision making. The dual-process theory predicts that non-analytical reasoning should remain intact even though analytical reasoning processes may decline. Indeed, since physicians accumulate more experience over time, dual-process theory predicts that aspects of decision making that are especially driven by non-analytical decision making may in fact improve with experience (Eva, 2002). Reliable evidence indicates that the quality of healthcare provided decreases with the age of the physician. A systematic review of 62 studies found that 45 studies (73%) reported a decrease in performance for some or all outcomes (Choudhry, Fletcher, Soumerai, 2005, EL: 2). Another 13 (21%) found no association. The remaining four found a non-linear (inverted U) trend or an increase in some or all outcomes. This pattern held across a wide variety of measures, including knowledge measures, health outcomes, and adherence to standards of care for diagnosis, screening, prevention, and therapy. However, one potential limitation of this

review is that it covered studies from a period of time during which evidence-based medicine and quality assurance techniques, such as performance evaluation, became widely adopted. So, the apparent age-related declines may instead be driven by the fact that the older physicians were trained prior to this shift. It is possible that the newer generation of physicians may not exhibit declines in quality of care as they age if they are able to stay up to date with the evidence through more innovative mechanisms.

Though an updated systematic review has not been conducted since 2005, a number of notable recent studies reinforce this pattern of decreases in quality of care provided by older physicians. A population-based study of adherence to guidelines for antibiotic prescribing in treating urinary tract infections in children in Taiwan found that adherence dropped gradually from 87% in physicians younger than 35 to 45% in physicians older than 55 (Chen, Wu, Li, Liu, Woung, et al., 2011, EL: 5). Holmboe et al. (2008, EL: 5) found that physicians more than 20 years out of medical school performed considerably worse on the maintenance of certification assessment compared to physicians fewer than 20 years out. Physicians who scored lower on the assessment also exhibited worse performance on measures of treating patients: whether they had diabetes patients obtain eye exams, lipid tests, and HbA1c tests, whether they had female patients receive a mammogram in the past year, and whether they had patients with coronary artery disease obtain a lipid test in the past year. St-Onge et al. (2016, EL: 5) similarly found worse diagnostic performance for clinical vignettes among older physicians. In sum, there does seem to be some reliable evidence from multiple sources of a general decrease in both conceptual knowledge and quality of care with increased age.

However, the specific reasons for the decreasing quality of care with age are less certain. One possibility, reviewed by Eva (2002, EL: 2) and also discussed in the prior section on *Basic Science of Aging*, is that performance decline may be caused by negative age-related changes in cognitive processing, particularly related to working memory and fluid intelligence. Another possibility is that older physicians fail to learn and/or retain changing standards of care. In fact, another study of scores on a certification assessment found that age predicted poorer performance on questions that tested knowledge for standards of care that had changed over the preceding 30 years, but age did not predict poorer performance on questions about standards of care that had not changed (Day et al., 1988, EL: 5; see also Holmboe, Lipner, & Greiner, 2008). However, this study is quite dated, and it is not certain that this finding would still hold among the current cohort of physicians. Conducting more studies of this nature could help elucidate the mechanisms of the apparent decreases in knowledge and performance with age.

Despite the evidence of declines in performance with age, there is other more mixed evidence. First, there are reasons to think that System 1 (non-analytical processing or pattern recognition) may remain stable or even improve with age and experience. In the previous section, we discussed how more automatic forms of memory or habitual responses, as well as fixed knowledge, remain intact until advanced ages of 80 or above (Eva, 2002, EL: 2). Being able to continue to rely on automatic forms of memory affirms the important role of non-analytical medical decision making. Second, the empirical studies on diagnosis present a mixed picture. Some studies have found that older physicians tend to both identify correct diagnoses very quickly and settle on a diagnosis quickly. This is a double-edged sword. On the one hand, it can lead to quick and accurate diagnoses. For example, Hobus et al. (1993, EL: 5; see discussion in Eva, 2002) found a positive relation between years of experience and diagnostic

accuracy, and Eva et al. (2010, EL: 5) also found greater diagnostic accuracy with age/experience. However, quickly settling on a diagnosis can also lead to premature closure (i.e., failing to consider alternatives after reaching a decision), and some research has found that older physicians focus more heavily on information presented earlier in a case (Eva & Cunningham, 2006, EL: 5).

In sum, the majority of the evidence suggests that older physicians perform worse in a variety of ways, even though gaining experience over one's career may mitigate this decline to some extent. However, physicians are tasked not only with maintaining current standards of care but also keeping up with changing standards, which we discuss in the next section.

Keeping up with Changing Standards of Care

One of the fundamental challenges in medicine is keeping up with the ever-changing standards of care. The Institute of Medicine report on *Best Care at Lower Cost: The Path to Continuously Learning Health Care in America* (McGinnis, Stuckhardt, Saunders, & Smith, 2013) concluded that diagnostic and treatment options are changing at an accelerating rate, making it ever more important to keep up with changing standards.

Two major reviews of 120 and 256 articles, respectively, have systematized the barriers to using current standards of care (Cabana et al., 1999, EL: 2; Cochrane et al., 2007, EL: 2). These reviews cover a number of broader barriers, such as systems-level and patient-level barriers that are less relevant here, and they also cover somewhat different subsets of barriers and use somewhat different terminology. Nevertheless, there is substantial agreement in terms of the cognitive and attitudinal barriers identified: familiarity, outcome expectancy, agreement, self-efficacy, and habits. In fact, there is a roughly temporal order to this list. Imagine a physician learning about a new treatment standard of care. First, the physician must become *familiar* with and *aware* of this new standard of care. Second, the physician must develop knowledge or skill, for example, knowledge about indications and dosages of a therapy. Third, the physician must form a high *outcome expectancy* and *agree* with the new standard of care--believe that following this new standard is more likely to lead to better outcomes than not following it. Fourth, the physician must feel confident that they can implement the standard, termed *self-efficacy*. Fifth, the physician must overcome *habits* or *inertia*, doing things the same way as they have always been done. Both habits and awareness are potentially related to the phenomenon of *proactive interference*, which was discussed previously in this chapter and will soon be discussed in the upcoming chapter *Testing Enhances Learning and Retention*. Simply because a physician has been exposed to a new practice, agrees with it, and feels that they can implement it, does not mean that they will always spontaneously remember to use it. Many years of doing something differently will make it harder to remember the new practice the next time it is called for.

How might a longitudinal assessment program affect these barriers? In the chapter *Testing Enhances Learning and Retention*, we will discuss the overwhelming evidence that repeated testing benefits learning and retention (Adesope, Trevisan, & Sundararajan, 2017, EL: 1; Rowland, 2014, EL: 1). As such, it is reasonable to assume that an intervention that includes regular testing events would be effective at reducing barriers to keeping up with changing standards of care that are due to familiarity, awareness, and knowledge. A longitudinal assessment program may also aid in ameliorating interference from previously learned practices. As we discuss in the chapter *Testing Enhances Learning*

and Retention, retrieving information during testing can increase its distinctiveness (Kuo & Hirshman, 1997, EL: 3; Peterson & Mulligan, 2013, EL: 3; Lehman, Smith, & Karpicke, 2014, EL: 4) and therefore less likely to be confused with other, related information. Further, because cognitive skills that are less frequently used may be particularly vulnerable to interference, longitudinal assessment can directly target these skills by targeting those areas more frequently.

It is less clear whether and how longitudinal assessment could impact the attitudinal barriers, such as outcome expectancy (not believing that the standard of care or guideline would be beneficial) and self-efficacy (not believing that one actually has the ability to implement the practice), which are not the primary goals of an assessment program. Nevertheless, it is possible that feedback to learners--with correct answers, explanations of the answers, and citations to resources to learn more--might be able to change opinions about whether the standard is beneficial and might make a physician feel that they have enough knowledge to implement the standard in person.

Chapter Summary

This chapter covered four main processes. First, we reviewed acquiring medical expertise and the dual-process theory of medical expertise. This theory proposes that medical decision making is a combination of fast intuitive thinking (non-analytical process) that is shaped by experience and slow analytical thinking guided by logic. According to this theory, the benefits of non-analytical thinking are that decisions can be reached very fast and are often correct. However, intuitive, non-analytical thinking also has downsides. Idiosyncratic experiences can shape a physician's decisions (e.g., about treatment decisions). And, idiosyncrasies of the illnesses of patients a physician does and does not see affect the maintenance of expertise; for instance, they could distort the physician's beliefs about the prevalence of a diagnosis and the likelihood of that diagnosis coming to mind.

Second, we discussed the basic science of maintaining expertise. One conclusion from this section is that a failure to bring relevant information to mind does not necessarily indicate it is permanently lost. Rather, it may become accessible again later, especially with the right internal or external cues. One important cause of retrieval failures is the inability to access knowledge quickly enough to be useful. Another is interference from competing concepts and skills. Indeed, interference may be inevitable insofar as repeatedly retrieving some information in response to a cue weakens other information related to that cue. This means that it is likely that repeatedly experiencing more common diagnoses (e.g., myocardial infarction) will reduce the likelihood of thinking about rare diagnoses (e.g., aortic dissection) when faced with a cue (e.g., chest pain). The literature both on interference and on analytical thinking suggest that a longitudinal continuing certification program can attempt to fill in potential gaps in experience by testing clinical cases that are somewhat rare but of high importance to patient care.

Third, we reviewed barriers to keeping up with changing standards of care. At a physician level (rather than patient- or systems-level), barriers include not being aware of a new standard, lack of knowledge of the standard, not believing that the new standard is better, not being confident in how to implement the new standard, and habits. One goal of continuing certification programs has traditionally been to assess whether physicians are keeping up with changing standards. The benefit of a longitudinal

assessment program is that it may serve as both as an assessment and educational program to make physicians more knowledgeable about new standards and their application in patient care.

Fourth, the basic science of aging suggests that whereas crystallized intelligence (e.g., medical knowledge) remains intact until the 80s, fluid intelligence (e.g., novel learning or using balancing multiple tasks in working memory) declines with age. The majority of the evidence suggests that the quality of healthcare declines with physician age. This could be driven in part by the decline in fluid intelligence or a failure to keep up with changing standards of care.

Future Directions

In reviewing this literature, we identified a number of directions for future research that were specifically lacking in the literature or not seen applied to the physician population of continuing certification. These include five proposed studies where we supply the ideas but not necessarily a fully developed research design.

1. What is the most Useful Feedback to Emphasize New Standards of Care?

Two benefits of a longitudinal assessment program are that it can potentially help physicians learn about standards of care that have changed since their training and that it can provide useful feedback to physicians about their relative strengths and weaknesses. We propose that boards prospectively classify items as testing new standards of care versus testing old--but still relevant--standards of care. (Alternatively, boards can try to classify items according to when the relevant standard of care emerged, though this may be more difficult to implement) This classification can then be used in three ways.

First, feedback about performance relative to the guidelines before and after its change will provide physicians with a sense of whether they are challenged more by staying current versus by maintaining older knowledge. Second, we propose that boards use an approach pioneered in a clever study conducted by the American Board of Internal Medicine (Day et al., 1988) to track the efficacy of the longitudinal assessment programs. This study looked at older versus younger physicians' performance on the continuing certification program assessment, contrasting questions for which standards of care have changed over time versus questions for which they have not. The finding was that older physicians performed worse for questions testing knowledge about standards that had changed, but not for standards that had remained the same. Ideally, if the assessment program works to keep physicians up to date, this interaction for older vs. younger physicians on changed vs. unchanged standards should diminish over time. This analysis could be conducted both before and after implementing the longitudinal assessment program to evaluate the contributions of the educational component of the longitudinal program. And, it could be conducted on an ongoing basis to measure the success of the program over time, with the goal of continuously optimizing the test to minimize the difference between older and younger physicians. Third, extending this analysis pioneered by Day et al. (1988) can help to uncover the reasons for poor performance. In particular, Day et al. found decreases in performance over time only on questions for which standards of care have changed over time, which seems to implicate challenges of staying current rather than aging or time since residency per se. However, since 1988, the landscape of CME has changed considerably, so it is not clear whether the

same pattern would be found. Furthermore, one possibility is that physicians selectively keep up with standards that they think are especially relevant to their practice. This possibility could be assessed by having physicians rate the relevance of each question, and testing whether relevance interacts with age and whether or not a standard has changed. Still other analyses would be possible if physicians also rate their confidence in their answers. (See the next chapter "Self-Assessment Is Not Enough" for more discussions on confidence.) For example, one possibility is that if a physician is wrong on a question that involves a new standard, but is highly confident, that might mean old knowledge is interfering with learning new knowledge or that they never learned the new standards. In contrast, if a physician is wrong but not very confident on a question that involves a new standard, that might mean that they have heard of the new standard but haven't fully learned it.

2. What is the Relationship between Response Time, Performance, and Age in Assessments?

A question that could easily be investigated in assessment programs has to do with the relationship between response times, performance, and age. Some prior research has examined the relationships between age, response time, accuracy, and case difficulty (Barrows et al., 1982, EL: 5; Elstein, Shulman, and Sprafka, 1978, EL:5; Gruppen, Woolliscroft, & Wolf, 1988, EL: 5; Norman et al., 1989, EL: 5; Pelaccia, et al., 2014, EL: 5). However, these findings are nuanced and not entirely consistent, and therefore this research could greatly benefit from broader case materials and from larger, more representative samples of physicians across many specialties. In fact, given many items on continuing certification program assessments already incorporate questions about diagnosis and treatment, and gather response time in order to address these questions.

3. *Can Small Interventions Help with Improving Clinical Reasoning Skills?*

The use of longitudinal assessment programs as a platform to test clinical reasoning interventions is feasible. In particular, the dual-process theory presumes that there are two modes of reasoning, one that is faster and one that is slower, and that these modes need to be coordinated (Croskerry, 2009a, EL: 2; 2009b, EL: 2; Norman and Eva, 2010, EL: 2; Pelaccia et al., 2011, EL: 2). For example, one proposal is that physicians need to learn how to switch from faster automatic judgment for routine problems to slower, more effortful reasoning for more unusual or ill-defined problems (Moulton et al., 2007; see also Graber, 2009). Pat Croskerry and colleagues have provided thoughtful overviews of potential ways to attempt to teach physicians to avoid common biases (e.g., Croskerry, Singhal, & Mamede 2013a, 2013b).

Though a number of clinical reasoning interventions have been proposed and tested, there is only mixed evidence whether clinical reasoning interventions work (e.g., Schmidt & Mamede, 2015, EL: 2; see also Isler, Yilmaz, and Dogruyol, 2020). Still, the Institute of Medicine report on Improving Diagnosis in Health Care still expresses interest in testing such interventions (National Academies of Sciences, Engineering, & Medicine, 2015, pp. 4-32 to 4-34). Longitudinal assessment programs could serve as a testing ground for brief interventions that could be embedded right before or during a question.

4. What Impact does Non-Analytical Reasoning have on Arriving at a Correct Diagnosis?

The studies proposed above would have direct application to the design of a longitudinal assessment program. We have also identified three opportunities that instead are aimed at further uncovering the basic mechanisms of non-analytical reasoning in medical decision making. These could, in turn, provide insights into a longitudinal assessment program. All three of these would be fairly easy to embed inside a longitudinal assessment, and some of them simply require data analysis without extensive planning.

We propose that experiments on the role of non-analytical reasoning in diagnosis can be directly embedded into a longitudinal assessment program, and by doing so, will help to firmly establish an empirical base of knowledge regarding non-analytical reasoning in medicine. The research on non-analytical reasoning in medicine has only been tested in a few studies with only modest sample sizes (Brooks et al., 1991, EL: 4; Hatala, Norman, & Brooks, 1999, EL: 4; Young, Brooks, & Norman, 2011, EL: 4). Though these findings are intuitive, and though they build upon an extensive literature on the basic science of categorization from cognitive science (Cohen & Lefebvre, 2017), there are a number of important gaps in knowledge. First, most of the basic science research has been conducted with abstract stimuli, and with undergraduates who are trained for only short periods of time, not complex real-world medical stimuli that require years for physicians to master. Thus, conducting larger studies with physicians will help to establish these phenomena within medicine with more certainty.

It is important to understand how big of an effect that irrelevant information has in biasing diagnosis. A longitudinal assessment provides a remarkable opportunity to design studies embedded into the programs to test the role of irrelevant information from prior cases in biasing the diagnosis of future cases. This research could test how long the bias lasts, how strong the bias is, how the prevalence of a disease or a physician's knowledge of a disease affects the bias, and whether age of the physician interacts with non-analytical reasoning. Further, given that most of the literature on non-analytical reasoning in diagnosis has focused on diagnosis based on visual information (e.g., reading ECGs, pathology, dermatology, radiology), another question is whether the bias is different for diagnosis that requires integrating multiple signs, symptoms, and lab reports (e.g., emergency medicine, internal medicine, etc.; Norman, Young, & Brooks, 2007). However, because of the few number of studies, and the fact that they have primarily focused on highly visual domains (dermatology and ECGs), it is not clear to what extent non-analytical reasoning is involved in less perceptual types of diagnosis.

Although there is broad consensus that non-analytic reasoning plays some role in diagnosis, we know comparatively little about *when*, *where*, and *how much* it matters. Delineating how experience both bolsters and biases subsequent decisions could help make physicians more aware of how such non-analytic factors could impact decision making, which could serve as a motivation to follow evidence-based guidelines or potentially be used in debiasing efforts.

5. *What is the Effect of Feedback on Outcome Expectancy and Self-Efficacy?*

Another opportunity is to study the feedback provided that explains whether an answer is right or wrong. Among the multiple barriers to keeping up with changing standards of care, the main barrier that a longitudinal assessment program is intended to address is awareness, familiarity, and knowledge about the new standard of care. However, it is possible that if written in the right way, the feedback

could also address other barriers, such as not believing that the new standard is better (outcome expectancy) and not being confident in how to implement it (self-efficacy). Experiments could be designed that manipulate the provided feedback, and questions could be embedded about a physician's feelings of outcome expectancy and self-efficacy in order to test ways to maximize the broader utility of the feedback for overcoming multiple barriers.

In summary, there are a number of potential ways that data collected from longitudinal assessment programs, or from interventions embedded inside the programs, could serve as ways to assess the efficacy of the programs themselves, provide guidance about how to improve feedback, and test basic questions about medical expertise and diagnosis that are hard to study in other settings and for which studies are rare and small. Many of these proposals can be accomplished with minimal changes to the duration of the program, and physicians may find them insightful if the results can demonstrate strong evidence regarding the roles of aging and keeping up with standards of care, speed versus accuracy, and debiasing techniques.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451-474.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063-1087.

Anderson, R. C., & Pichert, J. W. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, *17*(1), 1-12.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General*, *128*(2), 186-197.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396-408.

Bahrick, H. P. (1983). The cognitive map of a city: Fifty years of learning and memory. The Psychology of learning and motivation, *17*, 125-163.

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General*, *104*(1), 54-75.

Bailey, P. E., Henry, J. D., Rendell, P. G., Phillips, L. H., & Kliegel, M. (2010). Dismantling the "age–prospective memory paradox": The classic laboratory paradigm simulated in a naturalistic setting. *Quarterly Journal of Experimental Psychology*, *63*(4), 646-652.

Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. Clinical and investigative medicine. *Clinical and Investigative Medicine*, *5*(1), 49-55.

Basden, D. R., & Basden, B. H. (1995). Some tests of the strategy disruption interpretation of part-list cuing inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(6), 1656.

Bjork, R. A., & Richardson-Klavehn, A. (1989). On the puzzling relationship between environmental context and human memory. In C. Izawa (Ed.), *Current issues in cognitive processes: The Tulane Flowerree Symposium on Cognition* (p. 313–344). Lawrence Erlbaum Associates, Inc.

Brooks, L. R., Norman, G. R., & Allen, S. W. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General*, *120*(3), 278-287.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, 1116.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults?. *Journal of Memory and Language*, *30*(5), 542-579.

Cabana, M. D., Rand, C. S., Powe, N. R., Wu, A. W., Wilson, M. H., Abboud, P. A. C., & Rubin, H. R. (1999). Why don't physicians follow clinical practice guidelines?: A framework for improvement. *Jama*, *282*(15), 1458-1465.

Casscells, W., Schoenberger, A., & Graboys, T. B. (1978). Interpretation by physicians of clinical laboratory tests. *The New England Journal of Medicine*, *299*(18), 999–1001.

Castel, A. D. (2005). Memory for grocery prices in younger and older adults: The role of schematic support. *Psychology and Aging*, *20*(4), 718-721.

Castel, A. D. (2007). The adaptive and strategic use of memory by older adults: Evaluative processing and value-directed remembering. *Psychology of Learning and Motivation*, *48*, 225-270.

Castel, A. D., Benjamin, A. S., Craik, F. I., & Watkins, M. J. (2002). The effects of aging on selectivity and control in short-term recall. *Memory & Cognition*, *30*(7), 1078-1085.

Castel, A. D., Farb, N. A., & Craik, F. I. (2007). Memory for general and specific value information in younger and older adults: Measuring the limits of strategic control. *Memory & Cognition*, *35*(4), 689-700.

Castel, A. D., McGillivray, S., & Worden, K. M. (2013). Back to the future: Past and future era-based schematic support and associative memory for prices in younger and older adults. *Psychology and Aging*, *28*(4), 996-1003.

Charles, S. T., Mather, M., & Carstensen, L. L. (2003). Aging and emotional memory: the forgettable nature of negative images for older adults. *Journal of Experimental Psychology: General*, *132*(2), 310-324.

Chen, C. C., Wu, L. C., Li, C. Y., Liu, C. K., Woung, L. C., & Ko, M. C. (2011). Non-adherence to antibiotic prescription guidelines in treating urinary tract infection of children: A population-based study in Taiwan. *Journal of Evaluation in Clinical Practice*, *17*(6), 1030-1035.

Choudhry, N. K., Anderson, G. M., Laupacis, A., Ross-Degnan, D., Normand, S. L. T., & Soumerai, S. B. (2006). Impact of adverse events on prescribing warfarin in patients with atrial fibrillation: Matched pair analysis. *BMJ*, *332*(7534), 141-145.

Choudhry, N. K., Fletcher, R. H., & Soumerai, S. B. (2005). Systematic review: The relationship between clinical experience and quality of health care. *Annals of Internal Medicine*, *142*(4), 260-273.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, *5*(2), 121-152.

Cochrane, L. J., Olson, C. A., Murray, S., Dupuis, M., Tooman, T., & Hayes, S. (2007). Gaps between knowing and doing: understanding and assessing the barriers to optimal health care. *Journal of Continuing Education in the Health Professions*, *27*(2), 94-102.

Cohen, H., & Lefebvre, C. (Eds.). (2017). Handbook of categorization in cognitive science (2nd edition). Elsevier.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why?. *Current Directions in Psychological Science*, *19*(1), 51-57.

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Flix & H. Hagendorf (Eds.), *Human memory and cognitive capabilities: Mechanisms and performances* (pp. 409–422). Amsterdam, the Netherlands: North-Holland.

Croskerry, P. (2009a). Clinical cognition and diagnostic error: Applications of a dual process model of reasoning. *Advances in Health Sciences Education*, *14*(1), 27-35.

Croskerry, P. (2009b). A universal model of diagnostic reasoning. *Academic Medicine*, *84*(8), 1022–1028.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality & Safety*, *22*(Suppl 2), ii58-ii64.

Croskerry, P., Singhal, G., & Mamede, S. (2013). Cognitive debiasing 2: Impediments to and strategies for change. BMJ quality & safety, 22(Suppl 2), ii65-ii72.

Day, S. C., Norcini, J. J., Webster, G. D., Viner, E. D., & Chirico, A. M. (1988). The effect of changes in medical knowledge on examination performance at the time of recertification. In *Research in medical education: Proceedings of the Annual Conference on Research in Medical Education* (Vol. 27, p. 139-144).

Dixon, R. A. (1999). Exploring cognition in interactive situations: The aging of *N*+1 minds. In T. M. Hess & F. Blanchard-Fields (Eds.), *Social cognition and aging* (pp. 267-290). Elsevier.

Dixon, R. A., & Gould, O. N. (1996). Adults telling and retelling stories collaboratively. In P. B. Baltes & U. M. Staudinger (Eds.), *Interactive minds: Life-span perspectives on the social foundation of cognition* (pp. 221–241). Cambridge University Press.

Drachman, D. A. (2005). Do we have brain to spare?. *Neurology*, *64*(12), 2004-2005.

Durning, S. J., Artino, A. R., Holmboe, E., Beckman, T. J., van der Vleuten, C., & Schuwirth, L. (2010). Aging and cognitive performance: challenges and implications for physicians practicing in the 21st century. *Journal of Continuing Education in the Health Professions*, *30*(3), 153-160.

Ebbinghaus, H. (1885). Ueber das Gedächtnis.

Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases (pp. 249–267). Cambridge, UK: Cambridge University Press.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. Medical problem solving: An analysis of clinical reasoning. 1978. In :. Harvard University Press.

Erdelyi, M. H., & Becker, J. (1974). Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials. *Cognitive Psychology*, *6*(1), 159-171.

Ericsson, K. A. (2015). Acquisition and maintenance of medical expertise: a perspective from the expert-performance approach with deliberate practice. *Academic Medicine*, *90*(11), 1471-1486.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363-406.

Eva, K. W. (2002). The aging physician: Changes in cognitive processing and their impact on medical practice. *Academic Medicine*, *77*(10), S1-S6.

Eva, K. W. (2003). Stemming the tide: Cognitive aging theories and their implications for continuing education in the health professions. *Journal of Continuing Education in the Health Professions*, *23*(3), 133-140.

Eva, K. W., & Cunnington, J. P. (2006). The difficulty with experience: does practice increase susceptibility to premature closure? *Journal of Continuing Education in the Health Professions*, *26*(3), 192-198.

Eva, K. W., Link, C. L., Lutfey, K. E., & McKinlay, J. B. (2010). Swapping horses midstream: factors related to physicians' changing their minds about a diagnosis. *Academic Medicine: Journal of the Association of American Medical Colleges*, *85*(7), 1112.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.

Fraundorf, S. H., & Benjamin, A. S. (2016). Conflict and metacognitive control: the mismatch-monitoring hypothesis of how others' knowledge states affect recall. *Memory*, *24*(8), 1108-1122.

Fraundorf, S. H., Hourihan, K. L., Peters, R. A., & Benjamin, A. S. (2019). Aging and recognition memory: A meta-analysis. *Psychological Bulletin*, *145*(4), 339-371.

Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2012). The effects of age on the strategic use of pitch accents in memory for discourse: A processing-resource account. *Psychology and Aging*, *27*(1), 88-98.

Graber, M. L. (2009). Educational strategies to reduce diagnostic error: can you teach this stuff?. *Advances in Health Sciences Education*, *14*(1), 63-69.

Gruppen, L. D., Woolliscroft, J. O., & Wolf, F. M. (1988). The contribution of different components of the clinical encounter in generating and eliminating diagnostic hypotheses. In *Research in medical education: Proceedings of the annual conference on research in medical education* (Vol. 27, p. 242-247).

Hatala, R., Norman, G. R., & Brooks, L. R. (1999). Influence of a single example on subsequent electrocardiogram interpretation. *Teaching and Learning in Medicine*, *11*(2), 110-117.

Hargis, M. B., & Castel, A. D. (2018). Younger and older adults' associative memory for medication interactions of varying severity. *Memory*, *26*(8), 1151-1158.

Hertzog, C., Dixon, R. A., Hultsch, D. F., & MacDonald, S. W. S. (2003). Latent change models of adult cognition: Are changes in processing speed and working memory associated with changes in episodic memory? *Psychology and Aging*, *18*, 755-769.

Hobus, P. P. M., Schmidt, H.G. (1993). In Schmidt, H. G. The encapsulation framework in the presentation of physicians' recall of clinical cases. Presented at the annual meeting of the American Educational Research Association. Seattle, Washington, April 10-14, 2001.

Holmboe, E. S., Lipner, R., & Greiner, A. (2008). Assessing quality of care: Knowledge matters. *JAMA*, *299*(3), 338-340.

Horn, J. L., & Cattell, R. B. (1966). Age differences in primary mental ability factors. *Journal of Gerontology*, *21*(2), 210-220.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica*, *26*, 107-129.

Hoyer, W. J., & Verhaeghen, P. (2006). Memory aging. In J. Birren & K. Schaie (Eds.), *Handbook of the psychology of aging* (6th ed., pp. 209–232). New York, NY: Elsevier.

Isler, O., Yilmaz, O., & Dogruyol, B. (2020). Active reflective thinking with decision justification and debiasing training. *Judgment and Decision Making, 15*(6), 926-938.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515-526.

Kuo, T. M., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory and Language*, *36*(2), 188-201.

Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language*, *37*(4), 555-583.

Landauer, T. K. (1986). How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, *10*(4), 477-493.

Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis: Symbolic logic, probability, and value theory aid our understanding of how physicians reason. *Science*, *130*, 9–21.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787-1794.

Luo, L., & Craik, F. I. (2008). Aging and memory: A cognitive approach. *The Canadian Journal of Psychiatry*, *53*, 346–353.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science*, *25*(8), 1608-1618.

Marewski, J. N., & Gigerenzer, G. (2012). Heuristic decision making in medicine. *Dialogues in Clinical Neuroscience*, *14*(1), 77-89.

Mather, M., & Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences*, *9*(10), 496-502.

May, C. P., Rahhal, T., Berry, E. M., & Leighton, E. A. (2005). Aging, source memory, and emotion. *Psychology and Aging*, *20*(4), 571-578.

McGillivray, S., & Castel, A. D. (2017). Older and younger adults' strategic control of metacognitive monitoring: The role of consequences, task experience, and prior knowledge. *Experimental Aging Research*, *43*(3), 233-256.

McGinnis, J. M., Stuckhardt, L., Saunders, R., & Smith, M. (Eds.). (2013). *Best care at lower cost: The path to continuously learning health care in America*. National Academies Press.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81-97.

Moneta-Koehler, L., Brown, A. M., Petrie, K. A., Evans, B. J., & Chalkley, R. (2017). The limitations of the GRE in predicting success in biomedical graduate school. *PloS One*, *12*(1), e0166742.

Moscovitch, M. (1982). A neuropsychological approach to perception and memory in normal and pathological aging. In *Aging and Cognitive Processes* (pp. 55-78). Springer, Boston, MA.

Moulton, C.E.., Regehr, G., Mylopoulos, M., & MacRae, H. M. (2007). Slowing down when you should: a new model of expert judgment. *Academic Medicine*, *82*(10), S109-S116.

National Academies of Sciences, Engineering, and Medicine. (2015). *Improving diagnosis in health care*. National Academies Press.

Nelson, T. O. (1978). Detecting small amounts of information in memory: Savings for nonrecognized items. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(5), 453-468.

Nilsson, H., Juslin, P., & Olsson, H. (2008). Exemplars in the mist: The cognitive substrate of the representativeness heuristic. Scandinavian Journal of Psychology, 49(3), 201-212.

Norman, G. R., & Eva, K. W. (2010). Diagnostic error and clinical reasoning. *Medical Education*, *44*(1), 94-100.

Norman, G. R., Rosenthal, D., Brooks, L. R., Allen, S. W., & Muzzin, L. J. (1989). The development of expertise in dermatology. *Archives of Dermatology*, *125*(8), 1063-1068.

Norman, G., Young, M. & Brooks, L. (2007). Non-analytical models of clinical reasoning: The role of experience. *Medical Education*, *41*(12), 1140-1145.

Old, S. R., & Naveh-Benjamin, M. (2008). Memory for people and their actions: Further evidence for an age-related associative deficit. *Psychology and Aging*, *23*(2), 467-472.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, *17*(2), 299-320.

Pauker, S., & Kassirer, J. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, *302*, 1109–1117.

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Bertrand, C., Dory, V., & Charlin, B. (2014). How and when do expert emergency physicians generate and evaluate diagnostic hypotheses? A qualitative study using head-mounted video cued-recall interviews. *Annals of Emergency Medicine*, *64*(6), 575-585.

Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Medical Education Online*, *16*(1), 5890.

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1287-1293.

Postman, L., & Underwood, B. J. (1973). Critical issues in interference theory. *Memory & Cognition*, *1*(1), 19-40.

Rahhal, T. A., May, C. P., & Hasher, L. (2002). Truth and character: Sources that older adults can remember. *Psychological Science*, *13*(2), 101-105.

Rendell, P. G., & Craik, F. I. (2000). Virtual week and actual week: Age-related differences in prospective memory. Applied Cognitive Psychology: *The Official Journal of the Society for Applied Research in Memory and Cognition*, *14*(7), S43-S62.

Rendell, P. G., & Thomson, D. M. (1999). Aging and prospective memory: Differences between naturalistic and laboratory tasks. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *54*(4), P256-P269.

Roediger, H. L. (1978). Recall as a self-limiting process. *Memory & Cognition*, *6*(1), 54-63.

Rottman, B. M., Prochaska, M. T., & Deaño, R. C. (2016). Bayesian reasoning in residents' preliminary diagnoses. *Cognitive Research: Principles and Implications*, *1*(5), 1-5.

Rottman, B. M. (2017). Physician Bayesian updating from personal beliefs about the base rate and likelihood ratio. *Memory & Cognition*, *45*(2), 270-280.

Rowland, C. A. (2014). The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review*, *103*(4), 734-760.

Sadeh, T., Ozubko, J. D., Winocur, G., & Moscovitch, M. (2014). How we forget may depend on how we remember. *Trends in Cognitive Sciences*, *18*(1), 26-36.

Salthouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science*, *2*, 179-183.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, *103*, 403-428.

Salthouse, T. A. (2004). What and when of cognitive aging. *Current Directions in Psychological Science*, *13*(4), 140-144.

Salthouse, T. A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology*, *19*, 532-545.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology*, *27*, 763-776.

Schiff, G. D. (2008). Minimizing diagnostic error: the importance of follow-up and feedback. *The American Journal of Medicine*, *121*(5), S38-S42.

Schmidt, H. G., & Mamede, S. (2015). How to improve the teaching of clinical reasoning: A narrative review and a proposal. *Medical Education*, *49*(10), 961-973.

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*, 156-163.

Spencer, W. D., & Raz, N. (1995). Differential effects of aging on memory for content and context: a meta-analysis. *Psychology and Aging*, *10*(4), 527-539.

St-Onge, C., Landry, M., Xhignesse, M., Voyer, G., Tremblay-Lavoie, S., Mamede, S., Schmidt, H. & Rikers, R. (2016). Age-related decline and diagnostic performance of more and less prevalent clinical cases. *Advances in Health Sciences Education*, *21*(3), 561-570.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, *25*, 207-222.

Standing, L., Conezio, J., & Haber, R. N. (1970). Perception and memory for pictures: single-trial learning of 2,500 visual stimuli. *Psychonomic Science*, *2*, 43-53.

Stine-Morrow, E. A., Soederberg Miller, L. M., Gagne, D. D., & Hertzog, C. (2008). Self-regulated reading in adulthood. *Psychology and Aging*, *23*(1), 131-153.

Tullis, J. G., & Benjamin, A. S. (2015). Cueing others' memories. *Memory & Cognition*, *43*(4), 634-646.

Tullis, J. G., & Fraundorf, S. H. (2017). Predicting others' memory performance: The accuracy and bases of social metacognition. *Journal of Memory and Language*, *95*, 124-137.

Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *1*(4), 442-452.

Whelehan, D. F., Conlon, K. C., & Ridgway, P. F. (2020). Medicine and heuristics: cognitive biases and medical decision-making. *Irish journal of medical science, 189*, 1477-1484.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition*, *2*(4), 775-780.

Wickens, T. D. (1998). On the form of the retention function: Comment on Rubin and Wenzel (1996): A quantitative description of retention. *Psychological Review*, *105*(2), 379-386.

Williams, B. W. (2006). The prevalence and special educational requirements of dyscompetent physicians. *Journal of Continuing Education in the Health Professions*, *26*(3), 173-191.

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, *55*, 235-269.

Wixted, J. T. (2005). A theory about why we forget what we once knew. *Current Directions in Psychological Science*, *14*(1), 6-9.

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science*, *18*(2), 133-134.

Young, M. E., Brooks, L. R., & Norman, G. R. (2011). The influence of familiar non-diagnostic information on the diagnostic decisions of novices. *Medical Education*, *45*(4), 407-414.

Zacks, R. T., & Hasher, L. (2006). Aging and long-term memory: Deficits are not inevitable. In E. Bialystok & F.I.M. Craik (Eds.), *Lifespan Cognition: Mechanisms of Change* (pp. 162–177).

Zheng, T., Salganik, M. J., & Gelman, A. (2006). How many people do you know in prison? Using overdispersion in count data to estimate social structure in networks. *Journal of the American Statistical Association*, *101*(474), 409-423.

Chapter 3: Self-Assessment is not enough

Key Points

- Individuals do have some ability to assess their strengths and weaknesses, so there may be a benefit to offering physicians some control over topics studied and learned.
- However, complete control is not advised because of systematic biases in self-assessment.
- The poorest performers in a domain are the least accurate in their self-assessments and tend to overestimate their actual ability.
- Inaccuracy of self-assessments has been demonstrated in both controlled laboratory environments as well as in the applied medical domain.
- Explicit instruction is not enough to remove self-assessment biases.

Overview

Much of the work in cognitive psychology suggests that physicians' ability to accurately self-assess their knowledge is critical to multiple aspects of acquiring and retaining medical expertise over time. These include when deciding what material to study (and how long to study that material) for continuing certification program assessments, when deciding among CME options, and when deciding whether or not to look up additional information for making a decision for an individual patient, and when deciding whether or not to refer a patient to a sub-specialist versus treating a patient oneself.

Within the literature on medical expertise, the notion of *self-assessment* has been critiqued for being poorly defined (Eva & Regehr, 2005). It is true that self-assessment is a multi-faceted construct and can refer to related but distinct processes. We will begin by discussing a notable framework within cognitive psychology and defining important terms before evaluating relevant evidence. The influential framework of Nelson and Narens (1990, EL: 2) identifies two processes relevant to self-assessment. First, learners must *monitor*, or assess their current knowledge and level of performance. For example, when deciding whether they have sufficient expertise to treat a patient versus refer them elsewhere, a physician might monitor their expertise by judging whether they can bring relevant information to mind, remembering their experiences treating similar patients, and/or mentally enumerating their areas of medical expertise. Second, learners must *control* their activities, or choose learning and performance strategies informed by this knowledge of their strengths and weaknesses. For example, based on this assessment of expertise, the physician might treat the patient with their current knowledge, look up additional information, or refer the patient to a specialist. Together, these processes are termed *metacognition*, or reasoning about one's own thinking and knowledge.

Research from cognitive psychology supports the claim that accurate self-assessment matters for learning: There is evidence both that (a) monitoring is causally related to decisions about learning and that (b) those decisions in turn alter the type and amount of learning that occurs. For instance, monitoring of knowledge appears to have a causal role in determining what learners study and how

much time they spend on it (Finn & Metcalfe, 2008, EL: 3); across domains and participant groups, learners generally choose to study material they have judged that they do not know as well (the *discrepancy reduction* strategy; Dunlosky & Hertzog, 1997, EL: 5; Son & Metcalfe, 2000, EL: 2; c.f., Metcalfe & Kornell, 2003, EL: 3). In turn, decisions about what to study matters for long-term retention; learners who focus their study time on difficult material end up with better overall mastery than learners who spend on their time on easy material (Tullis & Benjamin, 2011, EL: 5; c.f., Nelson & Leonesio, 1988, EL: 5). More broadly, having awareness of one's own thinking (i.e., metacognition) is predictive of academic success when controlling for general intelligence (Ohtani & Hisasaka, 2018, EL: 1).

A key implication for the longitudinal assessment of medical expertise is that physicians' ability to self-assess has direct consequences for their behavior. If physicians do not accurately monitor their knowledge, they will make poor decisions about what to study for continuing certification program assessments and, more broadly, what to review for everyday practice.

Monitoring Accuracy Has Two Components

Before we can draw any conclusions about how accurately people can self-assess their knowledge, we first must consider how accuracy can be measured. Laboratory studies have assessed the monitoring component of metacognition by having participants (a) complete some task (e.g., answering science questions) and (b) rate their level of performance. A critical question in research on monitoring has been how closely perceived performance aligns with actual performance: If learners' self-assessments are accurate, then higher confidence should predict a higher probability of correctly responding, and lower confidence a lower probability.

Often there are concerns raised over whether adults really need to be told what to study since they feel they know their areas of strength and weakness. Methodologists (e.g., Juslin, Olsson, Winman, 1996; Lichtenstein & Fischhoff, 1977; Murphy, 1973; Nelson, 1996; Nelson & Dunlosky, 1991; Schraw, 2009; Yates, 1982) have delineated how the accuracy of monitoring can be assessed on at least *two* dimensions. *Absolute accuracy* (or *calibration*) is how well a learner can predict their overall level of performance. For example, if I predict that I will get a B average this term, do I earn a B average (good calibration), or do I get an A or C average (poorer calibration)? It identifies whether learners are overconfident, under-confident, or appropriately confident in their skills. In terms of self-assessing medical expertise, good calibration would be demonstrated if physicians who were more confident in their medical skills did in fact perform better. That is, calibration measures the ability to monitor *inter-individual* variation in cognitive skills. This kind of monitoring would be important when physicians judge whether their knowledge is "good enough"; that is, is their current knowledge good enough to provide effective care for the patient population that they see, or do they need to look up additional information or acquire additional training?

Assessing absolute accuracy requires learners to provide judgments on a scale that can be directly compared to objective criterion performance. For example, rating confidence using only terms such as "somewhat confident" does not permit a true assessment of whether a learner is overconfident or under-confident because there is no objective definition of what it means to be "somewhat

confident." To measure the absolute accuracy of monitoring on test performance (where objective accuracy is on a 0-100% scale), the confidence scale would also need to refer to the probability of correct responding (e.g., on a 0-100% Likert scale). Although this would represent a change from how confidence is collected in many assessments of medical expertise, there are several potential advantages to assessing absolute accuracy, most critically including giving physicians feedback on whether they are overconfident or under-confident, as well as asking novel research questions, such as "how does absolute accuracy vary across performance outcomes?".

Relative accuracy (or *resolution*) is how well a learner can identify which topics or domains they know comparatively well--that is, their areas of expertise. For example, if I think I am more knowledgeable about diabetes than thyroid issues, is that true (good resolution), or am I in fact better with the thyroid than diabetes issues (bad resolution)? This latter aspect of metacognitive monitoring is often assessed with a correlation between predicted and actual performance for some given target skill (Nelson, 1984). In self-assessing medical expertise, good resolution would be demonstrated, if across topic areas, physicians expressed more confidence in the specific topics they were indeed better at. This kind of monitoring is important when physicians choose which topics to study for continuing certification program assessments or which CME activities to participate in.

Regehr et al. (1996) argued for focusing on relative accuracy with physicians, where *intraindividual* differences are examined. Indeed, learners have been found to have relatively poor accuracy in predicting global performance, but are much more accurate at predicting their performance on specific questions (Eva & Regehr, 2011, EL: 3). One way of doing this is to require many self-assessments from a single individual, where each self-assessment is of some sub-area of medicine. Therefore, the goal would be to make it a comparative assessment of accuracy. For instance, "am I more knowledgeable at diabetes or hypertension?"

Metacognitive Monitoring Can Be Reasonably Accurate

Confidence Predicts Accuracy

In many cases, monitoring can be reasonably accurate, though imperfect: On average, higher confidence in one's cognitive skills predicts a somewhat greater probability that one is correctly answering a question or correctly completing a task, both in terms of absolute and relative accuracy. This is true across multiple types of performance. For example, people can monitor their *episodic memory*--knowledge of specific events, such as an individual patient's symptoms and diagnosis--with reasonable accuracy such that, broadly speaking, the more confident someone is in their memory, the more likely it is to be accurate (e.g., Banks, 2000, EL: 5; Benjamin, Diaz, & Wee, 2009, EL: 5; Egan, 1958, EL: 5; Tweed, Purdie, & Wilkinson, 2020, EL: 5; Wickelgren & Norman, 1966, EL: 5; Wixted, 2007, EL: 5; Wixted & Wells, 2017, EL: 5). It is also broadly true for semantic knowledge--that is, more general world knowledge, such as the name of a nation's capital or the appropriate drugs to treat a particular syndrome (Berdie, 1971: EL 5; Goldsmith & Koriat, 2007, EL: 5; Koriat & Goldsmith, 1996, EL: 5; Metcalfe, 1986, EL: 5; Smith & Clark, 1993, EL: 5). Indeed, even when learners are unable to bring desired

information to mind in the moment, they can accurately monitor whether they are likely to be able to retrieve that information in the future (the *feeling of knowing*; Freedman & Landauer, 1966, EL: 5; Gruneberg & Monks, 1974, EL: 5; Hart, 1965, EL: 5; Hart, 1967, EL: 5; Metcalfe, 1986, EL: 5; Nelson & Narens, 1980a, EL: 5; Nelson & Narens, 1980b, EL: 5; Smith & Clark, 1993, EL: 5).

Of course, when physicians choose what to study or practice, they need to evaluate not just their immediate knowledge, but their ability to retain that information in the future. Laboratory studies have tested this ability, too, by adapting the confidence-monitoring paradigm reviewed above into the *judgments of learning* paradigm. In this paradigm, learners first study novel material and/or review existing knowledge for a future test or task. These materials similarly vary across studies and include science facts, examples of to-be-learned categories (e.g., different species of birds), and word pairs, among others. After studying each item, the learner provides--either immediately or after a delay--a judgment of learning (JOL), or an assessment of how likely they are to be able to respond correctly *on the future test*. For example, learners would rate how likely they are to remember a science fact, or to be able to classify the species of a bird photograph. Lastly, the learner takes some form of test or assessment on the material. When these JOLs are made at a delay after initial learning, they also strongly predict later performance; however, when JOLs are made immediately, they poorly predict later performance, for reasons we discuss below (Nelson & Dunlosky, 1991, EL: 5; Rhodes & Tauber, 2011, EL: 1).

The implication of these laboratory studies is that physicians are likely to be able to self-assess their skills and knowledge with a moderate, though imperfect, degree of accuracy. This conclusion has been echoed by several reviews of the medical literature (Gordon, 1991: EL 1; Davis et al., 2006: EL 2), which have found that physician's self-assessments do predict their objective performance, but only weakly to moderately. (Note, however, that these measures did not always distinguish relative from absolute accuracy.) Where absolute accuracy diverges from the ideal, it is often in the direction of physicians being overconfident (Gordon, 1991: EL 1).

So, depending on one's perspective, the glass of self-assessment could be seen as either half empty or half full. On the one hand, the imperfections of metacognitive monitoring--including some systematic biases that we review below--mean that self-assessment alone is likely insufficient. On the other hand, given that learners do have some ability to monitor themselves, that capability could be leveraged in designing longitudinal continuing certification program assessments; for instance, by allowing physicians some control over which topics to be tested on. Assuming that the early assessments are fairly low stakes, physicians may leverage their self-assessment skills to choose topics that they struggle with. Additionally, physicians may have some insights into what topics are not relevant for their practice. For example, if an orthopedist has restricted their practice to adult hips and knees, it may not make sense to ask questions about adult ankles, elbows, shoulders, or spines or about pediatric problems except, perhaps, on topics that are likely to come up when they are taking call.

Would such learner control of which materials to study be helpful? Laboratory studies find that learner control of which materials to study is better than simply allocating study time equally or based on normative difficulty (Koriat, Ma'ayan, Nussinson, 2006, EL: 3; Mazzoni & Cornoldi, 1993, Experiment 3, EL: 4; Tullis & Benjamin, 2011, EL: 3). However, in a meta-analysis of classroom studies, Karich, Burns, and Maki (2014, EL: 1) found weak to non-existent evidence that such practices benefit students. Given the ambiguity of the available evidence, it is an open question whether physicians' own self-assessments

are more or less accurate at identifying topics that should be studied compared to an algorithm based on their prior performance on the topic.

People Can Accurately Control Reporting

Above, we have shown that people can--to some degree--self-assess the accuracy of a specific task response. Another important kind of monitoring is to determine whether one should respond at all. For example, physicians must decide whether to diagnose a patient based on their current knowledge or instead consult a colleague or external resource. Ward, Gruppen, and Regehr (2002) argue that it is more important for physicians to know when to stop and seek external resources (such as peers or the medical literature) than it is to have precise accuracy in monitoring one's own cognitive skills.

Fortunately, people also have some ability to self-assess as to when to respond to questions at all. Koriat and Goldsmith (1994, EL: 3) developed a two-phase laboratory procedure to test the accuracy of self-assessment as to when to respond. In an initial phase, participants answer general world-knowledge questions (e.g., *what is the chemical process responsible for the formation of glucose in the plant cell?*) but had the option to withhold responses; payment for participation is structured such that participants lost money for incorrect responses but not for withholding responses. In the second phase, participants revisit each question and are required to respond. This permits comparison between participants' accuracy when allowed to withhold responses versus when required to respond. Critically, questions for which participants withhold responses in phase 1 are much less likely to be answered correctly in phase 2, indicating that they were successfully able to self-assess what they did not know (Goldsmith & Koriat, 1999, EL: 3; Kelley & Sahakyan, 2003, EL: 5; Koriat & Goldsmith, 1994, EL: 5; Koriat & Goldsmith, 1996, EL: 5; Koriat, Goldsmith, & Halamish, 2008, EL: 5; Goldsmith & Koriat, 2007, EL: 5). Similarly, Eva and Regehr (2011, EL: 3) found that when learners were provided with an opportunity to skip a test question that was outside their knowledge set, they reasonably chose items that they otherwise would have answered incorrectly.

Another way that people can adjust their response is to monitor and control the *grain size* at which to estimate or report a judgement. For example, imagine a physician trying to estimate how long an infection would take to clear up. The physician could provide a specific estimate (5 weeks), a narrow range (4 to 6 weeks), or a wider range (2 to 8 weeks). Fortunately, people can also adeptly self-assess the appropriate grain size. The two-phase procedure described above yields similar evidence for effective metacognitive control when, rather than being given the option to withhold responses, participants are instead allowed to control the grain size of reporting; e.g., reporting that the Berlin Wall fell in the interval *1985 to 1995* when less confident versus reporting *1989* when more confident (Goldsmith, Koriat, & Pansky, 2005, EL: 5; Goldsmith, Koriat, & Weinberg-Eliezer, 2002, EL: 5; Koriat, Goldsmith, & Halamish, 2008, EL: 5; Neisser, 1988; Yaniv & Foster, 1997, EL: 5).

The need to self-assess when to report leads to the speculative suggestion that it may be beneficial for assessments of medical expertise to additionally assess whether physicians can judiciously employ such responses and perhaps even train this metacognitive skill. Eva and Regehr (2005) note the importance of knowing the edge of one's own knowledge and when an individual should stop and

consult an external resource. This ability also has practical significance as this phenomenon occurs during physicians' work. For example, a physician may recognize that a certain patient's condition requires the knowledge set of a specialist. In the proposed studies below, we describe one method that might be used to implement such an assessment.

Metacognitive Monitoring Is Subject to Systematic Biases

Although monitoring can be reasonably accurate in some cases, as we discussed above, research has also documented several important errors and biases in self-assessment. We review several key biases before turning to theoretical accounts of biases in metacognitive monitoring. People underestimate the degree to which their cognitive skills will change in the future. On the one hand, people greatly underestimate how much they will forget between the time they learn information and the time that they need to use it (Koriat, Bjork, Sheffer, & Bar, 2004, EL: 3), likely because recently acquired knowledge feels strong and salient in the moment. On the other hand, when learners start with low initial knowledge, they *under*estimate how much they can learn in the future because that knowledge initially feels difficult and inaccessible. Even as people practice and gain skill, their JOLs tend to reflect their initial struggles (the *under confidence-with-practice effect*; Koriat, 2008b, EL: 3; Koriat, Sheffer, & Ma'aayan, 2002; c.f., Serra & Dunlosky, 2005, EL: 3).

The tendency for people to treat their present state of knowledge as if it will continue forever has been termed the *stability bias* (Kornell & Bjork, 2009, EL: 3). This bias is likely to influence physicians' self-assessment of medical expertise in two ways: First, physicians may underestimate how much they may forget after their initial training, and so the accuracy of their self-assessment years later when participating in continuing certification programs may be inflated in the absence of external feedback. Second, they may also underestimate the degree to which their skills and knowledge are amenable to learning and practice—even in their current areas of weakness and even when practices need to update to conform to advances in medicine. This may lead physicians to forego beneficial training or review unless externally prompted to do so.

A corollary to the fact that people underestimate forgetting is that self-assessment is better at a delay. One of the most robust phenomena in monitoring is the *delayed-JOL* effect (Rhodes & Tauber, 2011, EL: 1; Nelson & Dunlosky, 1991, EL: 5). JOLs made immediately after initial learning show low relative accuracy whereas *delayed JOLs*--those made later (i.e., when JOLs are gathered in a later, second study session)--can be quite accurate in predicting memory. This difference holds across participant age groups and retention intervals (Rhodes & Tauber, 2011, EL: 1) and can be explained in terms of the ease-of-processing heuristic (Begg et al., 1989, EL: 5). Immediately after studying, knowledge is still active in the learner's working (or short-term) memory[2] and feels fluent and accessible. But, over time, the contents of working memory are lost, thus rendering immediate fluency a poor index of later performance (Benjamin, Bjork, & Schwartz, 1998, EL: 3). By comparison, what comes to mind sometime after learning is much more diagnostic of long-term retention (Begg et al., 1989, EL:

[2] Working memory is a temporary memory system with limited capacity for information and is distinct from long-term memory, where stored information decays very little over time.

5). An implication for long-term retention is that self-assessments should be performed before learning; confidence ratings asked immediately after a CME course, or immediately after feedback on a continuing certification program question, are unlikely to be indicative of a physician's expertise.

Recommendation 3-1: If confidence ratings are to be collected, they should be collected before feedback is provided; confidence ratings after feedback are likely to be inaccurate.

Some of the ABMS Member Boards ask physicians to rate how relevant an individual item is to their practice, and sometimes these relevance ratings are used to determine which questions should versus should not be presented again. There are conflicting reasons to ask a question about relevance both before and after getting feedback. An argument for asking it after receiving feedback is that a physician may misunderstand how relevant a question is until receiving the answer. For example, if they come to an incorrect diagnosis, and conclude that the question is irrelevant to their practice, but the correct diagnosis is within the scope of their practice, they would be able to recognize the relevance only after receiving feedback.

However, an argument against asking it after receiving feedback is that the feedback may bias their response - they may rate it as less relevant if they get the question wrong as a form of dissonance reduction - not wanting to feel that they got something wrong that is core to their area of practice. Furthermore, it may be especially powerful feedback if a physician rates a question as highly relevant, and then gets it wrong.

Learners Evaluate Information Sources Based on Superficial Fluency

Learners may sometimes judge the reliability or utility of information sources based on relatively superficial sources of fluency (Alter & Oppenheimer, 2009, EL: 2; Oppenheimer, 2008, EL: 2). For example, students judge themselves as learning more from a lecture when the teacher stands upright and makes eye contact, even when this does not influence actual learning (Carpenter, Wilford, Kornell, & Mullaney, 2013, EL: 3; see also Fiechter, Fealing, Gerrard, & Kornell, 2018, EL: 3). An implication is that when designing a continuing certification program, the physicians who use the program will likely perceive they are learning more if care is taken to present a fluent, easy-to-use experience.

Recommendation 3-2: Any longitudinal assessment system should be designed with fluency and ease of use in mind.

Learners Neglect Optimal Learning Conditions

Learners often fail to appreciate optimal learning conditions (Finn & Tauber, 2015, EL: 2). For example, categorization tasks (e.g., learning to categorize a set of symptoms as one disease versus another) are often learned better by intermixing (*interleaving*) the to-be-learned categories rather than presenting them one at a time (*blocking*; Bjork & Bjork, 2019, EL: 3; Brunmair & Richter, 2019, EL: 1; c.f.,

Kurtz & Hovland, 1956, EL: 4). However, given the choice, learners often choose blocked practice and view it as superior to interleaving (Kirk-Johnson, Galla, & Fraundorf, 2019, EL: 3; Kornell & Bjork, 2008a, EL: 3; Kornell et al., 2010, EL: 3; Wahlheim et al., 2012, EL: 3; Yan et al., 2016, EL: 3; Zulkiply et al., 2012, EL: 3), which has been attributed (Kirk-Johnson, Galla, & Fraundorf, 2019, EL: 3; Yan et al., 2016, EL: 3) to the fact that blocked learning creates a sense of fluency in the moment even though it is less effective for long-term learning, retention, and application.

Similarly, although retrieval practice potentiates long-term retention (as we review elsewhere), learners typically judge tested materials as *less* well-learned than restudied materials (Kirk-Johnson et al., 2019, EL: 5; Roediger & Karpicke, 2006, EL: 5) and choose restudying over retrieval practice (Kirk-Johnson et al., 2019, EL: 5). And, generating or creating to-be-learned material (e.g., through fill-in-the-blank prompts) is more effective than simply passively reading it (the *generation effect*; Slamecka & Graf, 1978, EL: 3). However, because of the additional effort associated with generation, learners perceive generated material as *less* well-learned (Besken & Mulligan, 2014, EL: 3).

A general principle is that learners often mistake the initial effort required by effective study strategies (Schmidt & R. Bjork, 1992, EL: 3) as a sign those strategies are ineffective and consequently do not choose to use them (Kirk-Johnson et al., 2019, EL: 5). This implies that physicians left to study on their own may be studying in less effective or less efficient ways than they might if explicitly directed.

Accessing External Knowledge May Be Misperceived by Learners as Having Knowledge

Modern information technology allows physicians—and others—to quickly access external sources of information (e.g., via UpToDate.com). But quick access to information from the internet or other external sources (e.g., books) can create the illusion of internally possessing that knowledge (Fisher, Goddu, & Keil, 2015, EL: 3). Experimental evidence of the relationship between quick access and a feeling of knowing comes from laboratory studies that manipulated the speed at which web pages loaded in an online search task; the faster the page loaded, the better participants felt they could retain the information (Stone & Storm, 2019, EL: 3). Thus, if physicians have access to external resources when self-assessing, they may overestimate the extent of their own personal knowledge. This misattribution may be relatively benign if the resources that physicians access during self-assessment are the same that they will use on the job; in this case, self-assessment would be still representative of later performance. But, it does imply that the only external resources provided during the self-assessment should be those that physicians will later use (e.g., UpToDate, WebMD, guidelines); otherwise, self-assessments are likely to be inaccurately influenced by those external resources.

Learners Stop Studying Too Soon

Learners often terminate study too quickly: They study too few items (Murayama, Blake, Kerr, & Castel, 2016, EL: 3), and, among the items they *do* study, they do not devote sufficient time or repetitions to optimize learning (Karpicke, 2009, EL: 3; Kornell & Bjork, 2008b, EL: 3). Some of this

behavior may simply reflect the fact that learners will not persist indefinitely in studying given other, competing activities (Kurzban, Duckworth, Kable, & Myers, 2013, EL: 6). However, it may also reflect errors in self-monitoring insofar as learners do not always recognize when continuing studying can increase learning (Murayama et al., 2016, Experiment 5, EL: 4). This metacognitive error has been argued to relate to the stability bias: Once learners have learned material sufficiently well enough to respond correctly in the moment, they terminate study because they do not recognize that their cognitive skills will decline over time (Kornell & Bjork, 2008b, EL: 3). Thus, one benefit of external assessment is that it can induce additional--and beneficial--practice beyond what learners would naturally engage in.

Poor Performers Overestimate Their Performance

Another important bias that has been identified in monitoring of absolute accuracy is the Dunning-Krueger effect (*Figure 3-1*): People with low skill often greatly overestimate their performance (Dunning, Johnson, Ehrlinger, Kruger, 2003, EL: 5; Kruger & Dunning, 1999, EL: 4). The *Dunning-Kruger effect* refers to this "unskilled and unaware" phenomenon whereby those who perform poorly in a domain are also unaware they are doing poorly. (By contrast, high performers if anything *underrate* their performance; Kruger & Dunning, 1999, EL: 5). This phenomenon has been found across many domains including college social science (Dunning et al., 2003; EL: 5), formal logic (Kruger & Dunning, 1999, EL: 4), humor (Kruger & Dunning, 1999, EL: 5), English grammar (Kruger & Dunning, 1999, EL: 5), face recognition (Zhou & Jenkins, 2020; EL: 5), and medicine (Berner & Graber, 2008, EL: 2; Davis, Maznabuab, Fordis, Harrison, Thorpe et al., 2006, EL: 2; Hodges, Regehr, & Martin, 2001: EL 5; Sears, Godfrey, Lucktar-Flude, Ginsburg, Tregunno et al., 2014, EL: 2).

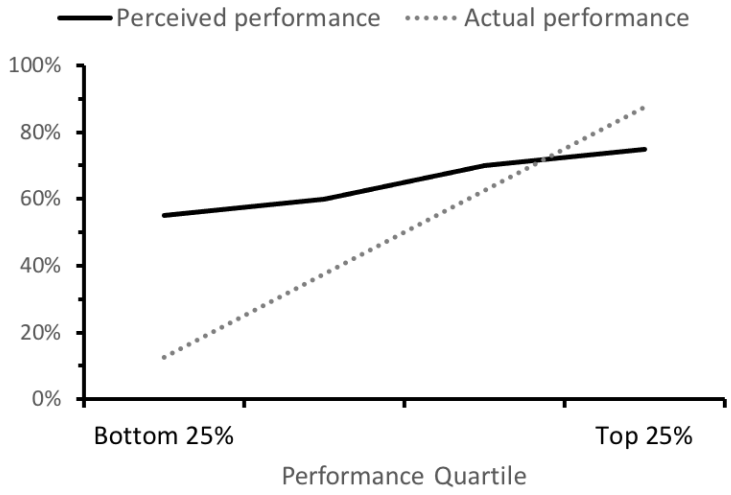


Figure 3-1. Prototypical Dunning-Kruger effect (not representing data from any specific study).

What causes the Dunning-Kruger effect? In most domains, the knowledge required for effective metacognitive monitoring in a domain is often the same as, or at least similar to, the knowledge for

effective cognitive performance (Kruger & Dunning, 1999, EL: 5; Dunning, 2011, EL: 2). For instance, imagine students factoring quadratic equations in an algebra class. To check if they have the right answer, students need to know the same rules they used to solve the problem to begin with; thus, a student who has learned the wrong rules will both produce the wrong answer *and* be unable to tell the answer is wrong. Low skill results in a "double curse" of both inaccurate performance and inaccurate self-assessment. Therefore, physicians who are weak in domain knowledge may be unaware of this fact and unable to correctly self-assess their lack of expertise, though accurate feedback may help raise their awareness.

Highlighting the importance of accurate self-assessment, physician overconfidence has been linked to diagnostic errors (Berner & Graber, 2008, EL: 2). As discussed in the previous chapter, *Cognitive Skills Need to be Kept Current*, physicians are often quite successful in their clinical decision-making. However, it has been well demonstrated that, at times, physicians can be overconfident in their diagnoses, decision-making, and assessments more generally (Berner & Graber, 2008, EL: 2). Because the right answer (e.g., a clinical diagnosis) so often arrives quickly to the mind of the physician (Barrows et al., 1982, EL: 5; Elstein, Shulman, and Sprafka, 2013, EL: 5; Gruppen, Woolliscroft, & Wolf, 1988, EL: 5; Pelaccia, et al., 2011, EL: 5), they might not always appropriately judge a wrong answer that also arrives quickly (and easily). Indeed, the ability to accurately judge whether one knows something can be challenging. Evidence from the healthcare domain has that learner's self-assessment accuracy for factual knowledge varied widely, but on average was fairly poor (Gordon, 1991, EL: 2). Self-assessment accuracy was even poorer for clinical performance than for factual knowledge. Gordon (1992, EL: 5) posited that self-assessment might be an underdeveloped cognitive skill for many, but that--like other cognitive skills--may be improved with targeted intervention.

Other Factors Can Influence What Learners Choose to Study

Cognitive psychology also provides evidence that choices of self-regulated study are guided by variables beyond those that would maximize learning and retention. Learners also preferentially practice material that they find *interesting*, regardless of how well they have learned it, and even when they know that learning is necessary for an upcoming task (Son & Metcalfe, 2000, EL: 3). Learners also fall into habits and routines of studying, such as reviewing material in the order it was originally presented, regardless of what needs the most practice or what is most important to learn (Ariel, Dunlosky, & Bailey, 2009, EL: 3; Ariel, Al-Harthy, Was, & Dunlosky, 2011, EL: 3). Learners will also preferentially choose easier material to practice, at the expense of practicing a weaker subject area (Miller, 2005, EL: 3). The preference for studying easier material may in part be due to the fact that materials are chosen based on subjective perceptions of competencies in lieu of more accurate, objective measurements. Simply asking physicians to choose their areas of study (e.g., for CME courses or for a continuing certification program) may be insufficient because physicians may choose what they find interesting or what they routinely do rather than where they may need the most continuing education. A mixture of some choice with other prescriptive material is an appropriate compromise.

Theoretical Mechanisms

Why are self-assessments not always objectively correct, and what accounts for the biases discussed above? Cognitive psychology has generally rejected a *direct-access* view of metamnemonic monitoring (Koriat, 1995, EL: 5; Koriat, 1997, EL: 5): Learners do not have the ability to directly "read off" the strength of their memory traces. Some of the starkest evidence against direct access is the fact that certain circumstances, such as very difficult questions for which the most common response is incorrect, can reverse the confidence-accuracy relationship such that answers given more confidently are actually *less* likely to be correct (Koriat, 2008a, EL: 5). This would not be possible if self-assessment were an objective assessment of knowledge.

Instead, cognitive psychology suggests an *inferential* view of metamemory (Schwartz, Benjamin, & Bjork, 1997, EL: 2; Koriat, 1997, EL: 5): Learners make an "informed guess" about their skill and knowledge based on various heuristics; critically, these heuristics are often, but not always, correct (Benjamin, Bjork, & Schwartz, 1998, EL: 6). For example, a strong predictor of memory confidence is simply the amount of information that comes to mind, whether it is right or wrong (Koriat, 1993, EL: 5). This could be explained by a heuristic whereby people base their confidence judgments on the amount of information that comes to mind. This strategy will generally produce accurate self-assessments because people often bring to mind more information about material they know well, but the strategy is not guaranteed to do so.

The inferential nature of metamnemonic monitoring implies that not *all* self-assessment will be accurate and that physicians may benefit from external feedback as to their accuracy. Further, the heuristic strategies that people use to infer their memory are subject to systematic biases, some of which can account for the recurring errors in self-assessment discussed above. In general, individuals use heuristics to aid in their decision-making. A few famous examples include the anchoring heuristic, where recently experienced information "anchors" subsequent decision-making, and the representative heuristics, where perceptions of representativeness of categories influence decision-making. Note that heuristics are good most of the time, which is likely why they exist in the first place. However, there are some rarer cases where heuristics fail to produce optimal outcomes.

Because self-assessment is merely an "informed guess," it is subject to various biases. One particularly influential bias is what Kornell, Rhodes, Castel, and Tauber (2011, EL: 3) have termed the *ease-of-processing heuristic*: Material that is experienced as subjectively fluent or easy to process in the moment is judged as better understood and learned (Alter & Oppenheimer, 2009, EL: 2; Begg, Duft, Lalonde, Melnick, & Sanvito, 1989, EL: 3; Oppenheimer, 2008, EL: 2; see also the closely related heuristic of *easily learned, early remembered*: Koriat, 2008b, EL: 4).

Evidence that learners use this heuristic comes from a wealth of experiments in which manipulations of fluency that are irrelevant to actual learning are nevertheless shown to affect JOLs. For instance, learners give higher JOLs to items that are written in a larger font (Kornell et al., 2011, EL: 3; Rhodes & Castel, 2008; EL: 3; c.f., Mueller, Dunlosky, Tauber & Rhodes, 2014, EL: 3), that are louder (Rhodes & Castel, 2009, EL: 3), that have greater visual clarity (Besken, 2016, EL: 3; Besken & Mulligan,

2013, EL: 3), even though each of these variables was unrelated to genuine memory within the respective experiments. Conversely, learners can disregard features that *do* matter for retention but that do not enhance immediate fluency (Sungkhasettee, Friedman, & Castel, 2011, EL: 3), such as the number of future study opportunities (Kornell et al., 2011, EL: 3).

We emphasize that the ease-of-processing heuristic is likely to be accurate in many cases--often, material that feels fluent and effortless *is* better learned (Koriat, 2008b, EL: 4). Nevertheless, it can also explain many of the biases reported above. Because learners use their current cognitive accessibility as a proxy for long-term learning, they underestimate both how much that accessibility may decline with forgetting or increase with study, yielding the stability bias. And, because initial fluency is an imperfect index of what contributes to long-term learning (Benjamin et al., 1998, EL: 3; Soderstrom & Bjork, 2015, EL: 3), a reliance on initial fluency may lead learners to misperceive optimal learning conditions. Similarly, fluent presentation and external sources of knowledge both create a feeling of cognitive ease that learners may mistake for genuine understanding.

Explicit Instruction Does Not Remove Self-Assessment Biases

In the sections above, we reviewed how people often use their subjective, in-the-moment experience as a heuristic to self-assess their knowledge and learning. Such judgments have been termed *non-analytic* because they are not necessarily based on conscious, verbalized introspection (Kelley & Jacoby, 1996, EL: 3). Perhaps one solution to the biases of non-analytic judgments would be to simply warn physicians about how the accuracy of their self-assessment may be flawed. Indeed, cognitive psychology does suggest that beyond these non-analytic "gut feelings," people also hold explicit beliefs that can be verbalized about which circumstances favor learning and performance. These can be used as the basis of *analytic* judgments (Fraundorf & Benjamin, 2014, EL: 4; Kelley & Jacoby, 1996, EL: 3; Koriat et al., 2004, EL: 3). For example, some learners may adopt spaced repetition because they have been taught that it is an effective study strategy, regardless of their own experience using this method (Lu & Fraundorf, in preparation, EL: 3).

However, self-assessment using explicit analytic beliefs is not a panacea. First, we cannot assume that people already know the best learning strategies. Non-scientists' beliefs about effective learning and memory are often inaccurate, as revealed by surveys of the general public (Simons & Chabris, 2011, EL: 5; Simons & Chabris, 2012, EL: 5; Yan, Thai, Bjork, 2014, EL: 5), of college students (Hartwig & Dunlosky, 2012, EL: 5; Karpicke, Butler, & Roediger, 2009, EL: 5; McCabe, 2011, EL: 5; Morehead, Rhodes, & DeLozier, 2016, EL: 5), and even of college instructors (Morehead et al., 2016, EL: 5). For example, most people--including physicians (Armson et al., 2020: EL 5)— describe self-testing primarily as a way to assess their current knowledge and not as a way to potentiate learning (Hartwig & Dunlosky, 2012, EL: 5; Kornell & Bjork, 2007, EL: 5; Kornell & Son, 2009: EL 5; McCabe, 2011, EL: 5; Morehead et al., 2016, EL: 5; Yan et al., 2014, EL: 5). Therefore, they are unlikely to spontaneously make use of the testing effect. Why do people have such mistaken beliefs about effective learning strategies?

One reason may be that they were simply never taught good strategies: About two-thirds of the U.S. population report they never received formal instruction on how best to learn (Yan et al., 2014, EL: 5).

Second, even when learners hold accurate explicit beliefs about their memory (e.g., if they believe that testing potentiates long-term retrieval), those beliefs are not always *activated* and used in self-assessment (McDaniel & Einstein, 2020: EL 2). For instance, although presumably all adults understand to some degree that information is forgotten over time, those asked to predict how much they remember a full year later give estimates no different than people asked to predict what they will remember a mere week later. Only when the question specifically uses the word "forgetting" does this belief become activated and influence predictions (Koriat et al., 2004, EL: 3). Thus, even when people are explicitly told that in-the-moment fluency can be a misleading basis for self-assessment and instructed to disregard it, they are not entirely successful in doing so (e.g., Yan, Bjork, & Bjork, 2016, EL: 3).

A key implication for the maintenance of medical expertise is that we cannot expect physicians to naturally know how best to self-assess or keep their knowledge current. Further, simply instructing physicians on how best to self-assess may be insufficient because even if physicians acquire accurate analytic beliefs (e.g., that testing benefits long-term retention), those beliefs will not always be used in self-assessment. Thus, external prompts for practice and self-assessment may be critical.

Chapter Summary

Metacognitive control of learning comprises two processes: monitoring and control. Monitoring refers to the ability to accurately self-assess one's own knowledge and abilities. Control refers to an individual's choices in how one guides their learning and performance strategies. Prior research supports that accurately self-assessing one's own abilities and knowledge is important to guiding (controlling) one's learning and maintaining one's expertise. For instance, self-assessment abilities are associated with the quality of learning strategies used by an individual and consequently learning outcomes.

Nevertheless, the literature we reviewed suggests that individuals do not have direct access to the strength of individual memories or knowledge and instead only have an "informed guess." These "informed guesses" often lead to systematic biases. For example, information that feels easier to process in-the-moment can lead individuals to overconfidence in how much they will remember in the future. Thus, self-assessments of knowledge immediately after learning something tend to be less accurate than delayed judgments. Relatedly, learners often stop studying too soon and underestimate the requisite amount of practice needed to adequately learn target information. The tendency to judge learning based on in-the-moment fluency can also lead to choosing sub-optimal learning strategies liked blocked studying (e.g., studying only hypertension, instead of several interspersed subject areas) and re-reading, rather than more effective strategies such as interleaving and retrieval practice, because those sub-optimal study practices feel less effortful at the time of study.

Another notable phenomenon in the self-assessment literature is the Dunning-Kruger effect, the robust finding that poorest performers are the least accurate in their self-assessments and tend to

overestimate their actual ability. Additionally, the top performers tend to underestimate their ability, though their absolute accuracy is not as poor as the poorest performers'. The Dunning-Kruger effect has been replicated in many domains, including in medicine with physicians.

Merely instructing learners about the existence of these biases is not enough to remediate them, although some preliminary evidence suggests that the chance to experience different learning conditions with feedback has promise as a more effective way to improve self-assessment accuracy. Given the available evidence that self-assessment can be biased and not easily remediated, we believe that externally guided learning for physicians in a longitudinal assessment program is central to creating a successful learning platform.

Future Directions

In reviewing this literature, we identified a number of directions for future research that were specifically lacking in the literature or not seen applied to the physician population of continuing certification. These include five proposed studies where we supply the ideas but not necessarily a fully developed research design.

1. Is Asking about Relevance best Before or After Feedback?

Earlier in the chapter we reviewed reasons for asking whether a question is relevant to a physician's practice before getting feedback (they may change their answer after getting feedback to feel better about getting it wrong) versus after (they may not know how relevant it is or is not until knowing the right answer such as the correct diagnosis). It may be worthwhile to study whether it makes a difference. For example, for certain questions relevance could be asked before, after, or potentially both, to see if there are differences between groups, and if physicians change their relevance ratings after getting a question wrong. They could also be prompted to explain why they changed their answer.

2. What is the most Appropriate Response Scale for Confidence Judgments?

Currently, physicians' confidence judgments are collected on different scales across member Boards. There is likely to be utility in exploring the optimal means of assessing confidence; for instance, which scales are interpreted most consistently? As we discussed above, confidence scales that include some reference to an objective standard of performance (e.g., "75% confident I'm right") would allow measures of absolute accuracy (e.g., overconfidence vs. under-confidence) to be collected and provided as feedback. Further, it may be useful to determine how many different intervals or categories of confidence can be differentiated by learners--is there a meaningful difference between, for instance, "very confident" and "extremely confident"? This issue is important because, given imprecision in how people translate internal confidence into external ratings (Benjamin, Diaz, & Wee, 2009: EL 2), a scale with too many categories may in fact decrease the accuracy of confidence ratings (Benjamin, Tullis, & Lee, 2013: EL 3). Lastly, it may be valuable to determine whether the highest level of confidence (e.g., "I'm virtually certain") represents a qualitatively distinct state of special accuracy, as proposed by certain dual-process theories of recognition (Parks & Yonelinas, 2007: EL 2; Yonelinas, 1994: EL 3; Yonelinas, 2002: EL 2; c.f., Wixted 2007: EL 2).

3. How does Physician Customization Enhance or Degrade the Assessment?

The intent of longitudinal assessment is to provide both a defensible pass-fail decision (to be used for making continuing certification decisions) and formative feedback (to help physicians continue to improve the breadth, depth, and currency of their medical knowledge throughout their career). Providing useful formative feedback requires tremendous flexibility so that each participant receives customized feedback targeted to the areas in which they need or wish to improve. Conversely, making defensible pass-fail decisions is simplified if there is a high degree of standardization so that all physicians attempting to maintain their certification are responsible for similar content mastery reflecting the certificates they hold. At times, these purposes seem antithetical.

In addition to studies on the impact that formative feedback has upon learning, it is even more critical to establish and maintain the quality of the summative aspects that will be used to make the pass-fail decisions. The pass-fail decision is often the hurdle that prevents some physicians from remaining certified, and in those instances the assessment organization will need firm evidence to justify that decision. Also, Boards often find it desirable to permit physicians to customize their assessment somewhat to make the assessment more representative of their practice. This customization may help with providing better formative feedback and giving physicians a greater sense of relevance to their practice, but it can sometimes degrade the fit between the measurement and the intended meaning of the certificate. This customization is more feasible in longitudinal programs than point-in-time assessments because they engage in the program every year instead of once every 5 or 10 years. Therefore, validity studies and analyses of psychometric quality should continue to be conducted to ensure that quality of the summative component has not been compromised. A few such studies are suggested below.

- Is the precision of the participants' scores sufficient to make defensible pass-fail decisions?
- Is the number of questions scored for summative purposes sufficient to represent the specialty or subspecialty?

4. Does Self-Assessment Accuracy Differ across Levels of Granularity of the Subject Matter?

Eva and Regehr (2011, EL: 3) propose a distinction between *self-assessment* at the global level (e.g., "How good a physician am I?") and *self-monitoring* of specific topic areas (e.g., "How much do I know about hypertension?"). In laboratory studies with college students, they found that learners could predict their performance much more accurately for specific questions than at a global level.

This distinction is relevant if physicians' confidence ratings are to be used for any purpose, such as controlling which topics get focused on in a longitudinal assessment. At what level of granularity must these confidence ratings be collected to be accurate? We propose investigating self-assessment accuracy across different levels of granularity. For instance, a physician can be asked to self-assess their competency globally as a physician (the highest level), at a topic level (e.g., hypertension; medium level), and at an item level (e.g., a targeted question about hypertension; the lowest level). The practical question is whether an accurate self-assessment can be obtained by querying physicians at a more general level or only at the item level.

5. Does Self-Assessment Accuracy Differ across Objective versus Comparative Feedback?

Another dimension on which self-assessment is whether they are made relative to an *objective* standard (e.g., "What percent correct will you get on this assessment?") or to a *social* or *comparative* standard (e.g., "How well do you think you will perform on this assessment relative to other doctors?" or "What percentile will you score in?"; Festinger, 1954, EL: 2). Some evidence from outside medicine suggests that people are more sensitive to their objective standing than their comparative standing (Hoelzl & Rustichini, 2005, EL: 4; Kruger & Burrus, 2004, EL: 3; Moore & Kim, 2003, EL: 3; Windschitl, Kruger, & Simms, 2003, EL: 3) and, perhaps as a result, are more responsive to objective than comparative feedback (Moore & Klein, 2008, EL: 3). Nevertheless, it may be of interest to collect physicians' self-assessments in both objective and comparative terms to determine which method yields more accurate self-assessment.

References

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*(3), 219-235.

Ariel, R., Al-Harthy, I. S., Was, C. A., & Dunlosky, J. (2011). Habitual reading biases in the allocation of study time. *Psychonomic Bulletin & Review*, *18*(5), 1015-1021.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General*, *138*(3), 432–447.

Armson, H., Roder, S., Wakefield, J., & Eva, K. W. (2020). Toward practice-based continuing education protocols: Using testing to help physicians update their knowledge. *Journal of Continuing Education in the Health Professions, 40*(4), 248-256.

Banks, W. P. (2000). Recognition and source memory as multivariate decision processes. *Psychological Science*, *11*(4), 267-273.

Barrows, H. S., Norman, G. R., Neufeld, V. R., & Feightner, J. W. (1982). The clinical reasoning of randomly selected physicians in general medical practice. Clinical and investigative medicine. *Clinical and Investigative Medicine*, *5*(1), 49-55.

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language*, *28*(5), 610-632.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, *127*(1), 55-68.

Benjamin, A. S., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise: applications to recognition memory. *Psychological Review*, *116*(1), 84-115.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(5),1601-1608.

Berdie, R. F. (1971). Self-claimed and tested knowledge. *Educational and Psychological Measurement*, *31*, 629-636.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine*, *121*(5), S2-S23.

Besken, M. (2016). Picture-perfect is not perfect for metamemory: Testing the perceptual fluency hypothesis with degraded images. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1417-1433.

Besken, M., & Mulligan, N. W. (2013). Easily perceived, easily remembered? Perceptual interference produces a double dissociation between metamemory and memory performance. *Memory & Cognition*, *41*(6), 897-903.

Besken, M., & Mulligan, N. W. (2014). Perceptual fluency, auditory generation, and metamemory: Analyzing the perceptual fluency hypothesis in the auditory modality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 429-440.

Bjork, R. A., & Bjork, E. L. (2019). The myth that blocking one's study or practice by topic or skill enhances learning. In C. Barton (Ed.), *Education Myths: An Evidence-Informed Guide for Teachers*. John Catt Educational Ltd.

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029-1052.

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*(6), 1350-1356.

Davis, D. A., Mazmanian, P. E., Fordis, M., Van Harrison, R. T. K. E., Thorpe, K. E., & Perrier, L. (2006). Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. *The Journal of the American Medical Association*, *296*(9), 1094-1102.

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *52*(4), P178-P186.

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in Experimental Social Psychology* (Vol. 44, pp. 247-296). Academic Press.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science*, *12*(3), 83-87.

Egan, J. P. (1958). Recognition memory and the operating characteristic. USAF Operational Applications Laboratory Technical Note.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (2013). *Medical problem solving: An analysis of clinical reasoning.* Harvard University Press.

Eva, K. W., & Regehr, G. (2005). Self-assessment in the health professions: a reformulation and research agenda. *Academic Medicine*, *80*(10), S46-S54.

Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education*, *16*(3), 311-329.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117-140.

Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, *58*(1), 19-34.

Finn, B., & Tauber, S. K. (2015). When confidence is not a signal of knowing: How students' experiences and beliefs about processing fluency can lead to miscalibrated confidence. *Educational Psychology Review*, *27*(4), 567-586.

Fiechter, J. L., Fealing, C., Gerrard, R., & Kornell, N. (2018). Audiovisual quality impacts assessments of job candidates in video interviews: Evidence for an AV quality bias. *Cognitive Research: Principles and Implications*, *3*(1), 47-52.

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*(3), 674-687.

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17-38.

Freedman, J. L., & Landauer, T. K. (1966). Retrieval of long-term memory:"Tip-of-the-tongue" phenomenon. *Psychonomic Science*, *4*(8), 309-310.

Goldsmith, M., & Koriat, A. (1999). The strategic regulation of memory reporting: Mechanisms and performance consequences. Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application, 373-400.

Goldsmith, M., & Koriat, A. (2007). The strategic regulation of memory accuracy and informativeness. *Psychology of Learning and Motivation*, *48*, 1-60.

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language*, *52*(4), 505-525.

Goldsmith, M., Koriat, A., & Weinberg-Eliezer, A. (2002). Strategic regulation of grain size memory reporting. *Journal of Experimental Psychology: General*, *131*(1), 73-95.

Gordon, M. J. (1991). A review of the validity and accuracy of self-assessments in health professions training. *Academic Medicine*, *66*(12), 762-769.

Gordon, M. J. (1992). Self-assessment programs and their implications for health professions training. *Academic Medicine*, *67*, 672-679.

Gruneberg, M. M., & Monks, J. (1974). 'Feeling of knowing' and cued recall. *Acta Psychologica*, *38*(4), 257-265.

Gruppen, L. D., Woolliscroft, J. O., & Wolf, F. M. (1988). The contribution of different components of the clinical encounter in generating and eliminating diagnostic hypotheses. In *Research in medical education: Proceedings of the annual conference on research in medical education* (Vol. 27, p. 242-247).

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology*, *56*(4), 208-216.

Hart, J. T. (1967). Memory and the memory-monitoring process. *Journal of verbal learning and verbal behavior*, *6*(5), 685-691.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement?. *Psychonomic Bulletin & Review*, *19*(1), 126-134.

Hodges, B., Regehr, G., & Martin, D. (2001). Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. *Academic Medicine*, *76*, S87–S89.

Hoelzl, E., & Rustichini, A. (2005). Overconfident: Do you put your money on it? *The Economic Journal*, *115*, 305-318.

Juslin, P. Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1304-1316.

Karich, A. C., Burns, M. K., & Maki, K. E. (2014). Updated meta-analysis of learner control within educational technology. *Review of Educational Research*, *84*(3), 392-410.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469-486.

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own?. *Memory*, *17*(4), 471-479.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic bases for judgment. *Journal of Memory and Language*, *35*(2), 157-175.

Kelley, C. M., & Sahakyan, L. (2003). Memory, monitoring, and control in the attainment of memory accuracy. *Journal of Memory and Language*, *48*(4), 704-721.

Kirk-Johnson, A., Galla, B. M., & Fraundorf, S. H. (2019). Perceiving effort as poor learning: The misinterpreted-effort hypothesis of how experienced effort and perceived learning relate to study strategy choice. *Cognitive Psychology*, *115*, 101237.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review*, *100*(4), 609-639.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General*, *124*(3), 311-333.

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349-370.

Koriat, A. (2008a). When confidence in a choice is independent of which choice is made. *Psychonomic Bulletin & Review*, *15*(5), 997-1001.

Koriat, A. (2008b). Easy comes, easy goes? The link between learning and remembering and its exploitation in metacognition. *Memory & Cognition*, *36*(2), 416-428.

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: the role of experience-based and theory-based processes. *Journal of Experimental Psychology: General*, *133*(4), 643-656.

Koriat, A., & Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: Distinguishing the accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, *123*(3), 297-315.

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *103*(3), 490-517.

Koriat, A., Goldsmith, M., & Halamish, V. (2008). Controlled processes in voluntary remembering.

Koriat, A., Ma'ayan, H., & Nussinson, R. (2006). The intricate relationships between monitoring and control in metacognition: Lessons for the cause-and-effect relation between subjective experience and behavior. *Journal of Experimental Psychology: General*, *135*(1), 36-69.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147-162.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219-224.

Kornell, N., & Bjork, R. A. (2008a). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory*, *16*(2), 125-136.

Kornell, N., & Bjork, R. A. (2008b). Learning concepts and categories: Is spacing the "enemy of induction"?. *Psychological Science*, *19*(6), 585-592.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*(4), 449-468.

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*(2), 498-503.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and the stability bias: Dissociating memory, memory beliefs, and memory judgments. *Psychological Science*, *22*(6), 787-794.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493-501.

Kruger, J., & Burrus, J. (2004). Egocentrism and focalism in unrealistic optimism (and pessimism). *Journal of Experimental Social Psychology*, *40*(3), 332-340.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121-1134.

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, *51*(4), 239-243.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). An opportunity cost model of subjective effort and task performance. *Behavioral and Brain Sciences*, *36*(6), 661-679.

Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, *20*, 159-183.

Lu, A. Z., & Fraundorf, S. H. (2020). How beliefs and perceptions influence study strategy decisions. Manuscript in preparation.

Mazzoni, G., & Cornoldi, C. (1993). Strategies in study time allocation: Why is study time sometimes not effective?. *Journal of Experimental Psychology: General*, *122*(1), 47-60.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. *Memory & Cognition*, *39*(3), 462-476.

McDaniel, M. A., & Einstein, G. O. (2020). Training learning strategies to promote self-regulation and transfer: The knowledge, belief, commitment, and planning framework. *Perspectives on Psychological Science*, *15*(6), 1363-1381.

Metcalfe, J. (1986). Premonitions of insight predict impending error. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(4), 623-634.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General*, *132*(4), 530-542.

Miller, S. H. (2005). American Board of Medical Specialties and repositioning for excellence in lifelong learning: maintenance of certification. *Journal of Continuing Education in the Health Professions*, *25*(3), 151-156.

Moore, D. A., & Kim, T. G. (2003). Myopic social prediction and the solo comparison effect. *Journal of Personality and Social Psychology*, *85*(6), 1121-1135.

Moore, D. A., & Klein, W. M. P. (2008). Use of absolute and comparative performance feedback in absolute and comparative judgments and decisions. *Organizational Behavior and Human Decision Processes*, *107*(1), 60-74.

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*(2), 257-271.

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory?. *Journal of Memory and Language*, *70*, 1-12.

Murayama, K., Blake, A. B., Kerr, T., & Castel, A. D. (2016). When enough is not enough: Information overload and metacognitive decisions to stop studying information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 914-924.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595-600.

Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical Psychology*, *1*(1), 35-59.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109-133.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science*, *2*(4), 267-271.

Nelson, T. O., & Leonesio, R. J. (1988). Allocation of self-paced study time and the" labor-in-vain effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 676-686.

Nelson, T. O., & Narens, L. (1980a). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and verbal Behavior*, *19*(3), 338-368.

Nelson, T. O., & Narens, L. (1980b). A new technique for investigating the feeling of knowing. *Acta Psychologica*, *46*(1), 69-80.

Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In Psychology of learning and motivation (Vol. 26, pp. 125-173). Academic Press.

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: a meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*(2), 179-212.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences*, *12*(6), 237-241.

Parks C. M. & Yonelinas A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted. *Psychological Review*, *114*, 188-202.

Pelaccia, T., Tardif, J., Triby, E., & Charlin, B. (2011). An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Medical Education Online*, *16*(1), 5890.

Regehr, G., Hodges, B., Tiberius, R., & Lofchy, J. (1996). Measuring self-assessment skills: an innovative relative ranking model. *Academic Medicine*, *71*(10), S52-S54.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General*, *137*(4), 615-625.

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review*, *16*(3), 550-554.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin*, *137*(1), 131-148.

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207-218.

Schraw, G., (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, *4*, 33-45.

Schwartz, B. L., Benjamin, A. S., & Bjork, R. A. (1997). The inferential and experiential bases of metamemory. *Current Directions in Psychological Science*, *6*(5), 132-137.

Sears, K., Godfrey, C. M., Luctkar-Flude, M., Ginsburg, L., Tregunno, D., & Ross-White, A. (2014). Measuring competence in healthcare learners and healthcare professionals by comparing self-

assessment with objective structured clinical examinations: A systematic review. *JBI Database of Systematic Reviews and Implementation Reports*, *12*(11), 221-272.

Serra, M. J., & Dunlosky, J. (2005). Does retrieval fluency contribute to the underconfidence-with-practice effect?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1258-1266.

Simons, D. J., & Chabris, C. F. (2011). What people believe about how memory works: A representative survey of the US population. *PloS One*, *6*(8), e22757.

Simons, D. J., & Chabris, C. F. (2012). Common (mis) beliefs about memory: A replication and comparison of telephone and Mechanical Turk survey methods. *PloS One*, *7*(12), e51876.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of experimental Psychology: Human learning and Memory*, *4*(6), 592-604.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25-38.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176-199.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204-221.

Stone, S. M., & Storm, B. C. (2019). Search fluency as a misleading measure of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Sungkhasettee, V. W., Friedman, M. C., & Castel, A. D. (2011). Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin & Review*, *18*(5), 973-978.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, *64(*2), 109-118.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*, 429–442.

Tweed, M., Purdie, G., & Wilkinson, T. (2020). Defining and tracking medical student self-monitoring using multiple-choice question item certainty. *BMC Medical Education*, *20*(1), 1-9.

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition*, *40*(5), 703-716.

Ward, M., Gruppen, L., & Regehr, G. (2002). Measuring self-assessment: Current state of the art. *Advances in Health Sciences Education*, *7*(1), 63-80.

Wickelgren, W. A., & Norman, D. A. (1966). Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology*, *3*(2), 316-347.

Windschitl, P. D., Kruger, J., & Simms, E. (2003). The influence of egocentrism and focalism on people's optimism in competitions: When what affects us equally affects me more. *Journal of Personality and Social Psychology*, *85*(3), 389-408.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152-176.

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest*, *18*(1), 10-65.

Yan, V. X., Bjork, E. L., & Bjork, R. A. (2016). On the difficulty of mending metacognitive illusions: A priori theories, fluency effects, and misattributions of the interleaving benefit. *Journal of Experimental Psychology: General*, *145*(7), 918-933.

Yan, V. X., Thai, K. P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: do they vary with mindset?. *Journal of Applied Research in Memory and Cognition*, *3*(3), 140-152.

Yan, V. X., Yu, Y., Garcia, M. A., Bjork, R. A. (2014). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition*, *42*, 1373–1383.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, *10*(1), 21-32.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance*, *30*, 132-156.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341–1354.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441-517.

Zhou, X., & Jenkins, R. (2020). Dunning–Kruger effects in face perception. *Cognition*, 203, 104345.

Zulkiply, N., McLean, J., Burt, J. S., & Bath, D. (2012). Spacing and induction: Application to exemplars presented as auditory and visual text. *Learning and Instruction*, *22*(3), 215-221.

Chapter 4: Testing Enhances Learning and Retention

Key Points

- Randomized controlled trials provide clear and abundant evidence that testing is superior to re-study for knowledge retention.
- The benefits of testing are not transient, but rather hold up over measurement periods as long as two years.
- Testing even strengthens learning and retention of more complex skills, such as (diagnostic) classification.
- Feedback after testing increases the learning benefits even more.

Overview

For over 100 years, psychologists have been aware of the learning benefits of testing one's own knowledge, including the earliest psychological studies on memory (Abott, 1909, EL: 4; Ebbinghaus, 1885, EL: 5). This phenomenon has been studied and reviewed with varying approaches over the years. Overall, this work demonstrates that retrieving information from memory through testing enhances subsequent retention more than does restudying (e.g., rereading or listening to) the same information (Adesope, Trevisan, & Sundararajan, 2017, EL: 1; Rowland, 2014, EL: 1; Yang, Luo, Vadillo, Yu, & Shanks, 2021: EL 1). This phenomenon is often referred to as *the testing effect*, although it has also been referred to as *test-enhanced learning*, *retrieval practice*, and *retrieval-based learning*.

Basic Experimental Design

The basic testing-effect experiment compares, at a minimum, two groups to which individuals are randomly assigned: a re-study group and a testing group (e.g., Carpenter, Pashler, Wixted, & Vul, 2008, EL: 4; Karpicke & Roediger, 2008, EL: 4; Roediger & Karpicke, 2006a, EL: 3, 2006b, EL: 3). The re-study group initially studies information and then has an additional study opportunity later. The testing group initially studies information and, instead of re-studying the material, is tested on it. (Some experiments also include a third control group that only initially studies information; e.g., LaPorte & Voss, 1975, EL: 4). The two groups then complete some assessment of memory or performance (see *Figure 4-1* for visual representation).

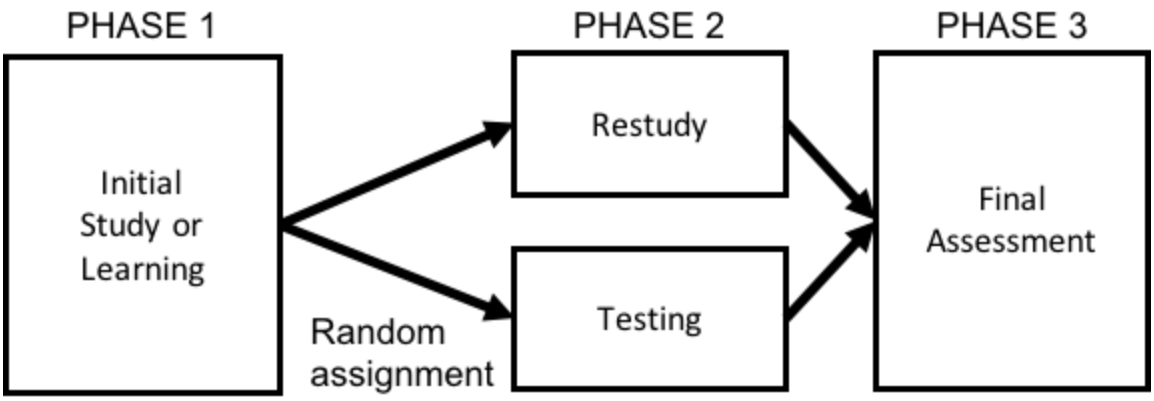


Figure 4-1. Schematic design of the typical testing-effect study procedure.

Meta-analytic reviews (Adesope et al., 2017, EL: 1; Rowland, 2014, EL: 1; Yang et al., 2021, EL: 1) provide evidence for the benefits of repeat testing or retrieval practice over restudy for long-term retention. This benefit holds across a wide variety of authentic educational domains (Yang et al., 2021: EL 1), including the natural sciences (Agarwal et al., 2012, EL: 3; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011, EL: 3; McDermott, Agarwal, D'Antonio, Roediger, & McDaniel, 2014, EL: 3; Yang et al., 2021, EL: 1), mathematics and statistics (Hopkins, Lyle, Hieb, & Ralston, 2016, EL: 4; Kang, McDaniel, & Pashler, 2011, EL: 4; Lyle & Crawford, 2011, EL: 4; Yang et al., 2021, EL: 1), geography and maps (Carpenter & Pashler, 2007, EL: 4; Rohrer, Taylor, & Sholar, 2010, EL: 3; Yang et al., 2021, EL: 1), psychology (McDaniel, Anderson, Derbish, & Morrisette, 2007, EL: 4; Wiklund-Hörnqvist, Jonsson, & Nyberg, 2014, EL: 4; Yang et al., 2021, EL: 1), and history (Agarwal et al., 2012, EL: 4; Carpenter, Pashler, & Cepeda, 2009, EL: 4; McDermott et al. 2014, EL: 3; Nungester & Duchastel, 1982, EL: 4; Roediger, Agarwal, McDaniel, & McDermott, 2011, EL: 3; Yang et al., 2021, EL: 1). In short, among the content domains studied, there is no domain for which there is systematic evidence that testing does *not* work to support retention of knowledge and cognitive skills. Indeed, and most critically for our purposes, a meta-analysis of classroom studies (Yang et al., 2021, EL 1) finds significant benefits of retrieval practice on learning of content in medicine (and also in nursing).

Recommendation 4-1: Testing should be used to support long-term retention of knowledge and cognitive skills.

Importantly, these randomized controlled trials of retrieval practice control for many potential confounders. For example, it would generally be inadequate to compare a testing group to a control group that received no form of practice prior to the final assessment since the control group would also spend less time with the subject matter. Thus, the meta-analysis by Rowland (2014, EL: 1) includes only randomized controlled trials that met the following stringent inclusion criteria: (a) equivalent amounts of time were spent with the target material across conditions prior to final assessment, (b) participants were tested on the same information that was initially studied, (c) identical information were presented in test and re-study conditions, (d) both groups exclude classroom learning. The fact that a testing effect

emerges with such controls indicates it is driven by the benefits of testing and not by mere re-exposure to information. A more recent meta-analysis conducted by Adesope, Trevisan, and Sundararajan (2017, EL: 1) was published with similar but not identical criteria. Adesope et al. included studies which: a) compared a practice test group and a group who either restudied information or engaged in a separate learning intervention, b) reported quantitative measurements of learning (e.g., group post-test scores), c) included necessary information to compute a study effect size, d) published to academic databases or other platforms from which they could be accessed, and e) featured some methodological control for preexisting differences (e.g., within-subject designs, pre-tests, matching procedures).

These studies show that testing is better than restudying, but how beneficial is retrieval practice via testing? Rowland's (2014, EL: 1) meta-analysis of 61 studies estimated the size of the testing effect as Hedges' g = 0.50; in other words, people randomly assigned to testing scored half a standard deviation (0.50) better than those assigned to restudy, constituting a medium effect size. Adesope et al. (2017: EL 1)'s more recent meta-analysis of 118 articles (272 independent effect sizes; 15,427 participants) found an even larger Hedges' g of 0.70. Retrieval practice or repeated testing is actually better at enhancing long-term retention and comprehension than some other popular educational techniques, such as concept mapping (Karpicke & Blunt, 2011 EL: 3). As such, being tested appears to be an effective way of enhancing physicians' long-term retention of medical expertise.

Benefits to Spaced Learning in Medicine

In recent years, there has been a growing call for a greater reliance upon testing as a studying and learning tool for students in the health professions (Brown, 2017, EL: 6; Cilliers, 2015, EL: 6; Chesluk et al., 2019; Fung, Joegi, & Fung, 2019, EL: 6; Griffith, Purkiss, Santen, & Burk-Rafel, 2017; EL: 6; Kulasegaram & Rangachari, 2018, EL: 3; Piza et al., 2019; EL: 5; Rapp, Maximin, & Green, 2014, EL: 6; Richmond, Cranfield, & Cooper, 2019, EL: 6). These calls typically include testing in regularly spaced intervals in contrast to 'cramming' study behavior (an issue we discuss in further detail below); the combination of testing and spacing over time has been termed *spaced repetition*.

Systematic review (Phillips, Heneka, Bhattarai, Fraser, & Shaw, 2019: EL 2) provides evidence that spaced repetition enhances practicing clinicians' acquisition of knowledge and their clinical behaviors. For example, Raupach et al. (2016; EL: 3) examined the impact of spaced learning on clinical reasoning skills among undergraduate students in their fourth year of medical education. Clinical reasoning is critical for physicians and includes behaviors related to diagnosing patients, evaluating test results, and implementing (or selecting) patient interventions. The researchers split the students into two groups (N = 124 at start of study, 70% retention rate) and designed a learning intervention where all learners would be exposed to the same content each week, but one group of students would be randomized to also receive questions about key features of the target content. During weeks where the additional learning questions were present, students also had access to explanations for correct/incorrect answers, which provided an opportunity for feedback. Raupach et al. found that students exhibited the greatest memory retention during spaced learning when target information was paired with questions about key features of the material (and had an opportunity for feedback). This study provides evidence that spaced learning with testing does more than just boost rote memorization,

but also provides benefits to more complex, high order cognitive processes as well. Additionally, as Raupach and colleagues note, other common methods of improving clinical reasoning (e.g., clinical rotations and small group sessions) are much more financially burdensome.

Larsen, Butler, and Roediger (2009, EL: 4) conducted a randomized controlled trial in which the participants were medical residents attending a conference. Participants first attended a learning session on two medical topics (status epilepticus and myasthenia gravis). After the learning session, participants were given a short-answer practice test (with feedback) on one topic and a review/study sheet on the other topic; the assignment of strategies to topics was randomized across two counterbalanced groups. Participants repeated the testing/restudying procedure, keeping the same assignment of topics to study strategies, roughly every two weeks until each topic had been revisited three times. Six months after initial learning, a final test was given to assess the amount of knowledge that was retained for the two topics. For both topics, repeated testing resulted in better memory than repeated restudy. Overall, testing had a large effect size of $d = 0.91$. This finding speaks to the value and relevance of testing and spaced learning for the medical realm.

Armson, Roder, Wakefield, & Eva (2020, EL: 4) conducted a similar randomized controlled trial with family physicians completing an online learning module. The online platform randomly assigned participants to either read an academic review article or take a quiz prior to engaging with the learning module. One week after the learning session, participants completed a post-test. Again, a large benefit of testing over reading ($d = 1.6$) was found across each of two topics (on each of two topics (opioids and chronic kidney disease).

Spaced learning has been shown to also benefit clinical behaviors. Shaw, Long, Chopra, and Kerfoot (2011, EL: 3) conducted a randomized controlled trial with health care providers in the wake of a medical conference. Participants included physicians (~45% had an MD degree), nurses, and physician assistants. Participants randomized to the spaced learning condition underwent regular adaptive learning testing--where test questions pertaining to the conference learning material that were missed were recycled back into a participant's queue for a later session. If a question was correctly answered twice in a row, it would not appear again. At the end of the spaced learning period, which lasted 18 weeks, all participants completed a survey which measured subjective implementation of practices related to material presented at the conference. Shaw et al. found that health care providers in the spaced learning condition reported greater adoption of the new material in their clinical practice. Additionally, 97% of participants stated interest in participating in supplemental spaced learning after future continuing medical education conferences. This work speaks to the benefits of spaced learning in the medical domain, in areas related to testing and in-person conferences.

Not all spaced learning interventions lead to differential learning outcomes. Timer, Steendijk, Arend, and Versteeg (2020, EL: 4) randomized 148 second year medical students to either a spaced lecture, which was a lecture divided into 15 minute segments, or a *massed* lecture (one long segment). Importantly, both conditions experienced the same information, with the only difference being whether the lecture was in *massed* versus in *smaller spaced out segments*. The researchers did not find evidence

to support that the smaller spaced out segments benefitted learning outcomes. One reason an effect may not have been found is due to the limited gap between the lecture segments (5 minutes) in the spaced condition. This study highlights the importance of both the interval length between spaced learning sessions and the nature of the sessions themselves (e.g., learning new information vs. rehearsing old information).

It is clear from the relatively limited literature examining the efficacy of retrieval practice in physicians that more rigorous scientific work is needed. The relatively limited studies that have been done often suffer from poor designs that limit causal attribution (e.g., cross-sectional, self-report, selection effects, or correlational methods), although a few well-controlled studies do exist (e.g., Larsen, Butler, & Roediger, 2009). Further, only a relatively small subset of studies in the medical domain has included participants other than medical students or residents. Given the growing emphasis on evidence-based studying practices, more research should be done to assess its efficacy in medicine. However, despite these valid limitations, it is reasonable to assume from the plethora of evidence provided from basic-science approaches that retrieval practice would be effective in medicine and in learning more generally. Given that the testing effect has been found to benefit learning in many domains (e.g., natural sciences, history, psychology, geology) and never found to harm learning in any domain, we believe the existing evidence firmly supports that testing will benefit cognitive skills in the domain of medicine. By practicing retrieving information from our memory, we strengthen our memories and increase our knowledge.

Moderators of Testing Effect

In light of the general benefits of testing, researchers have also explored potential moderators of the testing effect by varying the parameters of the basic testing-effect design presented in *Figure 4-1*. These potential moderators include the retention interval (i.e., the time in between learning and assessment), retrieval type (e.g., cued vs. free recall), and feedback type (i.e., immediate, delayed, or none), among others. It is important to note that not every possible combination of factors has been manipulated within the same study; that is, while many studies have explored the effects of feedback type, test format, and content domain, there may not be any study on the specific combination of explanatory feedback on free recall of chemistry facts. However, key findings have been reached about each of these moderators taken on its own, which we review below.

Retention Interval and "Cramming"

The benefit of testing over restudy for retention remains even when assessed eight to 24 months later (Agarwal et al., 2012, EL: 4; Kerfoot, 2009, EL: 4). In fact, the benefits of testing are if anything intensified with a longer *retention interval* (i.e., the time between learning and assessment), a phenomenon known as the *test-delay interaction* (e.g., Agarwal et al., 2012, EL: 4; Chan, 2010, EL: 4; Roediger & Karpicke, 2006a, EL: 3; Rowland, 2014, EL: 1; Runquist, 1983, EL: 3; Toppino & Cohen, 2009, EL: 3; Wheeler, Ewers, & Buonanno, 2003, EL: 3; Yeo & Fazio, 2019, EL: 3). Consistent with this point, Rowland's (2014, EL: 1) meta-analysis found a larger effect size for testing (versus restudy) when the

retention interval was longer than a day (Hedges' *g* = 0.69) compared to retention intervals of less than a day (Hedges' *g* = 0.41).

Interestingly, however, there is one circumstance in which testing is *not* more beneficial than restudy; namely when the final test immediately follows practice. Under these circumstances (i.e., "***cramming***" immediately before a test), re-study outperforms retrieval practice (e.g., Roediger & Karpicke, 2006a, EL: 3; Toppino & Cohen, 2009, EL: 4; Wheeler et al., 2003, EL: 3). That is, in the very short term, re-study may be better than testing, but testing is superior over the long term. Since physicians need to retain information over years if not decades, continuous testing will be more beneficial for retention than mere re-study.

Frequency, Length, and Repetition

Given that testing benefits long-term retention, one might ask how much testing we can feasibly ask learners to do. How long should each test be, and is there a point at which additional testing becomes harmful? Some research has suggested the existence of a *list length effect* whereby, as the amount of material to be learned increases (i.e., a longer practice test), the probability of learning any individual item decreases (Cary & Reder, 2003, EL: 3; Gillund & Shiffrin, 1984, EL: 4; Gronlund & Elam, 1994, EL: 4; Ohrt & Gronlund, 1999, EL: 3; Ratcliff, Clark, & Shiffrin, 1990, EL: 4; Strong, 1912, EL: 4). However, others have argued that the list-length effect disappears when various confounds are carefully controlled (Dennis & Humphreys, 2001, EL: 3; Dennis, Lee, & Kinnell, 2008, EL: 3; Kinnell & Dennis, 2011, EL: 3), and, at any rate, the *total* amount of material learned is greater with longer lists (Murayama, Blake, Kerr, & Castel, 2016, EL: 3; Ward, 2002; EL: 4). In sum, there does not appear to be any *cognitive* reason to avoid longer tests, and we recommend that the decision about test length instead be made based on measurement principles, time, and motivational constraints.

A related question concerns the *number* of tests; that is, how many times learners should be tested on the same material. Adding a second test or even more continues to enhance retention above and beyond the first (Roediger & Karpicke, 2006, EL: 3; Karpicke & Roediger, 2007, EL: 4; Pyc & Rawson, 2009, EL: 3; Wheeler & Roediger, 1992, EL: 3; Yang et al., 2021: EL 1). Even if learners answered correctly on the first test, further study can still enhance long-term retention, a strategy known as ***overlearning*** (e.g., Karpicke & Roediger, 2007, EL: 4; Karpicke, 2009, EL: 3; Kornell & Bjork, 2008, EL: 3; Postman, 1965, EL: 4; Pyc & Rawson, 2011, EL: 4; Rawson & Dunlosky, 2011, EL: 3; Vaughn & Rawson, 2011, EL: 3). Overlearning is thought to benefit retention because it provides further feedback and strengthens memory traces to buffer against future forgetting (Driskell, Willis, & Copper, 1992, EL: 1). Relative to the common strategy of dropping items from testing once they have been answered correctly a single time, overlearning has a medium to large benefit on long-term retention, *d* = 0.75 (Driskell et al. 1992, EL: 1).

Recommendation 4-2: Because *more* tests will always be better than *less* tests, considerations for how frequent testing occurs will need to be weighed against practical considerations.

Nevertheless, the benefit from the first test is much larger than the additional benefit from a second test (or from a second episode of practice more generally; Dunlosky & Hertzog, 1997, EL: 4;

Koriat, Sheffer, & Ma'ayan, 2002, EL: 3; Rawson & Dunlosky, 2011, EL: 3; Vaughn & Rawson, 2011, EL: 3; Yang et al., 2021: EL 1). In sum, there is a benefit to continuing to occasionally practice even learned concepts, but most of the benefits from retrieval practice could be realized with just one test.

Spaced Learning: Timing between Tests

When should learners be tested? Cognitive scientists have extensively studied the broader question of when to schedule learning--whether in the form of re-studying or testing. Practicing twice is better than practicing once (Madigan, 1969, EL: 4). But, a second learning session is much more beneficial when learning episodes are spaced over time (*distributed practice*) rather than back-to-back (*massed practice*; i.e., cramming), even when controlling for the total amount of time spent studying. This phenomenon has been often termed the *spacing effect* (Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, EL: 1; Crowder, 1976, EL: 3; Madigan, 1969, EL: 4) but has also been referenced with various other terminology in the literature. Versteeg et al. (2019, EL: 2) note that 'spaced education', 'spaced training', and 'distributed practice' have all been used to describe roughly the same phenomenon. For the purposes of the current review, we will use the term ***spaced learning*****.** Benefits of spaced learning cannot be attributed merely to inattention or boredom with massed study, since spaced learning is still better even when attention is measured and tightly controlled (Zimmerman, 1975, EL: 4). Rather, many contemporary theoretical accounts of spaced learning propose that distributed practice potentiates memory because each subsequent study episode reminds the learner of the previous episode or episodes, re-activating and strengthening them in memory (Benjamin & Tullis, 2010, EL: 4; Bjork & Bjork, 1992, EL: 3; Jacoby & Wahlheim, 2013, EL: 3; McKinley & Benjamin, 2020, EL: 3; Tullis, Benjamin, & Ross, 2014, EL: 3). Further, even when using spaced learning, spacing study episodes with longer gaps (*lags*) in between is generally better than spaced learning with relatively short gaps, which has been separately termed the *lag effect* (Cepeda et al., 2006, EL: 1; Crowder, 1976, EL: 3; Madigan, 1969, EL: 4; Melton, 1967, EL: 4). However, extremely long lags may be harmful (Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler, 2009, EL: 3; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008, EL: 3). The optimal lag depends on the *retention interval*: The longer that learners need to retain what they have learned, the longer the ideal gap in spaced learning (Cepeda et al., 2008, EL: 3). Since physicians generally need to retain their expertise for years if not decades, this implies that distributing practice over a long span of time would result in the most enduring medical knowledge. The spacing and lag effects extend to testing such that, given multiple tests, a longer lag between two tests leads to better retention (Pyc & Rawson, 2009, EL: 3).

Recommendation 4-3: It is preferable to have repeated testing opportunities spread out over time than one big test every five years or ten years but with no intermediary testing opportunities in between.

Test Format and Type of Knowledge

In general, testing appears to be effective across many testing formats (e.g., multiple choice, free recall). Benefits of testing for retrieval have been demonstrated for most basic memory tasks: *recognition* tasks in which the learner merely identifies a stimulus as previously encountered or not (e.g., multiple-choice or yes/no tests, deciding whether you recognize a person; Adesope et al., 2017, EL: 1; Rowland, 2014, EL: 1; Yang et al., 2021, EL: 1), *cued recall* tasks in which the learner supplies partial information in response to a cue (e.g., a fill-in-the-blank test, answering a question asked by a patient; Adesope et al., 2017, EL: 1; Rowland, 2014, EL: 1; Yang et al., 2021, EL: 1; c.f., Hinze & Wiley, 2011, EL: 4), and *free recall* tasks in which the learner must bring to mind information without any guide from the environment (e.g., an essay test; Adesope et al., 2017 EL: 1; Hinze & Wiley, 2011, EL: 4; Rowland, 2014, EL: 1; Yang et al., 2021, EL: 1). Further solidifying the benefits of testing, two meta-analyses (Adesope et al., 2017, EL: 1; Yang et al., 2021, EL 1) formally examined moderated effects of test format (e.g., free recall, cued-recall, short answer, etc.) in their meta-analysis with practice tests and final assessment tests and in both cases--and for all test formats--found a significant and positive effect size. For this reason, the specific format a test item uses is likely of less importance than the quality of the question presentation (e.g., clarity, readability, and veracity of text).

One finding particularly relevant to medical expertise is that testing benefits laboratory *classification* tasks, such as learning to classify different families of birds based on individual photo exemplars (Jacoby, Wahlheim, & Coane, 2010, EL: 4; Siler & Benjamin, 2019, EL: 3), somewhat analogous to diagnosing or classifying patients.

Some controversy has existed as to whether testing benefits more complex knowledge types and tasks, such as problem solving. Some researchers (van Gog & Kester, 2012, EL: 4; van Gog, Kester, Dirkx, Hoogerheide, Boerboom, Verkoeijen, 2015, EL: 4; van Gog & Sweller, 2015, EL: 3; Leahy, Hanham, & Sweller, 2015, EL: 4) have argued that the testing effect does not apply to domains high in "element interactivity"; that is, those in which individual concepts and facts relate tightly to one another. However, this classification has also been criticized (Karpicke & Aue, 2015, EL: 6) as being made on an ad-hoc basis without an independent definition of what constitutes a "complex" task.

Indeed, other researchers report benefits of retrieval practice on seemingly complex tasks, including route-finding (Rohrer et al., 2010, EL: 3), learning mathematical or statistical functions (Kang, McDaniel, & Pashler, 2011, EL: 4; Yeo & Fazio, 2019, EL: 3), troubleshooting problems in a simulated chemical processing plant (Darabi, Nelson, & Palanki, 2007, EL: 4), and--perhaps most relevant to the medical field--resuscitation skills (Kromann, Jensen, & Ringsted, 2009, EL: 4). Further, it is important to note that even in the small number of experiments that did not yield a testing effect, there is generally no *negative* effect of testing on long-term retention, merely an absence of a statistically reliable benefit over repeated study (van Gog, et al., 2015, EL: 4; Leahy et al., 2015 EL: 4; Rawson, 2015, EL: 3; Yeo & Fazio, 2019, EL: 3; c.f., van Gog & Kester, 2012, EL: 4). On the whole, meta-analytic evidence suggests testing benefits complex problem-solving tasks and other types of high-level conceptual knowledge (Yang et al., 2021: EL 1).

In general, then, the testing effect appears to play out for many different formats and types of knowledge--including those relevant to longitudinal medical expertise, such as classification, medical procedures, and the basic formats used in standard computerized testing. Therefore, we should expect

that testing should very likely benefit medical expertise and would certainly fare no worse than alternative strategies.

Ordering of Test Questions

Which material should be practiced when? Cognitive psychologists who have studied this issue have often contrasted two extremes of scheduling material for practice. We follow Brunmair and Richter (2019) by defining a *blocked* schedule as one in which *all* problems or examples pertaining to one topic are presented before moving on to the next topic or concept--similar to the organization of most textbooks or courses in formal education. For instance, a physician may study many examples of hyperthyroidism, then many examples of diabetes. By comparison, an *interleaved* schedule is defined as any ordering in which the to-be-practiced concepts are intermixed such that examples of one category are not fully exhausted before moving onto the next. For example, a physician may review some hyperthyroidism cases and some diabetes cases mixed together (in any order), rather than grouped by diagnosis. Meta-analysis (Brunmair & Richter, 2019, EL: 1) suggests that, for most materials, interleaving results in superior learning than blocking (with the exception of learning vocabulary words). Overall, interleaving results in a medium improvement in learning relative to blocking, with a Hedges' *g* of 0.42, or about the same difference as between testing and restudy. Various intermediate schedules are also possible, such as beginning with blocked practice and then transitioning to interleaved (Yan, Soderstrom, Seneviratna, Bjork, & Bjork, 2017, EL: 5). Preliminary evidence suggests that an intermediate degree of interleaving is optimal in more complex domains, such as when topics are arranged in a hierarchical structure at multiple levels of organization (Yan & Sana, 2021: EL 4) or when individual items can be cross-classified in multiple topics (Abel, Brunmair, Weissgerber, 2021: EL 3).

One reason that interleaving is thought to benefit learning is that it calls attention to the *differences* between concepts (Brunmair & Richter, 2019, EL: 1; Carvalho & Goldstone, 2015, EL: 3; Carvalho & Goldstone, 2017, EL: 3; Kang & Pashler, 2012, EL: 3). For example, learning to distinguish the cause of similar patient symptoms (e.g., shortness of breath could reflect heart problems or lung problems) requires understanding about what the two diagnoses have in common, but especially what differentiates them. Likewise, learning to choose between two treatments that could both be used in a given situation requires understanding why they may both be used, but especially why there is a reason to choose one over the other. One particular recommendation is that, when there is a concern that two diagnoses or two treatments may be confused (i.e., interference, as discussed under Cognitive Skills Need to be Kept Current), those concepts should be interleaved together on the *same* assessment rather than blocked into different assessments. Thus, covering the breadth of the medical field rather than blocking topics is preferable for learning and retention.

Recommendation 4-4: Concepts should not be blocked within the assessment (i.e., having blocks of contiguous questions on a single topic and then never come back to this topic). Concepts that share overlapping features, and may be more prone to misidentification, should especially be interleaved with each other to aid in differentiation/reduce interference.

Transfer of Untested Material

Evidence suggests that retrieval practice supports *transfer*: a benefit to learning not just on the exact tested item, but on related items or material (Carpenter, 2012, EL: 3; Pan & Rickard, 2018, EL: 1; Yang et al., 2021: EL 1). For example, Kang, McDermott, & Roediger (2007, EL: 4) found that retrieval practice transfers between test formats: College students who practiced in the form of multiple-choice questions also showed benefits on a final short-answer test (relative to re-study or no-review conditions), and vice versa (see also Lyle & Crawford, 2011, EL: 4).

Retrieval practice can sometimes also transfer from the practiced information to other, related information. In a college neuroscience course, McDaniel, Anderson, Derbish, and Morrisette (2007, EL: 4) presented students with fill-in-the-blank quiz questions, such as *All preganglionic axons, whether sympathetic or parasympathetic, release ____ as a neurotransmitter*. Practice on these questions benefited subsequent assessment performance even when students were tested on a different piece of information from the same statement, such as *All _____ axons, whether sympathetic or parasympathetic, release acetylcholine as a neurotransmitter*. Indeed, even being tested on part of a science text can sometimes generalize to other, related facts from the text (Chan, 2010 EL: 4; Chan, McDermott, & Roediger, 2006, EL: 4; c.f., Pan & Rickard, 2018 EL: 1; Woolridge, Bugg, McDaniel, & Liu, 2014, EL: 3). Collectively, this body of evidence suggests that, at the very least, near transfer occurs in response to testing, where individuals benefit from prior testing episodes to related but different situations.

Finally, retrieval practice can transfer between levels of knowledge or analysis (Agarwal et al., 2013, EL: 4; Butler, 2010, EL: 3; Pan & Rickard, 2018, EL: 1; Rohrer et al., 2010, EL: 3). For example, practicing the notion of *competition* with a definition question ("What is the term for when two or more organisms vie for limited environmental resources?") also benefits application (e.g., "A group of 500 pandas are living in a reserve. Recent dry weather has reduced the bamboo population, which the pandas rely on. The pandas are in what type of relationship?"), and vice versa (Agarwal et al., 2013, EL: 4). These results imply that learners who use testing are not just memorizing the answers to specific test items; they are developing their understanding of the concept more broadly.

Recommendation 4-5: Being tested should improve physicians' retention not just of the specific tested material, but of other related material as well.

Individual Differences

More recent work has begun to examine whether the testing effect applies equally across groups of learners. Meyer and Logan (2013, EL: 4) found that older adults benefit from testing just as much as college-age learners. This finding is relevant to longitudinal assessment of medical expertise

because it suggests that testing may be beneficial even for physicians more advanced in their career and further removed from training.

One question that has been of particular interest is how the testing effect may be modulated by general academic aptitude or knowledge. On one hand, the boost provided by testing may be especially helpful for students who would otherwise struggle. Thus, a larger testing effect has sometimes been observed for learners lower in the ability to hold information in active memory (*working memory capacity*; Agarwal, Finley, Rose, & Roediger, 2017, EL: 5; c.f. Wiklund-Hörnqvist et al. 2014, EL: 5), reading comprehension (Callender & McDaniel, 2007, EL: 5), or general intelligence (Brewer & Unsworth, 2012, EL: 5). Working memory typically declines with age (Park, Lautenschalger, Hedden, Davidson, Smith, & Smith, 2002, EL: 5). As such, this factor may make testing particularly important for older physicians. On the other hand, the more poorly that learners perform on practice tests, the less they benefit from retrieval practice (Rowland, 2014, EL: 2); thus, low-performing students may do better by using other strategies (Carpenter, Lund, Coffman, Armstrong, Lamm, & Reason, 2016, EL: 5). Because physicians may be expected to be relatively high-performing learners, insofar as they were accepted to and graduated medical school, we might *not* expect them to be in this low-performing group that would not benefit from such testing. Nevertheless, future research should further explore the impact that individual differences have on learning outcomes from different forms of interventions.

Feedback

When learners are tested (either practice tests or final assessments), most will answer some of the items that they have studied or practiced correctly, but make errors on others. One concern sometimes expressed by educators (and learners) is that these self-generated errors may become (falsely) incorporated into learners' knowledge base, and perhaps it would preferable to adopt a more didactic approach that prevents learners from making mistakes (e.g., *errorless learning*; for further discussion, Metcalfe, 2017, EL: 3; Middleton & Schwartz, 2012, EL: 2).

Not surprisingly, evidence affirms that the degree to which testing enhances learning depends on how well learners perform on the test (Rowland, 2014, EL: 1). When no feedback is provided during testing, individuals receive a positive memory boost for correctly recalled information (Kornell, Bjork, & Garcia, 2011, EL: 3; Rowland, 2014, EL: 1; Spellman & Bjork, 1992, EL: 6). However, for the items with weak memory strength that are not correctly recalled on the no-feedback test, no memory boost occurs. In this way, tests without feedback may create an asymmetry or *bifurcation* in learning dependent upon memory strength (for individual pieces of information) prior to testing. In contrast, during restudy conditions, participants receive a memory boost for all items reviewed; however, it is a weaker boost than received for correctly recalled items in the test condition (but nevertheless greater than the non-recalled items in the "no feedback test" condition). This asymmetry can be alleviated by the addition of feedback after a retrieval practice attempt. Thus, although testing is beneficial even without feedback, testing *with* feedback is even better (Butler & Roediger, 2008, EL: 4; Rowland, 2014, EL: 1; Yang et al., 2021: EL 1; c.f., Adesope et al., 2017, EL: 1).

So long as feedback is given, errors generated by learners in using testing do not impair long-term performance (Butler, Karpicke, & Roediger, 2008, EL: 3; Huelser & Metcalfe, 2012, EL: 3; Kang, Pashler, Cepeda, Rohrer, Carpenter, & Mozer, 2011, EL: 3; Kornell, Klein, & Rawson, 2015, EL: 3; Kornell,

Hays, & Bjork, 2009, EL: 3; Kornell & Metcalfe, 2014, EL: 4; Metcalfe, 2017, EL: 3; Metcalfe & Kornell, 2007, EL: 4; Richland, Kao, & Kornell, 2009 EL: 3; c.f., Knight, Ball, Brewer, DeWitt, & Marsh, 2012, EL: 3, for more mixed results). Indeed, an unsuccessful retrieval attempt followed by feedback is *more beneficial* than simply reading the correct information without attempting retrieval (Kornell, Hays, & Bjork, 2009, EL: 4; Hays, Kornell, Bjork, 2013, EL: 4; Richland, Kornell, Kao, 2009 EL: 4). Therefore, the concern that errors made during learning undermine long-term knowledge is unfounded so long as feedback is given. Further, because corrective feedback allows learners to learn from even difficult tests, when feedback is used, learners can be presented with more challenging and demanding tests (e.g., short answer rather than multiple choice) that lead to better learning (Kang, McDermott, & Roediger, 2007, EL: 3). Thus, training that permits errors can be more effective than errorless learning (Keith & Frese**,** 2008, EL: 1) because it allows learners to capitalize on testing and practice effects. At a practical level, these findings imply that tests will most benefit physicians' retention of medical expertise if (a) feedback is given, especially for more difficult material, and (b) tests are appropriately challenging/difficult.

Benefits of Feedback

Why is feedback so effective at ameliorating errors? One reason is that, when an error is committed with high confidence, the resulting negative feedback can be especially memorable (the *hypercorrection effect*; Butler, Fazio, & Marsh, 2011, EL: 4; Butterfield & Metcalfe, 2001, EL: 5; Butterfield & Metcalfe, 2006, EL: 5; Cyr & Anderson, 2012, EL: 5; Fazio & Marsh, 2009, EL: 5; Fazio & Marsh, 2010, EL: 5; Iwaki, Matsushima, & Kodaira, 2013, EL: 5; Metcalfe, 2017, EL: 3; Metcalfe & Finn, 2011, EL: 5; Sitzman, Rhodes, Tauber, & Liceralde, 2015, EL: 5), although perhaps not for older (retirement-age) adults (Eich, Stern, & Metcalfe, 2013, EL: 5; Sitzman et al., 2015, EL: 5; c.f., Cyr & Anderson, 2012, EL: 5). The importance of such hyper corrective feedback accords with multiple theoretical perspectives on cognitive science, such as *error-based learning* views in which learning occurs to the degree that preceding expectations are incorrect (*prediction error*; e.g., Clark, 2013, EL: 3; Dell & Chang, 2014, EL: 3; Rumelhart & McClelland, 1986, EL: 3), and Bayesian views in which cognition can be viewed as updating a set of beliefs in accordance with the experienced "data" or world (e.g., Frank & Goodman, 2012, EL: 4; Jacobs & Kruschke, 2010, EL: 3; Tenenbaum, Kemp, Griffiths, & Goodman, 2011, EL: 3). Thus, feedback may be particularly effective in alleviating *intrusions*--the false "recall" of incorrect information--rather than failures to recall anything at all (Butler & Roediger, 2008, EL: 4). In other words, it is especially important to give feedback when learners respond incorrectly rather than when they decline to respond.

A second advantage of feedback, converse to the hypercorrection effect, is that if the learner *is* correct, but has low confidence in their response (e.g., a "lucky guess"), the presence of feedback increases the probability that this correct response will be retained later (Agarwal, Bain, & Chamberlain, 2012, EL: 4; Butler et al., 2008, EL: 3; Fazio, Huelser, Johnson, & Marsh, 2010, EL: 3; c.f., Pashler, Cepeda, Wixted, & Rohrer, 2005, EL: 4). Additionally, feedback is unlikely to negatively affect learning. However, when a learner is highly confident in their response *and correct*, feedback may be redundant (Hays, Kornell, & Bjork, 2010, EL: 4; Karpicke & Roediger, 2008, EL: 4). Thus, a recommendation would be to

provide feedback for correct as well as incorrect responses as the benefits outweigh any possible drawbacks.

Best Form of Feedback

Given the general value of feedback, questions emerge about what feedback should be provided and when. The form of the feedback clearly matters: Simply stating whether a response is correct or incorrect (*verification feedback*) confers little or no benefit whereas presenting the actual, correct answer benefits learning (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991, EL: 1; Fazio et al., 2010, EL: 3; Hausmann, Vuong, Towle, Fraundorf, Murray, & Connelly, 2013, EL: 5; Metcalfe, 2017, EL: 3; Moreno, 2004, EL: 3; Pashler et al., 2005, EL: 4; Whyte, Karolick, Neilsen, Elder, & Hawley, 1995, EL: 4) although this difference may be qualified by the learner's knowledge level (Hausmann et al., 2013, EL: 5).

Recommendation 4-6: Feedback should be provided for both correct and incorrect responses on a test.

A few studies have also examined the effects of additional elaborations that can be provided beyond correct-answer feedback. One popular technique is to present an explanation of why the correct answer is correct; however, most studies have found that such *explanatory feedback* does not appear to yield gains over providing the correct answer alone (Bangert et al., 1991, EL: 1; Corral & Carpenter, 2020, EL: 4; Kulhavy, White, Topp, Chan, & Adams, 1985, EL: 4; Mandernach, 2005, EL: 4; Smits, Boon, Sluijsmans, & van Gog, 2008, EL: 4; Whyte et al., 1995, EL: 4; c.f., Butler, Godbole, & Marsh, 2013, EL: 3, for somewhat more mixed results) and, indeed, providing additional feedback to read may be less efficient overall (Kulhavy et al., 1985, EL: 4). On the other hand, one study suggests that providing *examples* of an incorrectly-understood concept can enhance learning beyond presenting the answer alone (Finn, Thomas, & Rawson, 2018, EL: 3), but, to date, there is not much research on this approach. There is evidence that feedback should include the correct answer, not merely whether the learner answered it correctly or not, but further explanation beyond may be unnecessary.

Gnepp, Klayman, Williamson, and Barlas (2020, EL: 3) found evidence that individuals may be skeptical of negative feedback when the feedback provider's accuracy or credentials are in question. However, this study examined workplace feedback from a manager, and it likely contrasts with the relative objectivity offered by an automated system providing feedback about errors. Still, this evidence suggests there may be value to citing information sources in feedback to add a sense of authority and objectivity.

Recommendation 4-7: Feedback provided on tests should include evidence and/or references with claims.

Timing of Feedback

Some work has also examined the timing of feedback, generally contrasting immediate feedback with feedback that is delayed to some degree. In controlled laboratory studies, feedback delayed by several hours or days is often more effective than immediate feedback (Butler & Roediger, 2008, EL: 4; Kulik & Kulik, 1988, EL: 1; Schmidt & Bjork, 1992, EL: 3; Schooler & Anderson, 1990, EL: 4), or at least no worse (Kang et al., 2011, EL: 4; Metcalfe, Kornell, & Finn, 2009, EL: 4; Smits et al., 2008, EL: 4). Delayed feedback may better potentiate long-term retention and learning because it encourages learners to develop their own monitoring and self-assessment skills, rather than relying exclusively on external feedback (Schmidt, Young, Swinnen, & Shapiro, 1989, EL: 4). On the other hand, in *in vivo* classroom studies, the reverse seems to be true: immediate feedback is better than delayed (Kulik & Kulik, 1988, EL: 1; Lemley, Sudweeks, Howell, Laws, & Sawyer, 2007, EL: 4). This reversal has been attributed to the fact that, in a busy classroom environment, students may not even attend to feedback when it is delayed as their priorities may have since shifted (Kulik & Kulik, 1988, EL: 1; Metcalfe, 2017, EL: 3). What does this imply for longitudinal assessment of medical expertise? Given that physicians are likely motivated to attend to the feedback they receive, the literature above suggests that delayed feedback may be superior, but there is a need to test this specifically within the medical domain. Some evidence does suggest that a particularly effective strategy may be to interleave periods of testing with periods of restudy so that learners can restudy material they answered incorrectly (McDaniel, Bugg, Liu, & Brick, 2015, EL: 4; Metcalfe & Miele, 2014; EL: 4), then incorporate the corrected information into their next retrieval attempt.

Training People to Use Retrieval Practice

Most work on the testing effect has been focused on testing administered by educators and professional organizations. However, learners can also choose to test themselves as a learning strategy. Indeed, college students who report using more retrieval practice in their own self-regulated learning have higher GPA (Hartwig & Dunlosky, 2012, EL: 5), and learners who choose to employ more testing in laboratory studies show better retention (Karpicke, 2009, EL: 5). Although correlational, when combined with the RCT evidence for the testing effect described above, this result suggests that there exists meaningful variation among learners in whether they employ retrieval practice and that this variation has consequences for their academic performance and knowledge.

As a result, some researchers have sought to test whether it is possible to teach learners to use testing approaches for learning. Some evidence suggests that individuals who have more formal education in cognitive psychology (McCabe, 2011, EL: 5), who receive narratives from other learners who adopted retrieval practice (Hui, de Bruin, Donkers, & van Merriënboer, 2021a: EL 4), or who are assigned practice that allows them to experience the testing effect (Ariel & Karpicke, 2017, EL: 4; Einstein, Mullet, & Harrison, 2012, EL: 5; Hui, de Bruin, Donkers, & van Merriënboer, 2021b: EL 3; Tullis, Finley, & Benjamin, 2013, EL: 4) come to appreciate the value of testing and incorporate it into future study plans. A workshop specifically designed to teach retrieval practice as a study strategy has shown success in increasing both college students' intention to apply retrieval practice and their resulting assessment performance (Stanger-Hall, Shockley, & Wilson, 2011, EL: 4).

Research on student study behaviors within health profession education have provided evidence that students may not always know or engage in optimal study behaviors. Several recent studies found medical students do not prioritize retrieval practice (Coker et al., 2018; EL: 5; Jouhari, Haghani, & Changiz, 2016, EL: 5; Piza et al., 2019, EL: 5). Coker et al. (2018, EL: 5) found that 90% of surveyed pharmacy students believed their learning would benefit from regular retrieval practice, but only 60% of these students stated they actually engage in retrieval practice behavior. Interestingly, 85% of the pharmacy students stated they would like to see a retrieval practice system put in place, potentially speaking to the value of an external structure that places pressure for optimal study behaviors to occur. These findings mirror other studies, including Piza et al. (2019, EL: 5), which found that the majority of health profession students used study techniques that went against evidence-based study principles (e.g., not returning to material after a course has ended, re-reading highlighted text in books rather than quizzing oneself on it). In addition, Piza et al. found that the majority of the health profession faculty surveyed also held misconceptions about evidence-based study practices.

However, other studies have found that medical students do use good study habits. In a survey of medical students who finished a two-year preclinical curriculum, Deng, Gluckstein, and Larsen (2015, EL: 5) found that all study participants used practice multiple choice questions during studying. Additionally, practicing more multiple choice questions predicted greater performance on a medical licensing examination. These conflicting findings may speak to the varying student-body cultures across medical schools that can influence study habits. Whatever the cause, the contrasting findings suggest that greater standardization of evidence-based study habits across medical schools would likely benefit medical students.

West and Sadoski (2011, EL: 5) examined predictors of success in the first semester of medical school. They found that the two study skills of time management and self-testing, as measured by the Learning and Study Strategies Inventory (LASSI), better predicted success during the first year than the general aptitude measures of undergraduate GPA and MCAT scores. Baatar, Lacy, Mulla, Piskurich (2017, EL: 5) studied the relation of optional practice tests to first year medical student classroom performance in a correlational study. On four of the five assessments in the course, students who used the practice tests (typically, about 60%) outperformed the students who did not. Students who used *every* practice test throughout the semester scored much higher on the final assessment than those who did not.

Not all studies have demonstrated positive benefits of retrieval practice in the health sciences. LaDisa and Biesboer (2017, EL: 5) did not find a benefit in student learning after introducing a retrieval practice intervention within a pharmacotherapy course. However, this study used a cross-sectional design (i.e., the comparison groups were two courses, at two different points, with different students), limiting its ability to draw causal conclusions. This study also found that most students did not find the additional retrieval practice intervention helpful, which may speak to individuals' negative perception of testing more generally.

Retrieval practice is an important aspect of evidence-based study practices; however, it is not the only important aspect to consider. Burk-Rafel, Santen, and Purkiss (2017, EL: 5) surveyed medical students' study behaviors prior to the United States Medical Licensing Examination (USMLE) Step 1, during a designated assessment study period, and found that students were studying an average of 11 hours a day over a period of, on average, 35 days. This intensive studying, akin to "cramming," is likely

less effective than studying behavior that takes place over a longer period of time. Supporting this, Burk-Rafel et al. found that studying earlier than the designated study period, as well as spending more time reviewing books and attempting practice questions, were associated with higher USMLE Step 1 scores, even when controlling for prior MCAT scores. Physicians then should be guided to experience the learning benefits of self-testing, as this will help them adopt more effective study and learning procedures beyond the test itself.

Cognitive Mechanisms Underlying the Testing Effect

Understanding *how* and *why* retrieval practice works is important for applying it across situations: A strong theoretical account of retrieval practice generates predictions about when and where it should be useful, rather than requiring each new application (e.g., each new test format, subject matter, or group of learners) to be tested afresh. Further, a clear explanation of retrieval practice facilitates outreach to learners and educators.

The testing effect is consistent with several broad, long-standing cognitive principles. The benefits of practicing retrieval can be seen as an instance of *transfer-appropriate processing* (Roediger & Blaxton, 1987, EL: 3; Roediger & Butler, 2011, EL: 3). The activities that make for the most effective learning are generally those that match the way the material will be used later. For example, reading the driver's manual would be ideal practice for taking a written driver's exam, whereas behind-the-wheel experience would be ideal practice for actually driving. The best way to potentiate later retrieval, then, is to practice retrieval itself, rather than re-read or performs other activities that are less closely related. Supporting this account, multiple meta-analyses (Adesope et al., 2017, EL: 1; Yang et al., 2021, EL: 1; c.f., Rowland, 2014, EL: 1) have found similarity of initial and final test moderates the testing effect. When practice tests and final tests used identical test formats, a larger testing effect occurs, compared to when practice tests and final tests were in different formats (although both designs resulted in a memory boost for participants).

However, this value of testing may not always be obvious to learners (or educators). Although testing facilitates long-term retention, it may require initial processing that is less accurate or burdensome, as learners struggle with practice questions and sometimes answer them erroneously or not at all. Thus, retrieval practice can be viewed as a *desirable difficulty*: the principle that conditions that facilitate retention, including practicing retrieval, are often *more* difficult during initial acquisition (Schmidt & Bjork, 1992, EL: 3). As we note above, for immediate tests, testing is generally *less* effective than restudy, and it is only over the long-term that the benefits of testing emerge. More generally, performance during initial learning is not necessarily a reliable index of long-term learning (Soderstrom & Bjork, 2015, EL: 3). This principle is counter-intuitive to many learners, in part perhaps because many learners view retrieving information from memory as a process distinct from learning (Karpicke, Butler, & Roediger, 2009, EL: 5; Kornell & Bjork, 2007, EL: 5; Kornell & Son, 2009, EL: 5; Yan, Thai, & Bjork, 2014, EL: 5). In this view, practicing retrieval may help you identify what you do and do not know, but it does not potentiate learning in and of itself. An analogy that supports this naive theory might be that saving a computer file ("learning") and opening a file ("retrieval") are distinct, independent processes. However, the human brain does not operate exactly like a computer, and this naive "storehouse" theory is

inconsistent with another broad-standing principle of memory (Karpicke, 2012, EL: 6). Retrieval is in fact a potent *modifier* of memory (Anderson, Bjork, & Bjork, 1994, EL: 4) such that each retrieval event itself alters the state of the memory system by making some information more accessible to future retrieval. Psychological scientists have noted the similarity of this phenomenon to the observer effect in physics, where the mere act of observing a particle can alter its condition; similarly, the mere act of retrieving a memory alters its condition as well (Roediger & Karpicke, 2006b; Spellman & Bjork, 1992).

More recently, researchers have investigated the cognitive mechanisms involved in testing in particular. One reason that testing may benefit retention is that it may increase the number of ways that people can bring to mind the to-be-remembered information (e.g., Bjork, 1975, EL: 3; McDaniel & Masson, 1985, EL: 3; Pyc & Rawson, 2010, EL: 4; Rowland & DeLosh, 2014, EL: 3). For example, it may promote the development of *mediators* between the retrieval environment and the to-be-retrieved material (Pyc & Rawson, 2010, EL: 4). That is, given the need to remember the stages of mitosis (the environment or cue), one might remember *PMAT* (the mediator) in order to retrieve *protophase, metaphase, anaphase, telophase* (the to-be-retrieved target). More generally, retrieval practice may lead learners to *elaborate* on the to-be-remembered information, bringing to mind additional related information (Carpenter, 2009, EL: 4), which is generally an effective learning technique (Anderson & Reder, 1979, EL: 3). Another, possibly overlapping mechanism, may be that retrieval practice enhances the distinctiveness of individual learning episodes (Kuo & Hirshman, 1997, EL: 3; Peterson & Mulligan, 2013, EL: 3; Lehman, Smith, & Karpicke, 2014, EL: 4). For example, the life cycle of the malaria parasite comprises multiple stages, including *sporozoites* and *merozoites*, that psychologically can be easily confused; however, practicing retrieving them from memory makes them more distinct.

Although there remains work to be done to specify the exact cognitive mechanism(s) that underlie the testing effect, the extant literature already supports at least one theoretical conclusion: The testing effect is not an isolated phenomenon. Rather, it follows from broad principles of memory and cognition (transfer-appropriate processing, desirable difficulty, and retrieval as a modifier of memory) and can take effect through general cognitive mechanisms (elaboration, distinctiveness, and mediators). Because the testing effect is linked to general psychological principles, it is likely to be applicable across a variety of domains and populations, including retention of medical expertise. Nevertheless, the principle of transfer-appropriate processing also implies that the type of testing and retrieval practice that will be *most* beneficial is that which most closely resembles the desired outcome; e.g., practicing diagnosis to benefit diagnostic expertise.

Chapter Summary

The benefits of testing for learning have been known for over a hundred years and are supported across many domains by a robust literature. The act of retrieving information from one's memory potentiates memory, boosting subsequent recall ability, and proves to be a superior learning method to restudy, concept mapping, and many other educational techniques. The benefits of testing can be further improved by leveraging several important moderator variables: retention intervals, number of tests, test timing (i.e., spacing), interleaving of materials, and feedback. The positive effects of testing can be reinforced by increasing the retention interval length. Although it is challenging to

know exactly when a subsequent test should occur, given that physicians are expected to retain their knowledge over the course of an entire career (i.e., several decades), longer retention intervals should be prioritized over shorter intervals.

Engaging multiple tests can further boost learning beyond the baseline benefits of testing alone, although initial tests are more beneficial than subsequent tests. The literature we encountered supports the recommendation that there is *no harm* to longer and more frequent tests, albeit they inconvenience the test taker, but there is evidence of continued benefits of additional testing. In response, we recommend that the amount of testing should be limited by physician time and motivational constraints.

It is important for gaps to be placed between testing sessions in order to maximize learning outcomes (i.e., the spacing effect). Having tests distributed over time, versus in a contiguous block, should be a key feature to any longitudinal assessment program. A variety of item formats (e.g., short-answer, multiple-choice, etc.) have all been shown to benefit from testing. For this reason, the specific format a test item uses is likely of less importance than the presentational quality of the question (e.g., clarity, readability, and veracity of text). Despite the presence of some controversy as to whether the benefits of testing are limited to simple knowledge types (e.g., rote memorization of facts), evidence exists to support improvement in more complex tasks (e.g., problem solving) as well. Additionally, there is no reliable evidence in the literature we reviewed that testing ever hurts memory retention.

Interleaving designs have been shown to result in superior performance relative to blocked designs. Having topics switch from item to item is more beneficial than many questions about one topic before switching to a new topic (i.e., blocking). Furthermore, interleaving may be especially beneficial for easily confused topics. For this reason, we recommend using interleaving to bolster cognitive skills and knowledge for targeted areas within medicine (e.g., when two distinct conditions share similar symptoms). A critical goal for any longitudinal assessment program is that the benefits of testing extend beyond future tests and include performance-related outcomes in a physician's medical practice. To this end, the reviewed literature supports the benefits of being tested on different, but related information to that which was tested previously.

Feedback serves an important function in testing and is recommended for inclusion in any longitudinal assessment framework. The presence of feedback can allay concerns over errors generated during a test, and feedback is especially important when learners respond incorrectly (although feedback also allows learners to improve when they decline to respond). An unsuccessful retrieval attempt followed by feedback is more beneficial than simply reading the correct information without attempting retrieval. In instances where the learner is correct, but has low confidence in their response (e.g., a "lucky guess"), feedback increases the likelihood that the correct response will be later remembered. Because corrective feedback allows learners to learn from even difficult tests, when feedback is used, learners can be presented with more challenging and demanding tests. When providing feedback to a learner, explanations for correct/incorrect answers have not been reliably shown to aid learning beyond simply providing the correct response; however, the use of examples during feedback may be beneficial and is worth further investigation (see study proposal below). Additionally, the presence of citations for sources of information and reference materials may also be beneficial. The question of when a learner should receive feedback is an important one. We found evidence that delayed feedback may be superior to immediate feedback; however due to sparse

evidence in applied domains, we believe this should be tested within medicine and propose such a study below.

A benefit to creating a longitudinal assessment program may be that it results in physicians adopting more effective study and learning habits as they are guided to experience the learning benefits of self-testing.

Future Directions

In reviewing this literature, we identified a number of directions for future research that were specifically lacking in the literature or not seen applied to the physician population of continuing certification. These include four proposed studies where we supply the ideas but not necessarily a fully developed research design.

1. *What Type of Explanation of the Correct Answer in the Feedback Provided Improves Learning and Retention?*

There is some evidence that providing explanations during feedback does not reliably benefit the learner beyond simply receiving the correct response. However, the evidence is not strong in explaining why. A study that manipulates the type of explanations provided during feedback may offer insight into how to improve feedback. One possible design is to provide one group of learners with feedback that uses concrete examples to illustrate a point in addition to providing a technical explanation of how the item should be answered. A comparison group would only receive an explanation, but no illustrative example. Learning outcomes would be compared at a later date.

2. *What is the Optimal Timing for Delivering Feedback for Improving Learning and Retention?*

The optimal timing for feedback to be provided is ambiguous. Evidence from controlled environments demonstrates value to delayed feedback; however, we were unable to find complementary evidence from applied domains. Given this asymmetry, one recommendation is to run an experiment on this topic. One design is to give immediate feedback for some questions and delayed feedback for other questions, and compare the questions within an individual learner. Delayed feedback can be designed to provide learners feedback either after the test, 1 day later, or 1 week later.

3. *Is Interleaving of Different Content Domains Beneficial to Learning and Retention?*

Meta-analytic evidence for the benefits of interleaving comes primarily from classroom and laboratory studies with a relatively small number of categories or concepts (e.g., formula to find the volume of four different types of mathematical solids). However, the number of concepts to be tested on continuing certification program assessments is far larger. Given the hypothesis that interleaving promotes learning by facilitating contrast between confusable concepts, intermixing *all* concepts on continuing certification program assessments may not be optimal because related concepts are unlikely to be adjacent. Indeed, some recent studies (Abel et al., 2021: EL 3; Yan et al., 2021: EL 4) suggest that, in more complex domains, an intermediate rather than maximal degree of interleaving may be optimal for learning precisely because it better facilitates such discriminative contrast. However, this evidence is still early. Thus, we propose comparing the learning benefits of a fully random intermixing of topics

versus an order constructed so that potentially confusable topics appear in close proximity. We expect this latter schedule to yield the best long-term learning.

4. *Are Citations Important Elements of Feedback to Improve Learning and Retention?*

Assessments often provide citations alongside evidence for a claim. Some pertinent questions, then, are whether citations benefit learners during feedback, and if so, why. One possibility is that merely having citations builds confidence in the evidence. Another possibility is that the citations are only helpful if physicians actually read the reference. If the testing interface allowed for users to save and/or follow references, log data could be collected to measure these behaviors. The extent to which users engaged with references could be used to predict future performance and provided insight into the value of citations within tests.

References

Abel, R., Brunmair, M., & Weissgeber, S. C. (2021). Change one category at a time: Sequence effects beyond interleaving and blocking. *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Advance online publication.

Abott, E. E. (1909). On the analysis of the factor of recall in the learning process. *The Psychological Review: Monograph Supplements*, *11*(1), 159-177.

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, *87*(3), 659-701.

Agarwal, P. K., Bain, P. M., & Chamberlain, R. W. (2012). The value of applied research: Retrieval practice improves classroom learning and recommendations from a teacher, a principal, and a scientist. *Educational Psychology Review*, *24*(3), 437-448.

Agarwal, P. K., Finley, J. R., Rose, N. S., & Roediger III, H. L. (2017). Benefits from retrieval practice are greater for students with lower working memory capacity. *Memory*, *25*(6), 764-771.

Agarwal, P. K., Roediger, H. L., McDaniel, M. A., & McDermott, K. B. (2013). How to use retrieval practice to improve learning. *Saint Louis, MO: Washington University in St. Louis*.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1063-1087.

Anderson, J. R., & Reder, L. M. (1979). An elaborative processing explanation of depth of processing. *L.; S. Cermak and FIM Craik, Eds., Levels of Processing in Human Memory (Erlbam, 1979)*, 385-404.

Ariel, R., & Karpicke, J. D. (2017). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied*, *24*(1), 43–56.

Armson, H., Roder, S., Wakefield, J., & Eva, K. W. (2020). Toward practice-based continuing education protocols: Using testing to help physicians update their knowledge. *Journal of Continuing Education in the Health Professions, 40*(4), 248-256.

Baatar, D., Lacy, N. L., Mulla, Z. D., & Piskurich, J. F. (2017). The impact of integration of self-tests into a pre-clerkship medical curriculum. *Medical Science Educator*, *27*(1), 21–27.

Bangert-Drowns, R. L., Kulik, C. L. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213-238.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective?. *Cognitive Psychology*, *61*(3), 228-247.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, *2*, 35-67.

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory and Language*, *66*(3), 407-415.

Brown, D. (2017). An evidence-based analysis of learning practices: the need for pharmacy students to employ more effective study strategies. *Currents in Pharmacy Teaching and Learning*, *9*(2), 163–170. doi:10.1016/j.cptl.2016.11.003

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029-1052.

Burk-Rafel, J., Santen, S. A., & Purkiss, J. (2017). Study Behaviors and USMLE Step 1 Performance: Implications of a Student Self-Directed Parallel Curriculum. *Academic Medicine: Journal of the Association of American Medical Colleges*, *92*(11), S67–S74. doi:10.1097/ACM.0000000000001916

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(5), 1118-1133.

Butler, A. C., Fazio, L. K., & Marsh, E. J. (2011). The hypercorrection effect persists over a week, but high-confidence errors return. *Psychonomic Bulletin & Review*, *18*(6), 1238-1244.

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology*, *105*(2), 290-298.

Butler, A. C., Karpicke, J. D., & Roediger III, H. L. (2008). Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 918-928.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition*, *36*(3), 604-616.

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1491-1494.

Butterfield, B., & Metcalfe, J. (2006). The correction of errors committed with high confidence. *Metacognition and Learning*, *1*(1), 69-84.

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology*, *99*, 339–348.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(6), 1563-1569.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science*, *21*(5), 279-283.

Carpenter, S. K., Lund, T. J., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review*, *28*(2), 353-375.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*(3), 474-478.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *23*(6), 760-771.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, *36*(2), 438-448.

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*(1), 281-288.

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699-1719.

Cary, M., & Reder, L. M. (2003). A dual-process account of the list-length and strength-based mirror effects in recognition. *Journal of Memory and Language*, *49*(2), 231-248.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*(4), 236-246.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*(3), 354-380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*(11), 1095-1102.

Chan, J. C. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*(1), 49-57.

Chan, J. C., McDermott, K. B., & Roediger III, H. L. (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General*, *135*(4), 553-571.

Chesluk, B. J., Eden, A. R., Hansen, E. R., Johnson, M. L., Reddy, S. G., Bernabeo, E. C., & Gray, B. M. (2019). How physicians prepare for Maintenance of Certification exams: a qualitative study. *Academic Medicine*, *94*(12), 1931-1938.

Cilliers, F. J. (2015). Is assessment good for learning or learning good for assessment? A. Both? B. Neither? C. It depends? *Perspectives on Medical Education*, *4*(6), 280–281.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181-204.

Coker, A. O., Lusk, K. A., Maize, D. F., Ramsinghani, S., Tabor, R. A., Yablonski, E. A., & Zertuche, A. (2018). The effect of repeated testing of pharmacy calculations and drug knowledge to improve knowledge retention in pharmacy students. *Currents in Pharmacy Teaching and Learning*, *10*(12), 1609-1615.

Corral, D., & Carpenter, S. K. (2020). Facilitating transfer through incorrect examples and explanatory feedback. *Quarterly Journal of Experimental Psychology*, *73*(9), 1340-1359.

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

Cyr, A. A., & Anderson, N. D. (2012). Trial-and-error learning improves source memory among young and older adults. *Psychology and Aging*, *27*(2), 429-439.

Darabi, A. A., Nelson, D. W., & Palanki, S. (2007). Acquisition of troubleshooting skills in a computer simulation: Worked example vs. conventional problem solving instructional strategies. *Computers in Human Behavior*, *23*(4), 1809-1819.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 1-9.

Deng, F., Gluckstein, J. A., & Larsen, D. P. (2015). Student-directed retrieval practice is a predictor of medical licensing examination performance. *Perspectives on Medical Education*, *4*(6), 308–313. doi:10.1007/s40037-015-0220-x

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, *108*(2), 452–478.

Dennis, S., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language*, *59*(3), 361-376.

Driskell, J. E., Willis, R. P., & Copper, C. (1992). Effect of overlearning on retention. *Journal of Applied Psychology*, *77*(5), 615-622.

Dunlosky, J., & Hertzog, C. (1997). Older and younger adults use a functionally identical algorithm to select items for restudy during multitrial learning. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *52*(4), P178-P186.

Ebbinghaus, H. (1885). Ueber das Gedächtnis.

Eich, T. S., Stern, Y., & Metcalfe, J. (2013). The hypercorrection effect in younger and older adults. *Aging, Neuropsychology, and Cognition*, *20*(5), 511-521.

Einstein, G. O., Mullet, H. G., & Harrison, T. L. (2012). The testing effect: Illustrating a fundamental concept and changing study strategies. *Teaching of Psychology*, *39*(3), 190-193.

Fazio, L. K., Huelser, B. J., Johnson, A., & Marsh, E. J. (2010). Receiving right/wrong feedback: Consequences for learning. *Memory*, *18*(3), 335-350.

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review*, *16*(1), 88-92.

Fazio, L. K., & Marsh, E. J. (2010). Correcting false memories. *Psychological Science*, *21*(6), 801-803.

Finn, B., Thomas, R., & Rawson, K. A. (2018). Learning more from feedback, EL:aborating feedback with examples enhances concept learning. *Learning and Instruction*, *54*, 104-113.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998-998.

Fung, J. N. M., Joegi, A., & Fung, Y. K. (2019). Medical students' perspective: Influences on the choice of learning strategies. *Medical Teacher*, *42*(6), 713.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1-67.

Gnepp, J., Klayman, J., Williamson, I. O., & Barlas, S. (2020). The future of feedback: Motivating performance improvement through future-focused feedback. *PloS one*, *15*(6), e0234444.

Griffith, M., Purkiss, J., Santen, S. A., & Burk-Rafel, J. (2017). Creating an Evidence-Based Advising Program for Exams: a Student-led 10-Step Approach. *Medical Science Educator*, *27*(4), 877–880.

Gronlund, S. D., & Elam, L. E. (1994). List-length effect: Recognition accuracy and variance of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1355-1369.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement?. *Psychonomic Bulletin & Review*, *19*(1), 126-134.

Hausmann, R. G., Vuong, A., Towle, B., Fraundorf, S. H., Murray, R. C., & Connelly, J. (2013, July). An evaluation of the effectiveness of just-in-time hints. In *International Conference on Artificial Intelligence in Education* (pp. 791-794). Springer, Berlin, Heidelberg.

Hays, M. J., Kornell, N., & Bjork, R. A. (2010). The costs and benefits of providing feedback during learning. *Psychonomic Bulletin & Review*, *17*(6), 797-801.

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 290-296.

Hopkins, R. F., Lyle, K. B., Hieb, J. L., & Ralston, P. A. (2016). Spaced retrieval practice increases college students' short-and long-term retention of mathematics knowledge. *Educational Psychology Review*, *28*(4), 853-873.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory*, *19*(3), 290-304.

Huelser, B. J., & Metcalfe, J. (2012). Making related errors facilitates learning, but learners do not know it. *Memory & Cognition*, *40*(4), 514-527.

Hui, L., de Bruin, A. B. H., Donkers, J., & van Merriënboer, J. J. G. (2021a). Stimulating the intention to change learning strategies: The role of narratives. *International Journal of Education Research*, *107*, 101753.

Hui, L., de Bruin, A. B. H., Donkers, J., & van Merriënboer, J. J. G. (2021b). Does individual performance feedback increase the use of retrieval practice? *Educational Psychology Review.* Advance online publication.

Iwaki, N., Matsushima, H., & Kodaira, K. (2013). Hypercorrection of high confidence errors in lexical representations. *Perceptual and Motor Skills*, *117*(1), 219-235.

Jacobs, R. A., & Kruschke, J. K. (2010). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8–21. doi:10.1002/wcs.80

Jacoby, L. L., & Wahlheim, C. N. (2013). On the importance of looking back: The role of recursive remindings in recency judgments and cued recall. *Memory & Cognition*, *41*(5), 625-637.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1441-1451.

Jouhari, Z., Haghani, F., & Changiz, T. (2016). Assessment of medical students' learning and study strategies in self-regulated learning. *Journal of Advances in Medical Education & Professionalism*, *4*(2), 72–79.

Kang, S. H., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*(5), 998-1005.

Kang, S. H., McDermott, K. B., & Roediger III, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*(4-5), 528-558.

Kang, S. H., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*(1), 97-103.

Kang, S. H., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning?. *Journal of Educational Psychology*, *103*(1), 48-59.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: Deciding to practice retrieval during learning. *Journal of Experimental Psychology: General*, *138*(4), 469-486.

Karpicke, J. D. (2012). Retrieval-based learning: Active retrieval promotes meaningful learning. *Current Directions in Psychological Science*, *21*(3), 157-163.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review*, *27*(2), 317-326.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772-775.

Karpicke, J. D., Butler, A. C., & Roediger III, H. L. (2009). Metacognitive strategies in student learning: do students practise retrieval when they study on their own?. *Memory*, *17*(4), 471-479.

Karpicke, J. D., & Roediger III, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151-162.

Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*, 966–968. doi:10.1126/science.1152408

Keith, N., and Frese, M. (2008). Effectiveness of error management training: a meta-analysis. *Journal of Applied Psychology*, *93*, 59–69. doi: 10.1037/0021-9010.93.1.59

Kerfoot, B. P. (2009). Learning benefits of on-line spaced education persist for 2 years. *The Journal of Urology*, *181*(6), 2671-2673.

Kinnell, A., & Dennis, S. (2011). The list length effect in recognition memory: An analysis of potential confounds. *Memory & Cognition*, *39*(2), 348-363.

Knight, J. B., Ball, B. H., Brewer, G. A., DeWitt, M. R., & Marsh, R. L. (2012). Testing unsuccessfully: A specification of the underlying mechanisms supporting its influence on retention. *Journal of Memory and Language*, *66*(4), 731-746.

Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, *131*(2), 147-162.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*(2), 219-224.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory*, *16*(2), 125-136.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: A distribution-based bifurcation model. *Journal of Memory and Language*, *65*(2), 85–97. doi:10.1016/j.jml.2011.04.002

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(4), 989-998.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(1), 283-294.

Kornell, N., & Metcalfe, J. (2014). The effects of memory retrieval, errors and feedback on learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum* (p. 225–251). Society for the Teaching of Psychology.

Kornell, N., & Son, L. K. (2009). Learners' choices and beliefs about self-testing. *Memory*, *17*(5), 493-501.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effect of testing on skills learning. *Medical Education*, *43*(1), 21-27.

Kulasegaram, K., & Rangachari, P. K. (2018). Beyond "formative": Assessments to enrich student learning. *Advances in Physiology Education*, *42*(1), 5–14. doi:10.1152/advan.00122.2017

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary Educational Psychology*, *10*(3), 285-291.

Kulik, J. A., & Kulik, C. L. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58*(1), 79-97.

Kuo, T. M., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: Studies of the testing effect. *Journal of Memory and Language*, *36*(2), 188-201.

LaDisa, A. G., & Biesboer, A. (2017). Incorporation of practice testing to improve knowledge acquisition in a pharmacotherapy course. *Currents in Pharmacy Teaching and Learning*, *9*(4), 660–665. doi:10.1016/j.cptl.2017.03.002

LaPorte, R. E., & Voss, J. F. (1975). Retention of prose materials as a function of postacquisition testing. *Journal of Educational Psychology*, *67*(2), 259-266.

Larsen, D. P., Butler, A. C., & Roediger III, H. L. (2009). Repeated testing improves long-term retention relative to repeated study: A randomised controlled trial. *Medical Education*, *43*(12), 1174-1181.

Leahy, W., Hanham, J., & Sweller, J. (2015). High element interactivity information during problem solving may lead to failure to obtain the testing effect. *Educational Psychology Review*, *27*(2), 291-304.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1787-1794.

Lemley, D., Sudweeks, R., Howell, S., Laws, R. D., & Sawyer, O. (2007). The effects of immediate and delayed feedback on secondary distance learners. *Quarterly Review of Distance Education*, *8*(3), 251-260.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*(2), 94-97.

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior*, *8*(6), 828-835.

Mandernach, B. J. (2005). Relative effectiveness of computer-based and human feedback for enhancing student learning. *The Journal of Educators Online*, *2*(1), 1-17.

McCabe, J. (2011). Metacognitive awareness of learning strategies in undergraduates. Memory & Cognition, *39*(3), 462-476.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger III, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. Journal of Educational Psychology, 103(2), 399-414.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4-5), 494-513.

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention?. *Journal of Experimental Psychology: Applied*, *21*(4), 370-382.

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371-385.

McDermott, K. B., Agarwal, P. K., D'Antonio, L., Roediger III, H. L., & McDaniel, M. A. (2014). Both multiple-choice and short-answer quizzes enhance later exam performance in middle and high school classes. *Journal of Experimental Psychology: Applied*, *20*(1), 3-21.

McKinley, G. L., & Benjamin, A. S. (2020). The role of retrieval during study: Evidence of reminding from overt rehearsal. *Journal of Memory and Language*, *114*, 104128.

Melton, A. W. (1967). Repetition and retrieval from memory. *Science*, *158*(3800), 532-532.

Metcalfe, J. (2017). Learning from errors. *Annual Review of Psychology*, *68*, 465-489.

Metcalfe, J., & Finn, B. (2011). People's hypercorrection of high-confidence errors: Did they know it all along?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(2), 437-448.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*(2), 225-229.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*(8), 1077-1087.

Metcalfe, J., & Miele, D. B. (2014). Hypercorrection of high confidence errors: Prior testing both enhances delayed performance and blocks the return of the errors. *Journal of Applied Research in Memory and Cognition*, *3*(3), 189-197.

Meyer, A. N., & Logan, J. M. (2013). Taking the testing effect beyond the college freshman: Benefits for lifelong learning. *Psychology and Aging*, *28*(1), 142-147.

Middleton, E. L., & Schwartz, M. F. (2012). Errorless learning in cognitive rehabilitation: A critical review. *Neuropsychological Rehabilitation*, *22*(2), 138-168.

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional Science*, *32*(1-2), 99-113.

Murayama, K., Blake, A. B., Kerr, T., & Castel, A. D. (2016). When enough is not enough: Information overload and metacognitive decisions to stop studying information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(6), 914-924.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*(1), 18-22.

Ohrt, D. D., & Gronlund, S. D. (1999). List-length effect and continuous memory: Confounds and solutions.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, *144*(7), 710-756.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging*, *17*(2), 299-320.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 3-8.

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1287-1293.

Phillips, J. L., Heneka, N., Bhattarai, P., Fraser, C., & Shaw, T. (2019). Effectiveness of the spaced education pedagogy for clinicians' continuing professional development: A systematic review. *Medical Education*, *53*, 886-902.

Piza, F., Kesselheim, J. C., Perzhinsky, J., Drowos, J., Gillis, R., Moscovici, K., Danciu TE., Kosowska, A., & Gooding, H. (2019). Awareness and usage of evidence-based learning strategies among health professions students and faculty. *Medical Teacher*, *41*(12), 1411-1418.

Postman, L. (1965). Unlearning under conditions of successive interpolation. *Journal of Experimental Psychology*, *70*(3), 237-245.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?. *Journal of Memory and Language*, *60*(4), 437-447.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, *330*(6002), 335-335.

Pyc, M. A., & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test–restudy practice: Implications for student learning. *Applied Cognitive Psychology*, *25*(1), 87-95.

Rapp, E. J., Maximin, S., & Green, D. E. (2014). Practice corner: Retrieval practice makes perfect, *RadioGraphics*, *34*(7), 1869-1870.

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(2), 163-178.

Raupach, T., Andresen, J. C., Meyer, K., Strobel, L., Koziolek, M., Jung, W., Brown, J., & Anders, S. (2016). Test-enhanced learning of clinical reasoning: a crossover randomised trial. *Medical Education*, *50*(7), 711-720.

Rawson, K. A. (2015). The status of the testing effect for complex materials: Still a winner. *Educational Psychology Review*, *27*(2), 327-331.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough?. Journal of Experimental Psychology: General, 140(3), 283-302.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning?. *Journal of Experimental Psychology: Applied*, *15*(3), 243-257.

Richmond, A., Cranfield, T., & Cooper, N. (2019). Study tips for medical students. *BMJ*, *365*(April), 10–12.

Roediger III, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied*, *17*(4), 382-395.

Roediger III, H. L., & Blaxton, T. A. (1987). Effects of varying modality, surface features, and retention interval on priming in word-fragment completion. *Memory & Cognition*, *15*(5), 379-388.

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20-27.

Roediger III, H. L., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249-255.

Roediger III, H. L., & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*(3), 181-210.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(1), 233–239.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Rowland, C. A., & DeLosh, E. L. (2014). Benefits of testing for nontested information: retrieval-induced facilitation of episodically bound material. *Psychonomic Bulletin & Review*, *21*(6), 1516-1523.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & the A.T.P.R. Group (Eds.) *Parallel distributed processing: Exploration in the microstructure of cognition*, Vol. 2. Cambridge, MA: MIT Press, (pp. 216-271).

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, *11*(6), 641-650.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*(4), 207-218.

Schmidt, R. A., Young, D. E., Swinnen, S., & Shapiro, D. C. (1989). Summary knowledge of results for skill acquisition: Support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 352-359.

Schooler, L. J., & Anderson, J. R. (1990). The disruptive potential of immediate feedback. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*, 702-708, Cambridge, MA.

Shaw, T., Long, A., Chopra, S., & Kerfoot, B. P. (2011). Impact on clinical behavior of face-to-face continuing medical education blended with online spaced education: a randomized controlled trial. *Journal of Continuing Education in the Health Professions*, *31*(2), 103-108.

Siler, J., & Benjamin, A. S. (2019). Long-term inference and memory following retrieval practice. *Memory & Cognition*, 1-10.

Sitzman, D. M., Rhodes, M. G., Tauber, S. K., & Liceralde, V. R. T. (2015). The role of prior knowledge in error correction for younger and older adults. *Aging, Neuropsychology, and Cognition*, *22*(4), 502-516.

Smits, M. H., Boon, J., Sluijsmans, D. M., & Van Gog, T. (2008). Content and timing of feedback in a web-based learning environment: Effects on learning as a function of prior knowledge. *Interactive Learning Environments*, *16*(2), 183-193.

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus Performance: An Integrative Review. *Perspectives on Psychological Science*, 10(2):176-199.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, *3*(5), 315-317.

Stanger-Hall, K. F., Shockley, F. W., & Wilson, R. E. (2011). Teaching students how to study: a workshop on information processing and self-testing helps students learn. CBE—Life *Sciences Education*, *10*(2), 187-198.

Strong Jr, E. K. (1912). The effect of length of series upon recognition memory. *Psychological Review*, *19*(6), 447–462.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279-1285.

Timmer, M. C., Steendijk, P., Arend, S. M., & Versteeg, M. Making a Lecture Stick: the Effect of Spaced Instruction on Knowledge Retention in Medical Education, *Medical Science Educator*, *30*, 1211–1219.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: questions and answers. *Experimental Psychology*, *56*(4), 252-257.

Tullis, J. G., Benjamin, A. S., & Ross, B. H. (2014). The reminding effect: Presentation of associates enhances memory for related words in a list. Journal of Experimental Psychology: General, 143(4), 1-15.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: Guiding learners to predict the benefits of retrieval. *Memory & Cognition*, *41*(3), 429-442.

van Gog, T., & Kester, L. (2012). A test of the testing effect: acquiring problem-solving skills from worked examples. *Cognitive Science*, *36*(8), 1532-1541.

van Gog, T., Kester, L., Dirkx, K., Hoogerheide, V., Boerboom, J., & Verkoeijen, P. P. (2015). Testing after worked example study does not enhance delayed problem-solving performance compared to restudy. *Educational Psychology Review*, *27*(2), 265-289.

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review*, *27*(2), 247-264.

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval?. *Psychological Science*, *22*(9), 1127-1131.

Versteeg, M., Hendriks, R. A., Thomas, A., Ommering, B. W. C., & Steendijk, P. (2019). Conceptualising spaced learning in health professions education: A scoping review. *Medical Education*, (June), 1–12. https://doi.org/10.1111/medu.14025

Ward, G. (2002). A recency-based account of the list length effect in free recall. *Memory & Cognition*, *30*(6), 885-892.

West, C. P., Huschka, M. M., Novotny, P. J., Sloan, J. A., Kolars, J. C., Habermann, T. M., & Shanafelt, T. D. (2006). Association of perceived medical errors with resident distress and empathy: a prospective longitudinal study. *JAMA*, *296*(9), 1071–1078.

West, C., & Sadoski, M. (2011). Do study strategies predict academic performance in medical school? *Medical Education*, *45*(7), 696–703. doi:10.1111/j.1365-2923.2011.03929.x

Wheeler, M., Ewers, M., & Buonanno, J. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*(6), 571-580.

Wheeler, M. A., & Roediger III, H. L. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science*, *3*(4), 240-246.

Whyte, M. M., Karolick, D. M., Nielsen, M. C., Elder, G. D., & Hawley, W. T. (1995). Cognitive styles and feedback in computer-assisted instruction. *Journal of Educational Computing Research*, *12*(2), 195-203.

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*(1), 10-16.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*(3), 214-221.

Yan, V. X., & Sana, F. (2021). Does the interleaving effect extend to unrelated concepts? Learners' beliefs versus empirical evidence. *Journal of Educational Psychology*, *113*(1), 125-137.

Yan, V. X., Soderstrom, N. C., Seneviratna, G. S., Bjork, E. L., & Bjork, R. A. (2017). How should exemplars be sequenced in inductive learning? Empirical evidence versus learners' opinions. *Journal of Experimental Psychology: Applied*, *23*(4), 403.

Yan, V. X., Thai, K. P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: do they vary with mindset?. *Journal of Applied Research in Memory and Cognition*, *3*(3), 140-152.

Yang C, Luo L, Vadillo MA, Yu R, Shanks DR. Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. Psychological Bulletin. 2021 Mar. DOI: 10.1037/bul0000309.

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology*, *111*(1), 73-90.

Zimmerman, J. (1975). Free recall after self-paced study: A test of the attention explanation of the spacing effect. The American Journal of Psychology, *88*(2), 277-291.

Chapter 5: Goals and Consequences Motivate

Key Points

- Expectancy-value theory is a valuable framework for exploring what factors can influence an individual's motivation to learn from assessments.
- The expectation of being assessed can motivate individuals to meet targeted goals.
- The perceived benefits and costs of testing shape individuals' goal-setting behaviors and motivation for learning.
- Framing longitudinal assessment as a learning intervention may create intrinsic motivation within physicians.
- Assessment organization should endeavor to minimize anxiety that may surround a testing environment.

Overview

In order for physicians to be motivated to participate in self-regulated learning and longitudinal assessments they must see the value of participating and understand the consequences of not participating. One general approach that can help us understand the role of consequences in learning is the framework of cost-benefit analysis (Kurzban, Duckworth, Kable & Myers, 2013). This perspective suggests that people's willingness to perform a mental task, such as studying, declines with the perceived cost of foregoing other alternative activities (e.g., spending time with family, watching TV) but increases with the perceived benefits of the task (e.g., keeping current on knowledge to make better judgments on patient care). This general approach aligns well with the expectancy-value theory from social and educational psychology (Eccles & Wigfield, 2002, 2020; Wigfield & Eccles, 2000; Wigfield, Tonks, & Klauda, 2016). This theory is broadly used in the literature to explain, understand, and predict human motivation in learning and academic performance. Expectancy-value theory posits that for learners, pursuing an educational goal (i.e., their motivation to learn) is a function of the perceived benefits of pursuing the goal, the perceived costs of pursuing it, and the chance of succeeding if they do pursue the goal (i.e., expectancies).

Motivation to Learn = Expectancies * (Benefits - Cost)

Thus, all other things being equal, physicians--and other learners--should be more motivated to study and practice their skills when there is a clear benefit for doing so, when there is a clear cost to not doing so, and when there is a reasonable expectation of success. In this chapter, we describe each of these expectancy-value theory components in turn and discuss the implications and applications of those components for continuing certification programs in medicine.

In *Figure 5-1*, we present a model in which we bring together each of these three motivational factors, their hypothesized interrelations, and the motivation to learn. We begin by reviewing research on the effects of one's expectancy for passing the assessment and of self-efficacy beliefs on engagement and learning outcomes. We then review the perceived benefits of testing and explore the hypothesis that physicians will experience stronger motivation and learning to the degree to which the assessment

aligns with and confers value to them. We then consider related research on mindsets and achievement goals and how they can affect factors related to learning and performance outcomes. Next, we discuss the perceived costs of testing, such as the perception of external rewards, anxiety, and stereotype threat (i.e., a situation in which one is concerned about potentially confirming a stereotype related to an aspect of their identity), as well as approaches to mitigate those costs. An important point to note is that much of the research that we found has focused on testing as a particular point-in-time assessment and not as a repeated, longitudinal view of assessment that could take place over years (c.f., Bernacki, Aleven, & Nokes-Malach, 2013). We end with a discussion of directions for future work in the area of motivation and the development of medical expertise.

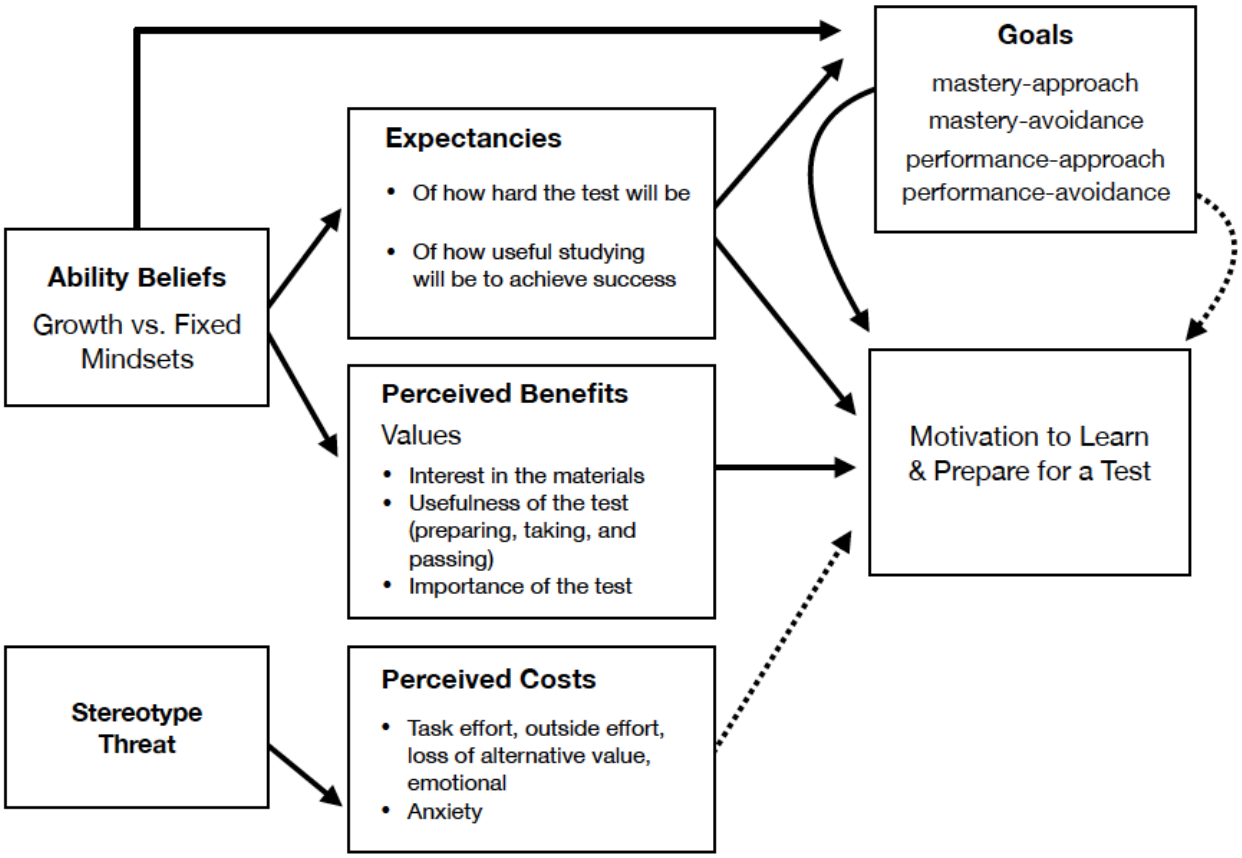


Figure 5-1. Expectancy-value model of the motivation to learn and interrelations to mindsets, stereotype threat, and achievement goals. The model is adapted from Eccles, Wigfield, and colleagues (Eccles & Wigfield, 2020; Wigfield & Eccles, 2000). Positive relations are denoted by the solid arrows and negative relations are denoted by dashed arrows.

Expectancies

Expectations of test difficulty affect engagement and performance

In the chapter on *Testing Enhances Learning and Retention*, we reviewed how the act of being tested can improve future retention. In addition, the *expectation* that one will be tested in the future can create the opportunity to engage in productive study and learning activities. Although just knowing about the existence of an upcoming assessment does not necessarily lead to better learning (Hyde & Jenkins, 1973, EL: 4; Postman, 1964, EL: 3), having a specific expectation of being tested in a particular way or on certain types of material can lead to better learning (McDaniel, Blischak, & Challis, 1994, EL: 3; Szpunar, McDermott, & Roediger, 2007, EL: 4).

The anticipated difficulty of the test matters, too. For example, laboratory researchers have often examined test difficulty by contrasting a *recall* task, in which learners must bring to mind the required information (e.g., a fill-in-the-blank item or essay), with a *recognition* task, in which learners must merely identify the information when it is presented (e.g., multiple-choice or true-false items). All other things being equal, recall is more difficult than recognition. Thus, people expecting a difficult recall test learn and remember more than people expecting an easier recognition test, regardless of the type of test they actually receive (Balota & Neely, 1980, EL: 4; Connor, 1977, EL: 3; d'Ydewalle, Swerts, & de Corte, 1983, EL: 4; Hall, Grossman, Elwood, 1976, EL: 3; Leonard & Whitten, 1983, EL: 3; Maisto, DeWaard, & Miller, 1977, EL: 4; Neely & Balota, 1981, EL: 4; Schmidt, 1988, EL: 4; c.f., Finley & Benjamin, 2012, EL: 3).

What accounts for this *test expectancy effect*? Benefits of intentional encoding (i.e., with the goal to learn) appear to be driven largely by the activities that learners engage in when preparing for a test (Hyde & Jenkins, 1969, EL: 4; Hyde & Jenkins, 1973, EL: 4; c.f., Neely & Balota, 1981, EL: 4). For example, deeper processing of the to-be-learned material such as focusing on the meaning of the words are better remembered than more surface processing activities such as focusing on the syntactic properties (Hyde & Jenkins, 1973, EL: 4). Learners expecting a more difficult test engage in more effective study behaviors, such as studying longer (d'Ydewalle et al., 1983, EL: 4; Thiede, 1996, EL: 3), continuing to practice after an initial quiz (Szpunar et al., 2007, EL: 4), and/or engaging in deeper, more meaningful practice (Hall et al., 1976, EL: 3; Leonard & Whitten, 1983, EL: 3; Schmidt, 1988, EL: 4). Conversely, even offering financial incentives does not increase learning when learners are required to use ineffective learning strategies (Craik & Tulving, 1975, EL: 4).

Together, these results suggest several principles about how the expectation of a test can influence physicians' motivation to learn. First, physicians may learn and retain more when they expect to be tested on those skills and when there is a consequence for not succeeding on the test. Second, the perceived difficulty of the test should be difficult enough to engender deeper, more effective learning when preparing for it. Finally, if the learning environment provides guidance on effective study behaviors and activities along with feedback, physicians will better be able to capitalize on the increased motivation to learn and perform well on a test.

Expectations of success affect how one studies and performs

In addition to the potential benefits of preparing for a specific type of test, research from the broader motivation literature suggests that a learner's beliefs about the likelihood of success on a given task has important consequences for their learning activities. Expectancy beliefs are theorized to be informed by both beliefs related to task outcome success (i.e., outcome expectancies) as well as beliefs

about one's personal capabilities to perform the task (i.e., self-efficacy). In the expectancy-value model, learners' beliefs of the likelihood of success on a given task have implications for their motivation to learn. For example, if a learner is given a task they perceive themselves as unlikely to succeed on (e.g., an extremely difficult test or a test in which not enough time is given to complete), then that individual will be less likely to engage in that activity or to prepare adequately because they are expecting that they will likely fail anyway. Although high failure rates are not traditionally a problem for continuing education programs, what constitutes subjective perceptions of success may be defined differently for different physicians (e.g., some may set high internal standards for success whereas others set lower standards). Prior work has shown that these expectancy beliefs have a large impact on academic performance (Meece, Wigfield, & Eccles, 1990, EL: 5; Penk & Schipolowski, 2015, EL: 5; Priess-Groben & Hyde, 2017, EL: 5; Wigfield & Eccles, 2000, EL: 2), persistence (Scheier & Carver, 1982, EL: 4) and choice (Bong, 2001, EL: 5; Durik, Vida, & Eccles, 2006, EL: 5; Simpkins, Davis-Kean, & Eccles, 2006, EL: 5).

Learners' beliefs about their capability to be successful on a particular task or in a specific domain have been described as one's *self-efficacy* beliefs (Bandura, 1986 1997). Albert Bandura and his colleagues have done much foundational work to understand these beliefs and their role in learning and achievement outcomes. High self-efficacy is associated with more productive learning behaviors (Bouffard-Bouchard, Parent, & Larvilee, 1991, EL: 5; Parajes, 2008, EL: 2; Pintrich & De Groot, 1990, EL: 5; Schunk & Parajes, 2002, EL: 2). For example, students who have high self-efficacy beliefs are more likely to engage in self-regulated learning and show persistence to learn new materials even in the face of difficulties or challenges (Bandura, 1997, EL: 2; Schunk & Parajes, 2002, EL: 2). These beliefs have also been shown to predict student retention and academic performance in school settings (Honicke & Broadbent, 2016, EL: 1) even after controlling for prior knowledge (Bailey, Lombardi, Cordova, & Sinatra, 2017, EL: 5; Kalender, Marshman, Schunn, Nokes-Malach, & Singh 2020, EL: 5). A number of factors have been hypothesized to influence one's development of self-efficacy, including performance feedback (e.g., test scores), observations of others, social persuasion messages, and physiological states (Bandura, 1997, EL: 2; Britner, 2008, EL: 2; Britner & Parajes, 2006, EL: 2; Usher & Pajares, 2008, EL: 2).

The prior work on anticipated test difficulty and beliefs about success on the test has several implications for the consideration of one's motivation to learn from a test. First, the assessment has to strike a balance of difficulty in that it is perceived as being challenging enough to confer the benefits of motivating one to engage in constructive study activities, but at the same time so it is not perceived as too difficult in that there would be no possibility for success. One way to communicate the level of difficulty is to provide representative examples of the test items to practice and receive feedback on. One factor that is likely to impact practicing physicians' performance and perception of difficulty is the fit between the topics of the questions and the kinds of patients they routinely see. This may mean that providing physicians with some control over the scope of the test content may help obtain an appropriate level of item difficulty, challenging but not too challenging. Further, as we will discuss in the value section below, relevance of the content to the physician is predicted to have an impact on motivation.

Second, because self-efficacy beliefs have a strong impact on how learners prepare and engage with the study materials, continuing certification programs have an opportunity to contribute to the positive development of self-efficacy. That is, the results of the assessment provide a form of performance feedback that could directly impact a physician's self-efficacy belief (e.g., getting a higher

score and thereby increasing self-efficacy). If continuing certification programs transition to more regularly spaced assessments, it is a further opportunity to develop self-efficacy by providing multiple instances of feedback over time. Each instance of feedback is an opportunity for an individual to adjust their appraisal of self-efficacy to be more in line with their performance (i.e., to improve by increasing study in areas of weakness so as to reduce the discrepancy). Of course, this could lead to a relative or momentary decrease in self-efficacy if a physician performs poorly on the assessment. To potentially mitigate negative aspects of such effects, test performance scores could be interleaved with encouraging messages to further promote self-efficacy (e.g., *you performed exceptionally well on X, but could improve some on Y*). Providing repeated performance feedback creates an interesting opportunity to see whether such feedback leads to more accurate self-assessment. It is also possible that someone could perform poorly repeatedly in longitudinal assessment leading to longer lasting negative self-efficacy. In these cases, assessment designers may consider flagging negative trajectories and consider interventions to help better prepare for the next longitudinal assessment.

Recommendation 5-1: It is important to develop an assessment that is challenging enough to confer motivational benefits, but not so difficult that it is perceived as likely to result in failure.

Perceived Benefits/Value: What is the value of the test to the learner?

In the expectancy-value model, the perceived value of a given task or assessment plays a critical role in one's motivation to prepare for and engage with that task or assessment. From this perspective, *values* are viewed as subjective qualities that an individual ascribes to a learning task or assessment (Wigfield, et al., 2016). These values are hypothesized to consist of multiple distinct components that include intrinsic task value (interest in the content of the test), utility value (usefulness of the test), and attainment value (importance of the test).

Intrinsic Task Value

Intrinsic task value is one's interest in the task or assessment for its own sake. Theories of interest typically discuss two different kinds: situational and individual (Krapp, 1999; Hidi & Harackiewicz, 2000; Hidi & Renninger, 2006; Schiefele, 1991). *Situational interest* is described as a momentary experience that is driven by environmental factors (e.g., a bright or loud stimulus) and correlated with both cognitive (e.g., attentional focus) and affective factors (e.g., positive or negative feelings) (Hidi & Harackiewicz, 2000, EL: 2). *Individual interest* is hypothesized to be a longer-lasting engagement and is associated with one's knowledge, values, and feelings for the particular target topic or task (Renninger, 2000, EL: 2). Much prior work has shown that individual interest in the task can increase self-reported effort (Renninger & Hidi, 2002, EL: 5), positive self-regulation (O'Keefe & Linenbrink-Garcia, 2014, EL: 5; Renninger & Hidi, 2019, EL: 2), and deep strategy use (Schiefele, Wild, & Krapp, 1995, EL: 5). It is also associated with better grades in school (Harackiewicz et al., 2008, EL: 5; Schiefele, Krapp, & Winteler, 1992, EL: 1).

In discrepancy theory, learners are intrinsically motivated to increase a valued competency when they learn of a *discrepancy* between their ability and a given goal or standard (Fox & Miner, 1999). Discrepancy theory posits that awareness of the gap between *what is* and *what ought to be* can cause feelings of discomfort or dissatisfaction within individuals that motivates them to alleviate this feeling by reducing a perceived gap in knowledge or ability (see also regulatory focus theory, Higgins, 1997, 2012). Similarly, research on metacognitive monitoring has shown that people are sensitive to gaps between perceived and desired knowledge (Dunlosky & Hertzog, 1998, EL: 5; Son & Metcalfe, 2000, EL: 3; Benjamin & Tullis, 2010, EL: 2). Critically, it is a learner's *perception* of a discrepancy that leads to high levels of motivation. In this view of learning, feedback would be especially important as it allows for learners to become aware of (i.e., perceive) "discrepancies."

In addition to work investigating natural variation in intrinsic task value and its positive relations to study and test outcomes, other research has shown that experimental interventions can facilitate interest and subsequent learning. For example, testing with feedback, in addition to directly enhancing learning, can also increase the desire to learn more about a topic (Abel & Bäuml, 2020: EL 3). Personalizing content can also increase interest and performance for a task (Bernacki & Walkington, 2018, EL: 4; Walkington & Bernacki, 2017, EL: 2). One way to personalize content is to create links to topical interests. Creating links can be done in a variety of ways. One is to have the learner reflect on the relevance of each item to their own interests. Another is to survey the learner's topics of interest and then, based on the survey responses, assign learning or test activities that match those topics (with the idea the test creator has constructed an assessment with multiple topic instantiations). One example of this latter type of intervention showed that students were more likely to engage with and learn math content when the word problems used topics that they had personal interests in (e.g., sports or music) (Bernacki & Walkington, 2018, EL: 4).

One implication from the research on intrinsic interest is that the more the content of the test (e.g., topics and patient scenarios) can match the interests of the physician the more motivated they will be to learn and keep current on that information. Further, the more they find the content interesting, the more engaged they will be with the material. This suggests that it would be desirable to collect information on what physicians interests are, in order to be able to match those interests. An additional consideration is that personalizing the material to a greater degree--matching a set of medical interests of the physician--may make physicians more motivated and engaged with preparing for and taking the assessment. Indeed, individual's interests and perceived competence have been shown to predict important career choices for medical students like medical specialty decisions (Williams, Saizow, Ross, & Deci, 1997, EL: 5). An implication of this work is that if physicians could select topic areas or problem scenarios to be tested on that matched their medical interests or personal practice, their motivation in preparing for and engaging with those questions should increase. An important caveat here would be to balance personalization and interest matching with appropriate coverage of material so that individuals do not only self-select topics they expect to do well in. Another technique that could potentially help select material in a longitudinal, spaced-repetition paradigm might include the collection of ratings of relevance, which could then be used to prioritize content of the information to be re-presented.

Recommendation 5-2: Longitudinal assessment programs should instill motivational benefits to

physicians by including topics that are especially important to the learner while balancing this with appropriate coverage of the content in the discipline.

Utility Value

Another aspect of value is called *utility value*, or the degree to which preparing for and taking the test is useful for some valued outcome; that is, as a means to an end (Eccles, 2009; Wigfield & Eccles, 2002). Utility value is often discussed in terms of the relevance of the task or test in relation to one's broader personal, educational, or professional goals. Is the task or test instrumental to helping one accomplish goals? Basic research on utility value has shown that it is positively associated with engagement, with learning, and with positive performance outcomes, such as higher grades (Harackiewicz, Smith, Priniski, 2016, EL: 2; Harackiewicz, Tibbets, Canning, & Hyde, 2014, EL: 2).

Several studies investigating utility value interventions have shown that, when utility value increases, so do academic performance outcomes (Harackiewicz & Priniski, 2018, EL: 2; Harackiewicz, Canning, Tibbetts, Priniski, & Hyde, 2016, EL: 4, Hulleman, Godes, Hendricks, Harackiewicz, 2010, EL: 4). For example, one intervention had students engage in a writing activity that encouraged them to make connections between their lives and what they were learning in their science course (Hulleman & Harackiewicz, 2009, EL: 4). A comparison group was instructed to write summaries of the material they were learning about. For students who entered the class with low initial performance expectations, the utility-value intervention increased their intrinsic interest and higher grades at the end of the course relative to writing summaries.

We focus on two types of utility value for the continuing medical education context. The first concerns the usefulness of preparing to take the assessment. That is, how does a physician value the act of preparing for the assessment? Do they view it as a helpful activity that is contributing to their medical training and skill development more generally, or just something they do because they have to? The more a physician sees connections between the activities of studying and their broader professional goals, the more motivated they will be to study. The second aspect concerns the value ascribed to the assessment itself. That is, does a physician view the test as useful to their broader educational goals (e.g., measuring their expertise, staying current) and professional goals (e.g., staying employed, being promoted)?

Regardless of the type of utility value, using relevant feedback will be important. Some longitudinal assessment platforms require the participant to rate each question regarding how relevant it was to their medical practice before letting them know if they answered it correctly or incorrectly. Periodically providing reminder-feedback to the participant regarding which questions they missed that they also rated as relevant to their practice may provide additional motivation for them to review those concepts. The periodic nature of the reminders should be congruent with the principles of spaced-repetition discussed in the earlier chapter. This feedback could be presented as summary feedback between assessment sessions.

An implication of this research is that physicians' motivation to learn is affected by their perception of the usefulness of the test for their career. If physicians see the program as relevant to their broader professional goals (e.g., further developing their cognitive skills and keeping current on

knowledge) they will be more motivated and more deeply engaged with the material. Likewise, physicians will be motivated to perform better if they view the assessment as useful to their professional goals (e.g., keeping one's certification). Interviews with physicians preparing and taking high stakes assessments show a range of perceptions of how relevant and related the content is to their current practice (Chesluk, Eden, Hansen, Johnson, Reddy, Bernabeo, & Gray, 2019, EL: 5). Some evidence suggests that utility value is amenable to intervention; for instance, in academic settings, it can be improved by having the learner briefly write about the usefulness of the assessment to them.

Recommendation 5-3: Basic research suggests that how testing organizations frame the assessment to physicians can impact how they perceive its usefulness (e.g., as an opportunity to develop relevant skills rather than as a required assessment).

Attainment Value

The third component of value is called *attainment value*, which is the benefit of doing well on a given task or assessment. In the current context, attainment value would capture how important it is to the individual to prepare for the assessment and perform well on it. This judgment will depend on the physician's perception of what the assessment is measuring (e.g., relevant medical knowledge and skills), how accurately it is measuring those competences, and the ramifications of passing or failing the assessment (e.g., often required for preferred employment).

Attainment value is theorized to have implications for one's self-concept and identity (Eccles, 2009; Eccles & Wigfield, 2020; Ryan & Deci, 2020). For example, self-determination theory implies that outcomes can confirm or deny aspects of one's identity (Ryan & Deci, 2020; La Guardia, 2009), including three core needs of autonomy, relatedness, and--most critical to our purposes--competence. If one perceives the assessment as measuring critical medical competence (i.e., places high attainment value on the assessment) and performs well, that result can be interpreted as confirming one's view of oneself as a competent, expert physician. Alternatively, if one perceives the assessment as important but performs poorly, it could call into question one's view of oneself as an expert or knowledgeable physician, or the validity / accuracy of the assessment.

Attainment value has been shown to be positively related to engagement (Putwain, Nicholson, Pekrun, Becker, & Symes, 2019, EL: 5), effort (Guo, Nagengast, Marsh, Kelava, Gaspard, Brandt, Cambria, Flunger, Dicke, Hafner, Brisson, & Trautwein, 2016, EL: 5), self-concept (Arens, Schmidt, & Preckel, 2019, EL: 5), and academic achievement (Trautwein et al., 2012, EL: 5; Meyer, Fleckenstein, & Koller, 2019; EL: 5). Therefore, whether or not one gets attainment value from the assessment has important implications for motivation and performance.

Laboratory research on memory and learning has also supported the relevance of attainment value by showing that the importance of the to-be-learned information has consequences for learning outcomes. In one lab paradigm, each to-be-learned item is experimentally assigned a point value that learners are awarded for successful retention, and learners are tasked with earning as many points as possible. Learners consistently remember more of the high-value items, demonstrating that value motivates better learning and retention (Castel, Benjamin, Craik, & Watkins, 2002, EL: 3; Castel,

Humphreys, Lee, Galván, Balota, & McCabe, 2011, EL: 4; Castel, Murayama, Friedman, McGillivray, & Link, 2013, EL: 3; Hennessee, Knowlton, & Castel, 2018, EL: 3; McGillivray & Castel, 2017, EL: 3). In part, this difference comes about because learners elect to study more of the high-value material and study it for longer amounts of time (Castel et al., 2013, EL: 3).

The implication of this work is that physicians' perception of the importance of the task and test affects their motivation to learn. The more that physicians see the test as measuring an important set of skills and knowledge, the more they will invest in performing well on the test. Further, if the assessment provides feedback relevant to physicians' self-concepts and identities (e.g., the identity of a skilled medical physician), they will show higher investment in developing and performing well on the assessment. Lastly, longitudinal assessments of medical expertise could encourage physicians to learn and retain particular skills by assigning them higher value, or apportioning more questions to these topics within the assessment (as is often already done).

Benefits of Pursuing Mastery and Achievement Goals

Achievement goals are the reasons why people engage in study and test activities. Achievement goals sometimes are described and investigated separately from expectancy-value theory and sometimes are included in an overarching model (Plante, O'Keefe, & Theoret, 2013). Some models place achievement goals as a factor that affects expectancies and values (Wigfield & Eccles, 2000) whereas others have conversely hypothesized that expectancies and values affect the adoption of achievement goals (Elliot, 1999; Greene, Miller, Crowson, Duke, & Akey, 2004). Some prior empirical work supports this second view by showing that expectancies and values have both direct effects on motivation for learning as well as indirect effects through achievement goals (Plante, O'Keefe, & Theoret, 2013, EL: 5). We highlight these relations in our model in figure 5-1.

Achievement goals can either be *mastery-oriented*, with a focus on improving and understanding the material in comparison to one's prior understanding, or *performance-oriented*, with a focus on demonstrating ability in comparison to others (Dweck, 1986; Elliot, 1999). Each of these two goals can be approach-or avoidance-based (Elliot, 1999). *Approach-based* goals are defined by striving toward a positive outcome and *avoidance-based* goals are defined by avoiding negative outcomes. Combining these different dimensions results in four different goals: a *mastery-approach* goal to learn as much as possible, a *mastery-avoidance* goal to avoid loss of knowledge or skills, a *performance-approach* goal to perform better than others, and a *performance-avoidance* goal not to perform worse than others (see Table 5-2 for a summary) (Elliot & McGregor, 2001; Elliot & Murayama, 2008).

Table 5-2. The 2 x 2 achievement goal framework Elliot and McGregor (2001) and Elliot and Murayama (2008).

	Mastery	Performance
Approach	Focus on self-improvement of skills understanding	Focus on performing better than others
Avoidance	Focus on not losing skills and understanding	Focus on not performing worse than others

Much prior work has examined these four achievement goals in relation to engagement, learning, and performance outcomes in both laboratory experiments and classroom studies. This literature has found that some types of achievement goals yield superior learning and performance than others. Specifically, performance-avoidance goals have been consistently related to negative outcomes, such as poor performance (e.g., grades and tests), as well as low self-efficacy, poor study habits, and procrastination (Elliot & Church, 1997, EL: 5; Elliot & McGregor, 1999, EL: 5; Elliot, McGregor, & Gable, 1999, EL: 5). Performance-approach goals represent a more intermediate level of performance; they have been related to some positive outcomes, such as better grades and assessment performance (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002, EL: 2; Linnenbrink-Garcia, Tyson, & Patall, 2008, EL: 2), but also some negative outcomes, such as less effective self-reported study behaviors (i.e., rote memorization) (Midgley, Arunkumar, & Urdan, 1996, EL: 5; Senko, Hulleman, & Harackiewicz, 2011, EL: 2).

In contrast, mastery-approach goals have been associated with positive outcomes such as self-reported interest and engagement (Elliott & Dweck, 1988, EL: 4; Elliot, McGregor, & Gable, 1999, EL: 5; Harackiewicz, Barron, Tauer, & Elliot, 2002, EL: 5; Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008, EL: 5). They have also shown positive relations to learning and transfer (Belenky & Nokes-Malach, 2012, 2013, EL: 5). For example, in one study, mastery-approach goals were the only achievement goals that predicted students learning a new statistics concept and then applying that concept to solve a new problem. However, there are also some mixed results regarding the relation between mastery-approach goals and student grades; some studies show positive relations and others show no relation (Belenky & Nokes-Malach, 2012, EL 5).

Mastery-avoidance goals have been the least studied of the four goal types but may be particularly relevant to certification boards as they pertain to avoiding the loss of knowledge and skills that were previously mastered. These goals have been associated with mixed results (Hulleman et al., 2010, EL: 4; Linbrink et al., 2008, EL: 2), including both positive outcomes, such as learning (Richey & Nokes-Malach, 2013, EL: 3), as well as negative outcomes, such as self-reported test anxiety (Elliot & McGregor, 2001, EL: 5). The effects of mastery-avoidance goals may critically depend on the learner's prior experience in the domain. For example, in one study that showed college students' mastery-

avoidance goals predicted learning outcomes, the target learning materials were middle-school physics concepts (e.g., electric circuits) that the students were likely to have already learned previously. One might hypothesize that this goal predicted learning because the students did not want to show that they did not know or understand the content that they had learned previously.

Other work has shown that achievement goals can be influenced by a number of factors (Ames, 1992, EL: 2) including that instructions can affect the goals that learners adopt in the moment (Elliot & Harackiewiz, 1996, EL: 3; Elliot & Dweck, 1988, EL: 4; Graham & Golan, 1991, EL: 3). For example, telling students that their goal for a given task is to "develop their understanding" or conversely "to perform better than others" impacts learning outcomes and engagement with the task (Bereby-Meyer & Kaplan, 2005, EL: 3; Elliot & Harackiewicz, 1996, EL: 3). This highlights how framing of the activity or task may influence what goals learners adopt in the moment. Other work has shown that the type of task can also impact the types of goals adopted. For example, a *discovery task*, in which the learner aims to find a principle that explains a data pattern, has been shown to promote the adoption of mastery-approach goals relative to a task presented as direct instruction followed by practice (Belenky & Nokes-Malach, 2012, EL: 4). The framing of the task is particularly relevant for the continuing certification of medical expertise because the instructions could easily be written to facilitate the adoption of a mastery goal. For example, physicians could be asked to focus on developing their understanding and trying to improve their score over time--aiming to achieve their personal best.

Growth Mindsets Benefits Motivation and Learning

Another important motivational factor that can impact how learners prepare for and engage with assessments is their mindset and beliefs about ability. *Mindset* is a broad term (Dweck & Yeager, 2019, EL: 2) used to describe a set of beliefs that can then impact one's expectations, and behaviors. Carol Dweck and her colleagues have been some of the leading researchers of mindsets and their impact on behavior, motivation, and cognition. One of the most powerful mindsets that have been investigated is people's beliefs about intelligence. In this work, Dweck and colleagues have focused on two types of beliefs. The first is a belief that intelligence is malleable and can change with experience in a domain, which has been called a *growth mindset*. The other is a belief that intelligence is inherited and cannot be changed through experience and has been called a *fixed mindset*. Growth mindsets and ability beliefs have been hypothesized to affect a learner's expectancies, values, and achievement goals, which in turn are hypothesized to affect the motivation to learn (see *Figure 5-1*). For example, a growth mindset is hypothesized to lead to positive self-regulated learning behaviors, such as effort in the context of challenge, which in turn lead to better learning and achievement outcomes (Blackwell, Trzesniewski, & Dweck, 2007, EL: 4&5). Research on mindsets has shown that a growth mindset predicts positive academic achievement (Costa & Faria, 2018, EL: 1; Blackwell, Trzesniewski, & Dweck, 2007, EL: 4&5; Gunderson, Gripshover, Romero, Dweck, Goldin-Meadow, & Levine, 2013, EL: 5, Henderson & Dweck, 1990, EL: 2, Paunesku, Walton, Romero, Smith, Yeager, Dweck, 2015, EL: 4; cf. Li & Bates, 2019, EL: 4). Growth and fixed mindsets have been shown to be related to students' self-reported interest (Haimovitz, Wormington, & Corpus, 2011, EL: 5), effort (Blackwell, Trzesniewski, & Dweck, 2007, EL: 5; Miele, Finn, & Molden, 2011, EL: 5; Miele & Molden, 2010, EL: 3), and learning goals (Blackwell, Trzesniewski, & Dweck, 2007, EL: 5; Haimovitz, Wormington, & Corpus, 2011, EL: 5). For example, at a

correlational level, a growth mindset during the middle-school years predicts learning goals (e.g., "An important reason why I do my school work is because I like to learn new things") and positive effort beliefs (e.g., "The harder you work at something, the better you will be at it"), which in turn predict positive study strategies (e.g., "I would spend more time studying for tests") and performance (e.g., achievement test scores) (Blackwell, Trzesniewski, & Dweck, 2007, EL: 5). The link between growth mindset and academic achievement is likely causal. Interventions designed to promote growth mindsets lead to positive changes in motivational and achievement outcomes (Blackwell, Trzesniewski, & Dweck, 2007, Expt. 2, EL: 4; Mueller & Dweck, 1998, EL: 4; Yeager, Romero, et al., 2016, EL: 4). In a recent national study with 12,490 ninth graders, researchers found that a brief one hour mindfulness intervention had a significant effect on GPA in mathematics and science courses particularly for lower-achieving students. These interventions emphasize messages that portray intelligence as malleable with experience and training.

In sum, mindsets about intelligence can have powerful downstream effects on motivational and learning outcomes and can directly impact expectancies, values, and goals. Therefore, physicians who believe their intelligence and skills are malleable are more likely to adopt good learning behaviors and goals, which would further their retention of cognitive skills. Physicians' adoption of a growth mindset may be fostered by the shift in continuing certification programs towards more regular spaced testing, which provide the opportunity to improve over time.

External Incentives Can Undermine Learning

The topics so far suggest an optimistic picture of how consequences can influence motivation. However, consequences can sometimes have deleterious effects. Theories of motivation (e.g., Bandura, 1986; Bénabou & Tirole, 2003, EL: 3; Ryan & Deci, 2000, EL: 3; Ryan & Deci, 2020, EL: 3) typically distinguish between *intrinsic motivation* (wanting to do something for its own sake) and *extrinsic motivation* (wanting to do something for other reasons). Contemporary theories of motivation (e.g., Howard, Gagné, Bureau, 2017, EL: 1; Ryan & Deci, 2020, EL: 3) further distinguish a continuum of extrinsic motivations ranging from a purely *external* focus on rewards and punishments to an *integrated* choice of activities that, while not necessarily enjoyable, are consistent with one's values (e.g., utility, attainment). For example, physicians might be motivated to keep their skills and knowledge current because they enjoy learning (intrinsic), because they value providing the best care to their patients (integrated), or because it would allow them to obtain or maintain a more prestigious employment position (external).

Broadly, intrinsic motivation is considered preferable, in the learning and maintenance of skills. Studies of individual differences have found that intrinsically-motivated students generally learn more and perform better, even controlling for prior knowledge (Taylor et al., 2014, EL: 1). In part, this is because motivation that can be classified as intrinsic or integrated leads people to work harder and persist longer in the face of difficulty (Hidi & Harackiewicz, 2000, EL: 2; Léon, Núñez, & Liew, 2015, EL: 5; Vallerand & Bissonnette, 1992, EL: 5), adopt better learning strategies (Léon et al., 2015, EL: 5), and procrastinate less (Katz, Eilot, & Nevo, 2014, EL: 5). By contrast, even when controlling for differences in intrinsic motivation, extrinsic motivation is associated with *poorer* academic performance (Lepper, Corpus, & Iyengar, 2005, EL: 5). The benefits of intrinsic motivation observed in these correlational

studies is echoed by randomized laboratory experiments, in which providing an extrinsic motivation (e.g., monetary rewards) tends to undermine performance on a variety of tasks, such as problem-solving, relative to manipulations that emphasize intrinsic motivation (e.g., discussing personal reasons to perform the task) or even *no* reward (Amabile, 1985, EL: 4; Glucksberg, 1962, EL: 4; Glucksberg, 1964, EL: 4; Gneezy & Rustichini, 2000, EL: 3; Kruglanski, Friedman, & Zeevi, 1971, EL: 4; McGraw & McCullers, 1979, EL: 4).

The potential concern, then, is that external consequences can in fact *undermine* this beneficial intrinsic and integrated motivation. Across a variety of tasks and participant groups, providing some type of incentive can shift learners to become extrinsically motivated such that, if the external reward is later removed, people report lower intrinsic motivation and are less apt to continue work (Bénabou & Tirole, 2003, EL: 3; Deci, 1971, EL: 3; Deci, Koestner, & Ryan, 1999, EL: 1; Deci, Koestner, & Ryan, 2001, EL: 3; Lepper & Greene, 1975, EL: 4; Lepper, Greene, & Nisbett, 1973, EL: 4; McGraw & McCullers, 1979, EL: 4; Pittman, Emery, & Boggiano, 1982, EL: 3; c.f., Cameron & Pierce, 1994, EL: 1; Eisenberger & Cameron, 1996, EL: 3). This phenomenon has been termed the *hidden cost of reward* (Deci, 1976). Thus, one concern could be that too heavily emphasizing external consequences for maintaining medical expertise could undermine physicians' intrinsic motivation to do so.

But, the literature suggests at least two strategies that may mitigate this possibility. First, because intrinsic/integrated motivation is better than a focus on purely external rewards, continuing certification programs could emphasize how maintenance of medical expertise may align with physicians' values (e.g., their best intentions in treating patients). Second, including some element of choice can yield a small-to-moderate increase in intrinsic motivation (Cohen's d = 0.25 for adults; Patall, Cooper, & Robinson, 2008, EL: 1)[3]. There are many different types of choices including: type of task, type of content, procedures for performing the task, relevant versus irrelevant aspects of the task. Self-determination theory predicts that choices are most effective when they support a sense of autonomy (i.e., they are meaningful and relevant) for the individual making that choice. Some have argued that this relevance should be balanced with the self-regulatory demands of the choice and therefore should simultaneously try to also minimize the cognitive demands of the choice (Patall et al., 2008, EL: 1). In order to boost physicians' motivation to participate in continuing certification programs it is important to demonstrate evidence of the efficacy of a longitudinal assessment program.

Perceived Costs of Testing

In the expectancy-value model, another important component of the motivation to learn is the perceived costs of the study activity or test. This component of the model has historically received less attention than components of expectancy and value). More recently, several efforts have been made to develop measurement tools that capture important aspects of cost (Conley, 2012; Flake, Barron, Hulleman, McCoach, & Welsh, 2015; Trautwein et al., 2012) and to better understand its role in the

[3] There is likely a trade-off in the amount of control over learning content provided to physicians. As discussed in the previous chapter, *Self-Assessment is Not Enough*, total control is inadvisable. Physicians' motivation may benefit from greater control over topics, but too much control is likely to result in poorer learning outcomes. We proposed a study to address this question at the end of *Self-Assessment is Not Enough*.

expectancy-value model (Barron & Hulleman, 2015; Eccles & Wigfield, 2020). Four aspects of cost have been identified: task effort, outside effort, loss of valued alternatives, and emotional cost (Flake et al., 2015). *Task effort* refers to the amount of time and energy of performing a task itself. *Outside effort* refers to the amount of time and energy required for other tasks than the focal task (e.g., family and work obligations, etc.), which may result in the perception of not having enough time to dedicate to the focal task. The *loss of valued alternatives* refers to what one has to give up to prepare for the task or test. For example, in the current context, a valued alternative lost in preparing for continuing certification program assessments may be leisure time (e.g., Galla, Plummer, White, Meketon, D'Mello, & Duckworth, 2014: EL 5). The last aspect is *emotional cost*, which refers to the potential stress and worry caused by the task. For example, anxiety in anticipation of a single high-stakes assessment would increase the perception of the emotional cost of the test. In interviews with physicians about how they prepared for and took continuing certification assessments the lack of time available because of outside effort involved in studying and the loss of valued alternatives were important themes that emerged (Chesluk et al., 2019, EL: 5).

In the expectancy-value model, the more perceived costs, the less likely one is to be motivated to learn and prepare for the assessment. In principle, then, the more that these perceived costs can be reduced, the better an individual's motivation to learn. Perhaps one way to mitigate the perceived costs would be to discuss those costs and expectations in advance in an effort to normalize them. Unfortunately, the research on perceived costs is still in its very early stages, so there are no clear intervention recommendations on how to reduce them; rather, much of the guidance of the model has focused on ways to increase value so that it outweighs the impact of the perceived costs (as we discussed above).

One aspect of continuing certification programs that may impact a physician's perceptions of cost are the monetary costs associated with certification. There is no research that we know of that has investigated the impact of financial cost of assessments on perceived costs within the expectancy-value framework. We hypothesize that any potential relations between monetary and perceived costs are likely impacted by one's other values for the test. For example, if one has high intrinsic, utility, or attainment value for the assessment that relation may mitigate any potential negative relation from the monetary cost.

In addition to learner's cognitive appraisals and perceptions of the test, the experience of testing can create additional in-the-moment performance costs (even if the learner is not aware of them). In particular, two phenomena that can be triggered in comprehensive high-stakes testing scenarios are test anxiety and stereotype threat.

Test Anxiety

The relationship between arousal, anxiety, and performance has been of interest for some time. Famously, the Yerkes-Dodson law of arousal and performance states that a moderate level of arousal leads to optimal performance whereas arousal that is too high or too low leads to suboptimal performance (Yerkes & Dodson, 1908, EL: 3). This "inverted U" model predicts poor performance at low levels of arousal because one is not adequately alert or engaged with the task and at high levels of arousal because one may experience anxiety and worry, which then interferes with performance. High

levels of arousal, anxiety, and worry have been investigated broadly across physical skills and performances as well as intellectual and academic contexts (Alpert & Haber, 1960, EL: 5; Beilock & Carr, 2001, EL: 3; Beilock, Schaeffer, & Rozek, 2017, EL: 2; Mandler & Sarason, 1952, EL: 4; Sarason, 1980, EL: 2).

Test anxiety is a multi-faceted construct consisting of physiological, psychological (e.g., emotional, cognitive), and behavioral components (Zeidner, 1998, 2009; von der Embse, 2018). It is hypothesized to emerge as worries or fear about receiving a negative evaluation on an evaluative test. Several models of test anxiety have been proposed and tested over time including interference (Alpert & Haber, 1960, EL: 5; Liebert & Morris, 1967, EL: 5), deficit (Tobias, 1985, EL: 2), and transactional models that incorporate components of the former two (Spielberger & Vagg, 1995, EL: 2; see von der Embse, 2018, EL: 1 for a review). More recently, biopsychosocial models have been proposed that focus on the interactive relations between biological, psychological, and social/environmental factors that trigger test anxiety in-the-moment (Segool et al., 2014, EL: 5; Jamieson, 2017, EL: 2).

Test anxiety, in general, is associated with poorer performance on classroom tests, GPA, IQ tests, and standardized tests (Ackerman & Heggedstad, 1997, EL: 1; Hembree, 1988, EL: 1; von der Embse, 2018, EL: 1). Although some arousal may be good, it is clear that many individuals approach standardized tests with levels of anxiety that are too high, such that deleterious effects are experienced (von der Embse, 2018, EL: 1). One reason that high levels of anxiety may be harmful to test taking is that it can reduce working memory resources (Beilock, 2008, EL: 2; Beilock & Carr, 2005, EL: 4; Moran, 2016, EL: 1). Moran (2016, EL: 1) examined the relationship between self-reported anxiety and working memory capacity in a meta-analysis (N = 22,061 participants) and found a small to moderate negative relationship (Hedges' g = -.33). However, there is still much debate about the boundary conditions of the relations and the exact mechanisms at play. One hypothesis is that anxiety impairs performance via multiple routes: worries impair verbal processing, and high arousal impairs spatial storage (Moran, 2016, EL: 1).

The deleterious effects of anxiety may intensify in response to high-pressure tests. Hinze and Rapp (2014, EL: 3) found that the benefits of retrieval practice (i.e., the testing effect) were diminished if there was significant performance pressure during episodes of memory retrieval. Additionally, Hinze and Rapp found that when episodes of retrieval practice were weighted more heavily (i.e., increasing performance pressure), final assessment test scores were lower than when retrieval practice episodes were more inconsequential. These findings illustrate the importance of reducing pressure during retrieval practice episodes in order to maximize learning outcomes. One possibility for reducing pressure would be to weigh the early retrieval practice questions less in the longitudinal assessment than later questions so if one does poorly early on they can identify areas of weakness and improve upon them. Moreover, the development of longitudinal assessment programs focused on learning benefits may reduce the perceived pressure of a singl, high-stakes test, relative to the moderate-stakes in any single portion of a spaced assessment.

Stereotype Threat

Stereotype threat refers to the diminished performance that can occur as a consequence of being reminded of a stereotype and how society expects that stereotype to perform. In a situation where a negative stereotype exists, actors of that stereotyped group may exhibit poorer performance because there is enormous anxiety about not wanting to reinforce a stereotype. Multiple mediating mechanisms have been proposed for this underperformance, including: the depletion of working memory resources (e.g., because working memory is being consumed to suppress negative thoughts), interference from attending to cognitive processes that are typically automatic, and strategies to protect one's self-concept (e.g., self-handicapping), among others (Spencer, Logel, & Davies, 2016, EL: 2). Stereotype threat has been observed in varying populations and domains, including women in math, Black Americans in higher education, white males in sports, and older adults in their episodic memory (Barber & Mather, 2013: EL: 3, Bouazzaoui, Fay, Guerrero-Sastoque, Semaine, Isingrini, et al., 2020, EL: 5; Nguyen & Ryan, 2008, EL: 1; Rahhal, Haher, & Colcombe, 2011: EL: 3; Steele & Aronson, 1995, EL: 5; Stone, Lynch, Sjomeling, & Darley, 1999, EL: 5). One meta-analysis (Nguyen & Ryan, 2008, EL: 1) found that the impact of stereotype threat on test performance was larger for ethnicity/race stereotypes than gender stereotypes (Cohen's *d* of .32 vs. .21, respectively). However, a more recent meta-analysis by Shewach, Sackett, and Quint (2019, EL: 1) found more similar effects across ethnicity and gender for stereotypes about cognitive ability (Cohen's *d* of .26 vs. .33, respectively).

Stereotype threat is more likely to occur when a test is more difficult (Nguyen & Ryan, 2008, EL: 1; Shewach, Sackett, & Quint, 2019, EL: 1). By contrast, Shewach, Sackett, and Quint (2019, EL: 1) found that when motivational incentives such a monetary reward are present, stereotype threat is much less pronounced than when monetary rewards are absent (Cohen's *d*s .14 vs. .41, respectively). This finding suggests that motivation likely plays an important role in stereotype threat.

Given the evidence for stereotype threat in testing, a few recommendations for longitudinal assessments may be beneficial. First, framing any longitudinal assessment in terms of its learning benefits may serve to lower anxiety and reduce perceptions of a test as "high-stakes" (see "reconstrual interventions" in Spencer, Logel, & Davies, 2016; EL: 2). Second, in situations where demographic information must be collected or presented, this should occur *after* any testing or at some other time not directly tied to the assessment. Third, testing materials should include a diverse cast of characters and should be carefully reviewed so as not to reinforce stereotypes through the testing content. If demographic information about the physicians is collected, it should occur considerably before testing (e.g., during initial sign-up) or after any testing, *not* immediately before.

Chapter Summary

In this chapter, we took an expectancy-value approach to thinking about the role of motivation in testing. We reviewed basic motivational theory and empirical work from laboratory and classroom settings and discussed their implications and applications in the context of continuing certification program assessments. This review suggests several motivational benefits of testing as well as some potential challenges posed by high-stakes standardized tests. Attention to instructional framing, test

purposes and values, and longitudinal assessment frameworks provide vehicles to further enhance motivational benefits and reduce costs.

Many of the motivational benefits for testing can be understood from the equation of having the perceived benefits of the test outweigh the perceived costs of preparing for and taking the assessment. We found that a sufficiently challenging test can facilitate both motivation to learn and later performance, as long as the test is not perceived as *too* difficult; that is, if learners perceive that investing effort is likely to increase success on the test. One way to make clear the level of difficulty is to describe the specific task items that will be used and to give representative problems to practice and receive feedback on.

We also reviewed a number of components of value (intrinsic, utility, and attainment) that should be attended to when designing the assessment. The ideal assessment should be perceived as relevant to the practitioner's interests (e.g., in terms of the topics and scenarios). It should be useful to furthering the practitioner's educational and professional goals, such as developing expertise and staying current. And, it should be perceived as important; that is, as an accurate measure of medical knowledge and skills and as an opportunity to confirm a physician's identity as a skilled medical expert. These values can be highlighted in the instructions and framing of the assessment and potentially in preparatory activities that might further reinforce them (e.g., a writing activity to discuss why this assessment is helpful to one's goals). Increasing these aspects of intrinsic and integrated value can also help to offset the perception of the test being simply an external punishment or reward, which can undermine motivation and engagement and diminish intrinsic interest. Similarly, framing the longitudinal assessment and feedback as an opportunity to learn and develop can further facilitate mastery-approach goal adoption and growth mindsets.

Complementing efforts to boost perceived value is an effort to mitigate perceived costs. It would be helpful to convey the task effort as reasonable and worthwhile. Perhaps describing the potential costs and expectations in advance to normalize them may help in reducing the overall perceived cost and therefore lead to higher motivation to learn.

Finally, we discussed other phenomena that can reduce motivation and performance: test anxiety and stereotype threat. The move to a longitudinal assessment scheme of more frequent testing may reduce test anxiety relative to less frequent, higher stakes tests, which in turn can help improve study behaviors and test performance. (See the Future Directions section, below, for some other possible types of interventions for test anxiety.) Similarly, reducing the stereotype threat is helpful for improving engagement and performance. Stereotype threat can be mitigated by highlighting the components of value previously described, emphasizing the assessment as an opportunity to improve (as opposed to a high-stakes test), not tying the collection of demographic data to the assessment, and including diverse demographic features in the testing clinical scenarios.

Future Directions

In reviewing this literature, we identified a number of directions for future research that were specifically lacking in the literature or not seen applied to the physician population of continuing

certification. These include five proposed studies where we supply the ideas but not necessarily a fully developed research design.

1. What is the Best Way to Measure Test Anxiety over Time?

Engagement and anxiety levels over time could be tracked by presenting Likert-scale questions periodically during the assessment (see Bernacki, Nokes-Malach, & Aleven, 2013, for an example in an intelligent tutoring system). For example, one could ask: *How engaged / anxious / focused / attentive do you feel right now*? Tracking engagement levels can serve several purposes. First, it can aid in creating assessments that instantiate optimal levels of engagement for learning outcomes. Second, it could similarly aid in creating a platform that is not overly anxiety-inducing and which individuals are more likely to enjoy. Third, collecting evidence on anxiety and learning may serve to recruit greater adoption of a longitudinal assessment program. For example, if it could be shown that longitudinal assessment leads to less self-reported anxiety than other testing methods, thus promoting the use of such methods.

2. *What are the Effects of Motivation in Longitudinal Testing?*

There is a lack of research on effects of motivation in longitudinal testing scenarios, including continuing certification programs. Thus, measuring whether the motivational components reviewed in this chapter---expectancies, perceived values and costs, achievement goals, and mindsets--predict performance could be extremely informative in assessment design decisions and potential interventions. For example, if utility value is strongly predictive of performance in this context, then test interventions (e.g., a brief writing task or instructional framing) could be tested to see whether perception of utility value could be increased. Further, measures of motivation could also serve as dependent measures for other interventions and test changes based on cognitive theory.

In addition, these data would contribute to basic science by characterizing motivation in the medical field and, more broadly, in an area where individuals have much more expertise in a domain than more novice populations. Such data would be relevant to theory testing and generalization and to understanding relations among motivational constructs.

3. *What is the Best Sequencing of Difficult Items to Increase Motivation in a Longitudinal Assessment?*

Prior work has suggested that the order of item difficulty in an assessment can affect one's motivation, preferences, and choices to engage in similar test activities later (Finn, 2010, EL: 3; Finn & Miele, 2016, EL: 3; Kimura, 2017, EL: 2). Research investigating participants solving math problems has found that participants prefer problem sequences with both difficult plus moderate problems than just the difficult problems alone (Finn, 2010, EL: 3 ; Finn & Miele, 2016, EL:3). Further, they found that participants preferred sequences in which the most difficult problems were in the middle of the sequence and not at the beginning or end (Finn & Miele, 2016, EL: 3). One could test whether this effect generalizes to continuing education contexts by testing the item difficulty and order of topics at the beginning, middle, or end of the item sequences. The prediction would be beginning or ending with easier problems would lead to higher levels of motivation to learn and engage with the material.

4. *What Interventions Increase Motivation and Performance in Longitudinal Assessments?*

A number of interesting interventions, including choice and personalization, have been designed to increase motivation and performance (Walkington & Bernacki, 2017, EL: 2; Patall, Cooper, & Robinson, 2008, EL: 1). Allowing individuals some choice of areas to be tested on could increase interest and engagement. Similarly, we hypothesize that personalizing the test to the individual taking it--matching test items to the context or contents of interest--would increase engagement, preparation, and performance outcomes. This would not involve any choice within the assessment system itself as the assessment system could automatically assign personalized test content based on an initial survey that the physician would take about their clinical practice. Another study could be related to a utility value intervention, such as a brief 10-minute writing exercise in which an individual writes about how the preparation and assessment is relevant to their education and professional goals. This intervention would be predicted to increase their utility value for the test and result in higher motivation to learn.

Building on interventions in the broader achievement goal literature (Elliot & Harackiewiz, 1996, EL: 3; Elliot & Dweck, 1988, EL: 4; Graham & Golan, 1991, EL: 3) one could examine how the task instructions and the structure of the assessment can help individuals adopt a mastery achievement goal. For example, instructions could emphasize understanding and improvement. The shift towards longitudinal assessment also allows for a focus on *intrapersonal comparison*; that is, focus on how a physician can improve compared to their past performance rather than relative to other physicians.

5. *How Best to Present Measurable Perceived Versus Actual Needs to Increase Motivation in Longitudinal Assessments?*

Discrepancy theory offers a method for measuring and instilling motivation in physicians based on needs (Fox & Miner, 1999). Although this method was originally proposed for other forms of continuing medical education, it can be adapted for longitudinal assessment programs. This method would have four steps: First, have physicians make a subjective rating for an aspect of clinical competency (e.g., knowledge of diabetes). Second, have a physician rate the subjective desirability of the target aspect of clinical competency (e.g., how important to you is it to have expert knowledge about diabetes?). Third, have a physician take an assessment of the target competency. Fourth, compute a discrepancy score between perceived and actual competency and present this information to the physician-learner. Thus, physicians who value a particular target competency, but who were unaware of a gap between their perceived and actual ability, may gain an intrinsic desire to improve. In addition, incorporating this information into a longitudinal assessment program provides learners with a means to address their gaps in knowledge and abilities. Specifically, highlighting gaps between perceived and actual competency could motivate a learner to focus on those particular areas or topics when studying for the next assessment.

References

Abel, M., & Bäuml, K.-H. T. (2020). Would you like to learn more? Retrieval practice plus feedback can increase motivation to keep on studying. *Cognition*, *201*, 104316.

Ackerman, P. L., & Heggestad, E. D., (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219-245.

Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *The Journal of Abnormal and Social Psychology, 61*(2), 207-215.

Amabile, T. M. (1985). Motivation and creativity: Effects of motivational orientation on creative writers. *Journal of Personality and Social Psychology*, *48*(2), 393-399.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*(3), 261-271.

Arens, A., Schmidt, I., & Preckel, F. (2019). Longitudinal relations among self-concept, intrinsic value, and attainment value across secondary school years in three academic domains. *Journal of Educational Psychology, 111*(4), 663-684.

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 576-587.

Bailey, J. M., Lombardi, D., Cordova, J. R., & Sinatra, G. M. (2017). Meeting students halfway: Increasing self-efficacy and promoting knowledge change in astronomy. *Physical Review Physics Education Research, 13*(2), 020140.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Barber, S. J., & Mather, M. (2013). Stereotype threat can both enhance and impair older adults' memory. *Psychological Science*, *24*(12), 2522–2529.

Barron, K. E., & Hulleman, C. S. (2015). Expectancy-value-cost model of motivation. In J. S. Eccles, & K. Salmelo-Aro (Eds.). *International Encyclopedia of Social and Behavioral Sciences, 2nd Edition: Motivational Psychology* (pp. 261-271). Amsterdam, Netherlands: Elsevier.

Beilock, S. L. (2008). Math performance in stressful situations. *Current Directions in Psychological Science, 17*(5), 393-343.

Beilock, S. L. & Carr, T. H. (2001). On the fragility of skilled performance: What governs choking under pressure? *Journal of Experimental Psychology: General, 130*(4), 701-725.

Beilock, S. L. & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science, 16*(2), 101-105.

Beilock, S. L., Schaeffer, M . W., & Rozek, C. S. (2017). Understanding and addressing performance anxiety. In A. J. Elliot, C. S. Dweck, & D. S. Yeager (Eds.) *Handbook of Competence and Motivation (2nd Edition): Theory and Application*. Guilford Press.

Belenky, D. M., & Nokes-Malach, T. J. (2012). Motivation and transfer: The role of mastery-approach goals in preparation for future learning. *Journal of the Learning Sciences*, *21(3)*, 399-432.

Belenky, D. M., & Nokes-Malach, T. J. (2013). Knowledge transfer and mastery-approach goals: Effects of structure and framing. *Learning and Individual Differences*, *25*, 21-34.

Bénabou, R., & Tirole, J. (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies*, *70*(3), 489-520.

Benjamin, A. S., & Tullis, J. G. (2010). What makes distributed practice effective? *Cognitive Psychology, 61*, 228-247.

Bereby-Meyer, Y., & Kaplan, A. (2005). Motivational influences of problem-solving strategies. *Contemporary Educational Psychology, 30*, 1-22.

Bernacki, M. L., Nokes-Malach, T. J., & Aleven, V. (2013). Fine-grained assessment of motivation over long periods of learning with an intelligent tutoring system: Methodology, advantages, and preliminary results. In R. Azevedo and V. Aleven (Eds.), *International Handbook of Metacognition and Learning Technologies* (pp. 629-644). NY: Springer.

Bernacki, M., & Walkington, C. (2018). The role of situational interest in personalized learning. *Journal of Educational Psychology*, *110*(6), 864–881.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246-263.

Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task value, and achievement goals. *Journal of Educational Psychology, 93*(1), 23–34.

Bouazzaoui, B., Fay, S., Guerrero-Sastoque, L., Semaine, M., Isingrini, M., & Taconnat, L. (2020). Memory Age-based Stereotype Threat: Role of Locus of Control and Anxiety. *Experimental Aging Research*, *46*(1), 39-51.

Bouffard-Bouchard, T., Parent, S., & Larvilee, S. (1991). Influence of self-efficacy on self-regulation and performance among junior and senior high-school age students. *International Journal of Behavior Development, 14*(2), 153-164.

Britner, S. L. (2008). Motivation in high school science students: A comparison of gender differences in life physical, and earth science classes. *Journal of Research in Science Teaching, 45*(8), 955-970.

Britner, S. L., & Parajes, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching, 43*(5), 485-499.

Cameron, J., & Pierce, W. D. (1994). Reinforcement, reward, and intrinsic motivation: A meta-analysis. *Review of Educational Research*, *64*(3), 363-423.

Castel, A. D., Benjamin, A. S., Craik, F. I., & Watkins, M. J. (2002). The effects of aging on selectivity and control in short-term recall. *Memory & Cognition*, *30*(7), 1078-1085.

Castel, A. D., Humphreys, K. L., Lee, S. S., Galván, A., Balota, D. A., & McCabe, D. P. (2011). The development of memory efficiency and value-directed remembering across the life span: A cross-sectional study of memory and selectivity. *Developmental Psychology*, *47*(6), 1553-1564.

Castel, A. D., Murayama, K., Friedman, M. C., McGillivray, S., & Link, I. (2013). Selecting valuable information to remember: Age-related differences and similarities in self-regulated learning. *Psychology and Aging*, *28*(1), 232-242.

Chesluk, B., Eden, A., Hansen, E., Johnson, M., Reddy, S., Bernabeo, E., & Gray, B. (2019). How physicians prepare for maintenance of certification exams: A qualitative study. *Academic Medicine*, *94*(12), 1931-1938.

Chesluk, B., Gray, B., Eden, A., Hansen, E., Lynn, L., & Peterson, L. (2019). That was pretty powerful: A qualitative study of what physicians learn when preparing for their Maintenance of Certification exams. *Journal of General Internal Medicine*, *34*(9), 1790-1796.

Conley, A. M. (2012). Patterns of motivation beliefs: Combining achievement goal and expectancy-value perspectives. *Journal of Educational Psychology, 104*(1), 32-47.

Connor, J. M. (1977). Effects of organization and expectancy on recall and recognition. *Memory & Cognition*, *5*(3), 315-318.

Costa, A., & Faria, L. (2018). Implicit theories of intelligence and academic achievement: A meta-analytic review. *Frontiers in Psychology, 9*, 829, 1-16.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268-294.

d'Ydewalle, G., Swerts, A., & De Corte, E. (1983). Study time and test performance as a function of test expectations. *Contemporary Educational Psychology*, *8*(1), 55-67.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology*, *18*(1), 105-115.

Deci, E. L. (1976). The hidden costs of rewards. *Organizational Dynamics*, *4*(3), 61-72.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627-668.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, *71*(1), 1-27.

Dunlosky, J., & Hertzog, C. (1998). Aging and deficits in associative memory: What is the role of strategy production? *Psychology and Aging, 13*(4), 597–607.

Durik, A. M., Vida, M., & Eccles, J. S. (2006). Task values and ability beliefs as predictors of high school literacy choices: A developmental analysis. *Journal of Educational Psychology, 98*(2), 382–393.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, *41*(10), 1040-1048.

Dweck, C. S., & Yeager, D. S. (2019). Mindsets: A view from two eras. *Perspectives on Psychological Science, 14*(3). 481-496.

Eccles, J. (2009). Who am I and what am I going to do with my life? Personal and collective identities as motivators of action. *Educational Psychologist*, *44*(2), 78-89.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109-132.

Eccles, J. S., & Wigfield, A. (2020). From expectancy-value theory to situated expectancy-value theory: A developmental, social cognitive, and sociocultural perspective on motivation. *Contemporary Educational Psychology, 61*, 101859.

Eisenberger, R., & Cameron, J. (1996). Detrimental effects of reward: Reality or myth? *American Psychologist*, *51*(11), 1153-1166.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, *34*(3), 169-189.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72*(1), 218-232.

Elliott, E. S. & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*(1), 5-12.

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 70*(3), 461-475.

Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 76*(4), 628-644.

Elliot, A. J., & McGregor, H. A. (2001). A 2 x 2 Achievement goal framework. *Journal of Personality and Social Psychology, 80(*3), 501-519.

Elliot, A.J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Educational Psychology, 91*(3), 549-563.

Elliot, A. J., & Murayama, K. (2008). On the measurement of achievement goals: Critique, illustration, and application. *Journal of Educational Psychology, 100*(3) 613-628.

Finn, B. (2010). Ending on a high note: Adding a better end to effortful study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1548–1553.

Finn, B. & Miele, D. B. (2016). Hitting a high note on math tests: Remembered success influences test preferences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 42,* 17-38.

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 632-652.

Flake, J. K., Barron, K. E., Hulleman, C., McCoach, B. D., & Welsh, M. E. (2015). Measuring cost: The forgotten component of expectancy-value theory. *Contemporary Educational Psychology, 41*, 232-244.

Fox, R. D., & Miner, C. (1999). Motivation and the facilitation of change, learning, and participation in educational programs for health professionals. *Journal of continuing Education in the Health Professions*, *19*(3), 132-141.

Galla, B. M., Plummer, B. D., White, R. E., Meketon, D., D'Mello, S. K., Duckworth, A. L. (2014). The Academic Diligence Task (ADT): assessing individual differences in effort on tedious but important schoolwork. *Contemporary Educational Psychology*, *39*(4), 314-325.

Glucksberg, S. (1962). The influence of strength of drive on functional fixedness and perceptual recognition. *Journal of Experimental Psychology*, *63*(1), 36-41.

Glucksberg, S. (1964). Problem solving: Response competition and the influence of drive. *Psychological Reports*, *15*(3), 939-942.

Gneezy, U., & Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, *115*(3), 791-810.

Graham, S., & Golan, S., (1991). Motivational influences on cognition: Task involvement, ego involvement, and depth of information processing. *Journal of Educational Psychology, 83*(2), 187-194.

Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology, 29*(4), 462–482.

Gunderson, E. A., Gripshover, S. J., Romero, C., Dweck, C. S., Goldin-Meadow, S., & Levine, S. C. (2013). Parent praise to 1- to 3-year olds predicts children's motivational frameworks 5 years later. *Child Development, 84*(5), 1526-1541.

Guo, J., Nagengast, B., Marsh, H. W., Kelava, A., Gaspard, H., Brandt, H., Cambria, J., Flunger, B., Dicke, A., Hafner, I., Brisson B., & Trautwein, U. (2016). Probing the unique contributions of self-concept, task values, and their interactions using multiple value facets and multiple academic outcomes. *AERA Open, 2*(1), 1-20.

Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition*, *4*(5), 507-513.

Haimovitz, K., Wormington, S. V., & Corpus, J. H. (2011). Dangerous mindsets: How beliefs about intelligence predict motivational change. *Learning and Individual Differences, 21*, 747-752.

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94*(3), 638–645.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94*(3), 562-575.

Harackiewicz, J. M., Canning, E. A., Tibbetts, Y., Priniski, S. J., & Hyde, J. S. (2016). Closing achievement gaps with a utility-value intervention: Disentangling race and social class. *Journal of Personality and Social Psychology*, *111*(5), 745-765.

Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology, 100*(1), 105-122.

Harackiewicz, J. M., & Priniski, S. J. (2018). Improving student outcomes in higher education: The science of targeted intervention. *Annual Review of Psychology*, *69*, 409-435.

Harackiewicz, J. M., Smith, J. L., & Priniski, S. J. (2016). Interest matters: The importance of promoting interest in education. *Policy Insights from the Behavioral and Brian Sciences, 3*(2), 220-227.

Harackiewicz, J.M., Tibbetts, Y, Canning, E.A., & Hyde, J.S. (2014). Harnessing values to promote motivation in education. In S. Karabenick and T. Urden (Eds.), Motivational Interventions, *Advances in Motivation and Achievement*, *Vol 18* (*pp* 71-105), Emerald Group Publishing.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47-77.

Henderson, V. L., & Dweck, C. S. (1990). *Motivation and achievement.* In S. S. Feldman & G. R. Elliott (Eds.), *At the threshold: The developing adolescent* (p. 308–329). Harvard University Press.

Hennessee, J. P., Knowlton, B. J., & Castel, A. D. (2018). The effects of value on context-item associative memory in younger and older adults. *Psychology and Aging*, *33*(1), 46-56.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, *70*(2), 151-179.

Hidi, S. & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*(2). 111-127.

Higgins, E. T. (1997). Beyond pleasure and pain. *American Psychologist, 52*(12), 1280-1300.

Higgins, E. T. (2012). *Regulatory focus theory.* In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of theories of social psychology* (p. 483–504). Sage Publications Ltd.

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology*, *28*(4), 597-606.

Honicke, T., & Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educational Research Review, 17*, 63-84.

Howard, J. L., Gagné, M., & Bureau, J. S. (2017). Testing a continuum structure of self-determined motivation: A meta-analysis. *Psychological Bulletin*, *143*(12).

Hulleman, C. S., Godes, O., Hendricks, B. L., & Harackiewicz, J. M. (2010). Enhancing interest and performance with a utility value intervention. *Journal of Educational Psychology, 102*(4), 880-895.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(5958), 1410-1412.

Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology*, *82*(3), 472-481.

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior*, *12*(5), 471-480.

Jamieson, J.P. (2017). Challenge and threat appraisals. In A. Elliot, C. Dweck, & D. Yeager (Eds.).Handbook of Competence and Motivation (2nd Edition): Theory and Application. New York, NY: GuilfordPress.

Kalender, Y., Marshman, E., Schunn, C., Nokes-Malach, T. J., & Singh, C. (2020). Damage caused by women's lowered self-efficacy on physics learning. *Physical Review Physics Education Research, 16 (1),* 010118.

Katz, I., Eilot, K., & Nevo, N. (2014). "I'll do it later": Type of motivation, self-efficacy and homework procrastination. *Motivation and Emotion*, *38*(1), 111-119.

Kimura, T. (2017).The impacts of computer adaptive testing from a variety of perspectives. *Journal of Educational Evaluation for Health Professions*, 14:12.

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education, 14*(1), 23-40.

Kruglanski, A. W., Friedman, I., & Zeevi, G. (1971). The effects of extrinsic incentive on some qualitative aspects of task performance 1. *Journal of Personality*, *39*(4), 606-617.

Kurzban, R., Duckworth, A., Kable, J. W., & Myers, J. (2013). Cost-benefit models as the next, best option for understanding subjective effort. *Behavioral and Brain Sciences*, *36*(6), 707-726.

La Guardia, J. G. (2009). Developing who I am: A self-determination theory approach to the establishment of health identities. *Educational Psychologist, 44*(2), 90-104.

León, J., Núñez, J. L., & Liew, J. (2015). Self-determination and STEM education: Effects of autonomy, motivation, and self-regulated learning on high school math achievement. *Learning and Individual Differences*, *43*, 156-163.

Leonard, J. M., & Whitten, W. B. (1983). Information stored when expecting recall or recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(3), 440-455.

Lepper, M. R., Corpus, J. H., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology*, *97*(2), 184-196.

Lepper, M. R., & Greene, D. (1975). Turning play into work: Effects of adult surveillance and extrinsic rewards on children's intrinsic motivation. *Journal of Personality and Social Psychology*, *31*(3), 479-486.

Lepper, M. R., Greene, D., & Nisbett, R. E. (1973). Undermining children's intrinsic interest with extrinsic reward: A test of the" overjustification" hypothesis. *Journal of Personality and social Psychology*, *28*(1), 129-137.

Li, Y., & Bates, T. C. (2019). You can't change your basic ability, but you work at things, and that's how we get hard things done: Testing the role of growth mindset on response to setbacks, educational attainment, and cognitive ability. *Journal of Experimental Psychology: General, 148*(9), 1640-1655.

Liebert, R. M., & Morris, L. W. (1967). Cognitive and emotional components of test anxiety: A distinction and some initial data. *Psychological Reports*, *20*(3), 975–978.

Linnenbrink-Garcia, L., Tyson, D. F., & Patall, E. A. (2008). When are achievement goal orientations beneficial for academic achievement? A closer look at main effects and moderating factors. *Revue Internationale De Psychologie Sociale, 21*, 19-70.

Maisto, S. A., Dewaard, R. J., & Miller, M. E. (1977). Encoding processes for recall and recognition: The effect of instructions and auxiliary task performance. *Bulletin of the Psychonomic Society*, *9*(2), 127-130.

Mandler, G., & Sarason, S. B. (1952). A study of anxiety and learning. *The Journal of Abnormal and Social Psychology*, *47*(2), 166–173.

McDaniel, M. A., Blischak, D. M., & Challis, B. (1994). The effects of test expectancy on processing and memory of prose. *Contemporary Educational Psychology*, *19*(2), 230-248.

McGillivray, S., & Castel, A. D. (2017). Older and younger adults' strategic control of metacognitive monitoring: The role of consequences, task, and prior knowledge. *Experimental Aging Research*, *43*(3), 233-256.

McGraw, K. O., & McCullers, J. C. (1979). Evidence of a detrimental effect of extrinsic incentives on breaking a mental set. *Journal of Experimental Social Psychology*, *15*(3), 285-294.

Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology, 82*(1), 60–70.

Meyer, J., Fleckenstein, J., & Koller, O. (2019). Expectancy value interactions and academic achievement: Differential relationships with achievement measures. *Contemporary Educational Psychology, 58*, 58-74.

Midgley, C., Arunkumar, R., & Urdan, T. C. (1996). "If I don't do well tomorrow, there's a reason": Predictors of adolescents' use of academic self-handicapping strategies. *Journal of Educational Psychology, 88*(3), 423–434.

Miele, D. B., Finn, B., & Molden, D. C. (2011). Does easily learned mean easily remembered? It depends on your beliefs of intelligence. *Psychological Science, 22*(3), 320-324.

Miele, D. B., & Molden, D. C. (2010). Naive theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General, 139*(3), 535-557.

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin*, *142*(8), 831-864.

Mueller, C. M., & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33–52.

Neely, J. H., & Balota, D. A. (1981). Test-expectancy and semantic-organization effects in recall and recognition. *Memory & Cognition*, *9*(3), 283-300.

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314-1334.

O'Keefe, P. A., & Linenbrink-Garcia, L. (2014). The role of interest in optimizing performance and self-regulation. *Journal of Experimental Social Psychology*, *53*, 70-78.

Pajares, F. (2008). *Motivational role of self-efficacy beliefs in self-regulated learning.* In D. H. Schunk & B. J. Zimmerman (Eds.), *Motivation and self-regulated learning: Theory, research, and applications* (p. 111–139). Lawrence Erlbaum Associates Publishers.

Patall, E. A., Cooper, H., & Robinson, J. C. (2008). The effects of choice on intrinsic motivation and related outcomes: A meta-analysis of research findings. *Psychological Bulletin*, *134*(2), 270-300.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science, 26*(6), 784-793.

Penk, C., & Schipolowski, S. (2015). Is it about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42*, 27-35.

Pintrich, P. R., & De Groot, E. V. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of educational psychology*, *82*(1), 33.

Pittman, T. S., Emery, J., & Boggiano, A. K. (1982). Intrinsic and extrinsic motivational orientations: Reward-induced changes in preference for complexity. *Journal of Personality and Social Psychology*, *42*(5), 789-797.

Plante, I., O'Keefe, P. A., & Theoret, M. (2013). The relation between achievement goal and expectancy-value theories in predicting achievement-related outcomes: A test of four theoretical conceptions. *Motivation and Emotion, 37*, 65-78.

Postman, L. (1964). Studies of learning to learn II. Changes in transfer as a function of practice. *Journal of Verbal Learning and Verbal Behavior*, *3*(5), 437-447.

Priess-Groben, H. A., & Hyde, J. S. (2017). Implicit theories, expectancies, and values predict mathematics motivation and behavior across high school and college. *Journal of Youth Adolescence, 46*, 1318-1332.

Putwain, D. W., Nicholson, L. J., Pekrun, R., Becker, S., & Symes, W. (2019). Expectancy of success, attainment value, engagement, and achievement: A moderated mediation analysis. *Learning and Instruction*, *60*, 117-125.

Rahhal, T. A., Hasher, L., & Colcombe, S. J. (2001). Instructional manipulations and age differences in memory: Now you see them, now you don't. *Psychology and Aging*, *16*, 697–706.

Renninger, K. A. (2000). *Individual interest and its implications for understanding intrinsic motivation.* In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (p. 373–404). Academic Press.

Renninger, K. A., & Hidi, S. (2002). *Student interest and achievement: Developmental issues raised by a case study.* In A. Wigfield & J. S. Eccles (Eds.), *A Vol. in the educational psychology series. Development of achievement motivation* (p. 173–195). Academic Press.

Renninger, K. A., & Hidi, S. E. (2019). *Interest development and learning.* In K. A. Renninger & S. E. Hidi (Eds.), *Cambridge handbooks in psychology. The Cambridge handbook of motivation and learning* (p. 265–290). Cambridge University Press.

Richey, J. E., & Nokes-Malach, T. J. (2013). How much is too much? Learning and motivation effects of adding instructional explanations to worked examples. *Learning and Instruction*, *25*, 104-124.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54-67.

Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 101860.

Sarason, I. G. (1980). (Ed.) *Test Anxiety: Theory, Research, and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Scheier, M. F., & Carver, C. S. (1982). Self-consciousness, outcome expectancy, and persistence. Journal of Research in Personality, 16(4), 409-418.

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*, *26*(3-4), 299-323.

Schiefele, U., Krapp, A., & Winteler, A. (1992). Interest as a predictor of academic achievement: A meta-analysis of research. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (p. 183–212). Lawrence Erlbaum Associates, Inc.

Schiefele, U., Wild, K. P., & Krapp, A. (1995, August). Course-specific interest and extrinsic motivation as predictors of specific learning strategies and course grades. In 6th EARLI Conference in Nijmegen.

Schmidt, S. R. (1988). Test expectancy and individual-item versus relational processing. *The American Journal of Psychology*, *101*(1), 59-71.

Segool, N. K., von der Embse, N. P., Mata, A. D., Gallant, J. (2014). Cognitive behavioral model of test anxiety in a high stakes context: An exploratory study. *School Mental Health, 6*, 50-61.

Senko, C., Hulleman, C. S., & Harackiewicz, J. M. (2011). Achievement goal theory at the crossroads: Old controversies, current challenges, and new directions. *Educational Psychologist*, *46*(1), 26-47.

Shewach, O. R., Sackett, P. R., & Quint, S. (2019). Stereotype threat effects in settings with features likely versus unlikely in operational test settings: A meta-analysis. *Journal of Applied Psychology*, *104*(12), 1514-1534.

Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. S. Eccles (Eds.), *A Vol. in the educational psychology series. Development of achievement motivation* (p. 15–31). Academic Press.

Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, *42*(1), 70.

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204.

Spielberger, C. D., & Vagg, P. R. (1995). Test anxiety: A transactional process model. In C. D. Spielberger & P. R. Vagg (Eds.), *Series in clinical and community psychology. Test anxiety: Theory, assessment, and treatment* (p. 3–14). Taylor & Francis.

Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology*, *67*, 415-437.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811.

Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology*, *77*(6), 1213-1227.

Szpunar, K. K., McDermott, K. B., & Roediger, H. L. (2007). Expectation of a final cumulative test enhances long-term retention. *Memory & Cognition*, *35*(5), 1007-1013.

Taylor, G., Jungert, T., Mageau, G. A., Schattke, K., Dedic, H., Rosenfield, S., & Koestner, R. (2014). A self-determination theory approach to predicting school achievement over time: The unique role of intrinsic motivation. *Contemporary Educational Psychology*, *39*(4), 342-358.

Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology: Section A*, *49*(4), 901-918.

Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist*, *20*(3), 135-142.

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: A latent interaction modeling study. *Journal of Educational Psychology*, *104*(3), 763.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research, 78*(4), 751-796.

Vallerand, R. J., & Blssonnette, R. (1992). Intrinsic, extrinsic, and amotivational styles as predictors of behavior: A prospective study. *Journal of Personality*, *60*(3), 599-620.

von der Embse, N., Jester, D., Roy, D., Post, J. (2018). Text anxiety effects, predictors, and correlates: A 30 year meta-analytic review. *Journal of Affective Disorders, 227*, 483-493.

Walkington, C., & Bernacki, M. L. (2017). Personalization of instruction: Design dimensions and implications for cognition. *Journal of Experimental Education, 86*(1), 50–68.

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68-81.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs and values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 92-120). San Diego: Academic Press.

Wigfield, A., Tonks, S., & Klauda, S. L. (2016). Expectancy-value theory. In K. R. Wentzel & D. Miele (Eds.), *Handbook of motivation in school* (2nd ed., pp. 55-74). New York: Routledge.

Williams, G. C., Saizow, R., Ross, L., & Deci, E. L. (1997). Motivation underlying career choice for internal medicine and surgery. *Social Science & Medicine*, *45*(11), 1705-1713.

Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., Lee, H.Y., O'Brien, J., Flint, K., Roberts, A., Greene, D., Walton, G.M., Dweck, C.S., & Trott, J. (2016). Using design

thinking to improve psychological interventions: The case of the growth mindset during the transition to high school. *Journal of Educational Psychology*, 108(3), 374.

Yeager, D. S., Hanselman, P., Walton, G.M., Murray, J.S., Crosnoe, R., Muller, C., Tipton, E., Schneider, B., Hulleman, C.S., Hinojosa, C.P., Paunesku, D., Romero, C., Flint, K., Roberts, A., Trott, J., Iachan R., Buontempo, J., Yang, S.M., Carvalho, C.M., Hahn P.R., Gopalan, M., Mhatre, P., Ferguson, R., Duckworth, A.L., Dweck, C. S. (2019). A national study reveals where a growth mindset improves achievement. *Nature, 573*, 364-369.

Yerkes, R. M., Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*. *18*(5), 459–482.

Zeidner, M. (1998). *Test anxiety: The state of the art.* New York: Plenum Press.

Zeidner, M. (2009). Test anxiety. In I. B. Weiner and W. E. Craighead (Eds.) *The Corsini Encyclopedia of Psychology* (pp. 1-3). John Wiley & Sons, Inc.

Chapter 6: Synthesis and Summary

In this final chapter, we synthesize the key points from the earlier chapters. First, we explain the key empirical findings through a new model of learning and assessment. Second, we summarize the key recommendations for longitudinal assessment programs based on the current state of research. Third, we suggest novel research that boards may consider implementing within longitudinal assessment programs to identify ways to improve the programs (while also contributing to the basic science of learning). Fourth, we discuss longitudinal assessment programs in the context of other forms of life-long learning that physicians experience, and we discuss how longitudinal assessment fills certain gaps in learning opportunities.

Assessment and Learning Model

Figure 6-1 synthesizes and summarizes the key points from our research. It is meant to capture many of the most important processes discussed in the prior chapters; however, due to the length of this report, it is not intended to be exhaustive. Though some parts of the figure can be understood on their own, the figure is also intended as a guide back to the text. The following subsections summarize and synthesize the four main chapters from this report, and thereby help explain the figure. The nodes and relations in the figure are explained in more detail in prior chapters.

In Figure 6-1, arrows denote the causal processes or mechanisms that explain the relationships among the variables. Arrows with solid lines represent increasing relations and arrows with dotted lines represent decreasing relations. The four boxes are used to group the variables discussed in each of the four chapters.

Cognitive Skills Must Be Kept Current

As physicians get further and further out of residency, three processes occur in parallel. First, physicians accumulate more clinical experience over time. This extensive clinical experience can exert a positive effect on patient care (particularly in areas in which physicians choose to focus their practice) because it allows for quick pattern recognition, which often produces accurate diagnoses and other useful clinical decisions. However, it can also be negative insofar as a physician's clinical experience is inherently idiosyncratic, and some physicians choose to narrow their practices over time, which may leave gaps in knowledge and introduce bias by distorting perceptions of prevalence. These idiosyncrasies, biases, and gaps in knowledge can lead physicians to make decisions that deviate from standards of care.

Second, over time, physicians also experience cognitive aging, one implication of which is that they will tend to rely more heavily on habitual routines, rather than learning new ones, and they may also have more difficulty balancing multiple tasks in working memory. Research shows that, on average, older physicians tend to provide poorer quality of care than younger physicians; however, the specific mechanisms of this finding are unclear because age is correlated with multiple other factors, including

time since residency, changes in standards of care, the accumulation of (varied) clinical experiences and changes in pattern recognition, and specialization or changes in clinical practice.

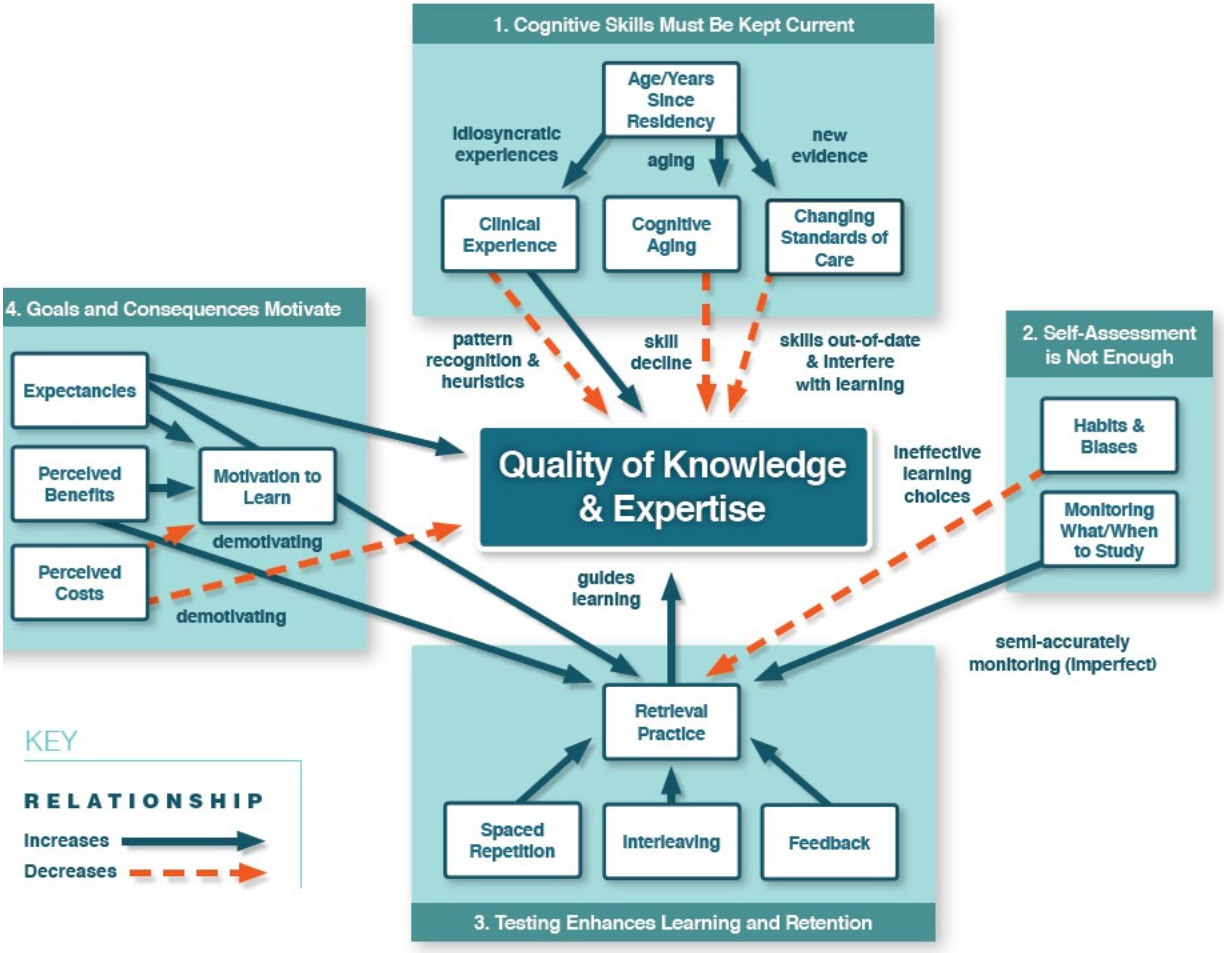


Figure 6-1. Synthesis of Topics Influencing Quality of Knowledge and Expertise

Third, physicians need to learn new standards of care as standards can change over time. Staying up to date can be difficult because it involves several processes. In particular, physicians must 1) be initially exposed to a new standard of care relevant to their practice, 2) gain knowledge of the new standard, 3) agree with the new standard, 4) feel confident that they can implement it and 5) remember to use the new standard when appropriate. Each individual barrier can be a challenge, and because there are multiple barriers, there are multiple potential points of failure to learn and implement new standards.

Self-Assessment is not enough

Is self-assessment enough to keep cognitive skills current? Prior research does support the importance of accurately self-assessing one's own skills and abilities. However, individuals do not have direct access to this information and instead use "informed guesses" which, though somewhat accurate,

suffer from systematic biases and cannot be easily addressed. For example, information that feels easy to process in-the-moment can lead individuals to overconfidence in their ability to remember it in the future. Another notable phenomenon well-established in the self-assessment literature is the Dunning-Kruger effect, the robustly supported finding that the poorest performers are also the least accurate in their self-assessments and tend to overestimate their actual ability. (By contrast, the top performers tend to *under*estimate their ability, though to a lesser degree.) People also tend to underestimate how much they will forget. This implies that physicians may think that they need less continued training to maintain a given level of knowledge than they actually do. And lastly, people tend to avoid many of the best learning strategies, such as self-testing, because the sense of difficulty they engender feels like poorer learning, and they instead prefer other forms of learning that feel better, but are actually less effective. This implies that, if given the choice, many physicians will study in ways that are less effective or efficient than if directed by a longitudinal assessment program.

For these reasons, continuing certification programs are needed to provide a more objective assessment of a physician's ability. At the same time, given that individuals do have some ability to accurately self-assess their own knowledge, this can potentially be leveraged by giving physicians some control over the topics included in the assessment.

Testing Enhances Learning and Retention

Whereas the goal of continuing certification programs has traditionally been to assess whether physicians are maintaining skills and keeping up with changing standards, the switch to longitudinal assessment presents the opportunity to serve learning as well as assessment purposes. Although assessments are often viewed merely as tools for decision making about one's performance level, strong evidence supports the experience of being tested as a powerful learning experience in its own right: The act of retrieving targeted information from memory strengthens the ability to use it again in the future, so that new and old standards of care can remain distinct and readily accessible.

Testing is further strengthened when followed by feedback, a phenomenon too often lacking in medical practice itself. By having tests spaced out over time: greater frequency of testing leads to deeper learning. However, the optimal frequency and number of tests a physician takes should be weighed against the burden to physicians. Research suggests that topics that are hard to distinguish can generally be better learned by intermixing rather than presenting them one-at-a-time, but there is a need for future research to identify the exact sequence that is optimal in medicine. Another benefit to creating a longitudinal assessment program may be that it results in physicians adopting more effective study and learning habits as they are guided to experience the learning benefits of self-testing.

Goals and Consequences Motivate

Testing can also serve as an important motivator. Physicians will be more motivated to study and practice their skills when the perceived benefits of doing so outweigh the perceived costs of not doing so. The expectation of specific, challenging assessments can lead people to study longer and more meaningfully; thus, testing should be challenging enough to engender deeper and more effective learning but also not so difficult as to lead to expectations of failure. For these reasons, longitudinal assessments could further motivate physicians' use of, and learning from, retrieval practice.

Physicians are also typically intrinsically motivated (i.e., internally driven) to learn and improve in their respective medical field. Emphasizing how maintenance of medical expertise aligns with physicians' values can increase the perceived benefits of preparing for and engaging with longitudinal assessment to further facilitate one's motivation to learn. For example, aligning the assessment with the physicians' interests (topics and scenarios), educational and career goals (e.g., developing expertise and staying current), and as an accurate measure of an important aspect of their knowledge and skills. Longitudinal assessment programs would benefit from emphasizing congruence between physicians' values and relevant learning outcomes, such as staying current on the latest knowledge relevant to the care provided by the physicians.

In addition, decreasing or mitigating the perceived costs of the assessment is also important. More frequent testing may help reduce test anxiety and stereotype threat relative to less frequent, higher stakes tests, which in turn can help improve study behaviors and test performance. Increasing a physician's motivation to learn, in turn, leads individuals to work harder, persist longer in the face of difficulty, adopt better learning strategies, and procrastinate less than when they are motivated by only external rewards.

A Cross-Cutting Theme - Feedback on Performance

One cross-cutting theme across multiple chapters of this report is the role of feedback. In the chapter *Cognitive Skills Need to be Kept Current,* we discussed how accurate and timely feedback is for the development of expertise in any domain. However, clinical systems often provide imperfect feedback mechanisms. For example, if a physician makes an incorrect diagnosis, the patient may never receive the correct diagnosis, and even if they do, the correct diagnosis may not be conveyed back to the physician who made the incorrect diagnosis or instituted inappropriate treatment. Schiff (2008, EL: 6) reports that physicians often learn about their diagnostic success in an ad-hoc manner (e.g., malpractice subpoenas, running into a colleague) and that, as a result, physicians lack a reliable system for learning from past errors. Therefore, imperfect feedback systems in medicine lead to suboptimal learning opportunities for physicians.

In the chapter *Self-Assessment is not enough*, we discussed how, in the absence of external feedback, people need to rely on their own internal monitoring to assess what they do vs. do not know. Though individuals do have some ability to monitor what they do versus do not know, this internal monitoring is imperfect in a variety of ways. Poor metacognitive accuracy is particularly problematic in high-stakes environments like medicine because it would mean that a physician is making incorrect decisions with high confidence. In terms of performance, the poorest performers in a domain tend to be least accurate in their self-assessments, overestimating their knowledge. This overestimation is believed to derive from the same lack of knowledge that caused them to perform poorly in the first place. Stepping back, it makes sense that insufficient feedback is the underlying cause of both poor knowledge/skills and subsequent overestimation of one's knowledge. Therefore, we expect that better learning through testing with feedback should improve both accuracy and metacognitive understanding of one's strengths and weaknesses.

In the chapter *Testing Enhances Learning and Retention*, feedback was discussed extensively. Though testing improves memory even without feedback of the correct answer, testing with feedback is

even more effective. We later identify a number of open questions (see below for proposed studies) regarding precisely how and when to provide feedback.

In the chapter *Goals and Consequences Motivate*, we discuss how feedback is critical to a number of aspects of motivation. Feedback is one important factor in the development of beliefs of self-efficacy. Both positive and negative feedback can influence one's beliefs of self-efficacy. Longitudinal assessments more generally provide an opportunity for individuals to get multiple pieces of feedback over time and the opportunity to improve self-efficacy with practice and sustained effort. Feedback is also critical to achievement goals and is needed to help one determine whether they are accomplishing their goals, for example, whether one is accomplishing a goal of self-improvement and increased knowledge one needs feedback to compare performances over time. Feedback also plays a critical role in the impact of mindsets on performance and behavior. For example, growth mindsets have been hypothesized to be particularly important for situations where one receives negative feedback. The type of feedback also matters. If one is given feedback that highlights future opportunities for growth and improvement, that feedback will be viewed differently than one-time, high-stakes evaluative feedback. The latter often is viewed as a contributing factor leading to high test anxiety.

Recommendations

Recommendations for longitudinal assessment have been made throughout the previous chapters. In Table 6-1 we compiled the recommendations so that they can be easily skimmed. The numbered items correspond to specific boxes within each chapter. Additional context for the recommendations is provided within each chapter's text and will be near each pop-out box. Recommendations are listed in the same order and with the same numbers as they appear in the chapters - the order is not intended to imply an order of importance.

Table 6-1: List of recommendations from All Chapters and Evidence Level

	Recommendations	**Source**	**Evidence Level ***
	Cognitive Skills Need to be Kept Current		
2-1	Physicians have idiosyncratic experiences, which can lead to biased judgments. Therefore, longitudinal assessment can attempt to fill in potential gaps in experience through rich clinical vignettes in a continuous framework.	Brooks et al., 1991; Choudhry et al., 2006; Hatala, Norman, & Brooks, 1999	4
2-2	Common concepts (e.g., common diagnoses, common treatment plans) may interfere with more rare concepts (e.g., rare diagnoses, less frequently used treatment plans). Due to interference, one goal for longitudinal assessment could be to test knowledge of the rare but important	Anderson, Bjork, & Bjork, 1994; Roediger, 1978; Watkins & Watkins, 1975; Postman & Underwood, 1973	2

	concepts that are especially easy to confuse with more common concepts.		

	Self-Assessment is Not Enough		
3-1	If confidence ratings are to be collected, they should be collected before feedback is provided; confidence ratings after feedback are likely to be inaccurate.	Benjamin, Bjork, & Schwartz, 1998	3
3-2	Any longitudinal assessment system should be designed with fluency and ease of use in mind.	Carpenter, Wilford, Kornell, & Mullaney, 2013; Fiechter, Fealing, Gerrard, & Kornell, 2018	3
	Testing Enhances Learning and Retention		
4-1	Testing should be used to support long-term retention of knowledge and cognitive skills.	Adesope, Trevisan, & Sundararajan, 2017; Rowland, 2014; Yang, Luo, Vadillo, Yu. & Shanks, 2021	1
4-2	Because *more* tests will always be better than *less* tests, considerations for how frequent testing occurs will need to be weighed against practical considerations.	Driskell, Willis, & Copper, 1992	1
4-3	It is preferable to have repeated testing opportunities spread out over time than one big test every five years or every ten years but with no intermediary testing opportunities in between.	Cepeda, Coburn, Rohrer, Wixted, Mozer, & Pashler, 2009	3
4-4	Concepts should not be blocked within the assessment (i.e., having blocks of contiguous questions on a single topic and then never come back to this topic). Concepts that share overlapping features, and may be more prone to misidentification, should especially be interleaved with each other to aid in differentiation/reduce interference.	Brunmair & Richter, 2019	1
4-5	Being tested should improve physicians' retention not just of the specific tested material, but on other related material too.	Pan & Rickard, 2018	1
4-6	Feedback should be provided for both correct and incorrect responses on a test.	Bangert-Drowns, Kulik, Kulik, & Morgan, 1991; Keith & Frese, 2008	1

4-7	Feedback provided on tests should include evidence and/or references with claims.	Gnepp, Klayman, Williamson, & Barlas, 2020	3
	Goals and Consequences Motivate		
5-1	It is important to develop an assessment that is challenging enough to confer motivational benefits, but not so difficult that it is perceived as likely to result in failure.	Shunk & Parajes, 2002	2
5-2	Longitudinal assessment programs should instill motivational benefits to physicians by including topics that are especially important to the learner while balancing this with appropriate coverage of the content in the discipline.	Walkington & Bernacki, 2017	2
5-3	Basic research suggests that how testing organizations frame the assessment to physicians can impact how they perceive its usefulness (e.g., as an opportunity to develop relevant skills rather than as a required assessment).	Spencer, Logel, & Davies, 2016	3

* The evidence levels ranges from 1 to 6, with 1 being the strongest evidence (meta-analyses) and 6 being the weakest evidence (opinion papers).

Proposed Studies

In this section we list the studies that may help us understand how to enhance continuing certification programs. Each of these studies is discussed at the end of the four chapters in some detail. We are not speculating on the outcomes of the proposed studies; only that the results should be informative. The study results may help policy-makers find ways to balance competing demands on their assessment program, such as, the need for standardization to uphold valid and reliable pass-fail decisions and the need for substantial flexibility to customize learning to meet the physicians' needs.

Cognitive Skills need to be Kept Current

1. *What is the most useful feedback to emphasize the new standards of care?*
2. *What is the relationship between response time, performance, and age in assessments?*
3. *Can small interventions help with improving clinical reasoning skills?*
4. *What impact does non-analytical reasoning have on arriving at a correct diagnosis?*
5. *What is the effect of feedback on outcome expectancy and self-efficacy?*

Self-Assessment is not enough

1. *Is asking about relevance best before or after feedback?*
2. *What is the most appropriate response scale for confidence judgments?*
3. *How does physician customization enhance or degrade the assessment?*
4. *Does self-assessment accuracy differ across levels of granularity of the subject matter?*
5. *Does self-assessment accuracy differ across objective versus comparative feedback?*

Testing Enhances Learning and Retention

1. *What type of explanation of the correct answer in the feedback provided improves learning and retention?*
2. *What is the optimal timing for delivering feedback for improving learning and retention?*
3. *Is Interleaving of different content domains beneficial to learning and retention?*
4. *Are citations important elements of feedback to improve learning and retention?*

Goals and Consequences Motivate

1. *What is the best way to measure test anxiety over time?*
2. *What are the effects of motivation in longitudinal testing?*
3. *What is the best sequencing of difficult items to increase motivation in a longitudinal assessment?*
4. *What interventions increase motivation and performance in longitudinal assessments?*
5. *How best to present measurable perceived versus actual needs to increase motivation in longitudinal assessments?*

The Role of Longitudinal Assessment in Comparison to Other Life-Long Learning Mechanisms

The bulk of this paper has focused on the basic science of learning and the affordances of longitudinal assessment for learning. However, over the course of a physician's career, they engage in multiple different forms of life-long learning. All physicians continue to learn through continuing medical education (CME) and through personal experience with patients. Some physicians work in settings in which they receive best practice alerts and or audit and feedback. None of these learning modalities is perfect and all have strengths and weaknesses. In this section, we attempt to characterize some of the most salient strengths and weaknesses of these different types of learning (summarized in Table 6-2).

In the next section, we outline six features of learning opportunities that we consider to be important when considering how likely the opportunity would be to lead to learning. In the following section, we discuss six different learning opportunities for physicians after residency training, and discuss the learning features present and absent for each. We are not implying that each type of learning opportunity should have each feature; there very well could be benefits of having multiple different learning opportunities with different emphases. Rather, our goal is simply to create a framework for thinking about the similarities and differences between the learning opportunities.

Features of Learning Opportunities

We consider the following six features to be of critical importance for facilitating and tracking learning (though there may also be other features that we have not listed). First, the chapter *Testing Enhances Learning* is focused on the benefits of testing - **retrieval practice** - for learning. Given that retrieval practice is so effective, we believe that it can be a critical component in lifelong learning.

Second, receiving **feedback** about one's judgments is considered to be critical for becoming an expert (Kahneman & Klein, 2009, EL: 2), yet in everyday practice physicians often do not get useful feedback about whether their diagnoses and treatment plans are correct (Schiff, 2008; EL: 6). The important role of feedback is discussed throughout this report and in a cross-cutting section earlier in this chapter. A key point is that, even though testing is beneficial on its own, testing plus feedback is considerably more effective, especially for correcting errors. Nevertheless, the best way to structure feedback (particularly if a user answers incorrectly) merits more study. Ideally, feedback would promote learning and retention and thereby increase the likelihood the knowledge is applied in future patient care situations for which it is relevant.

Third, within the chapter *Testing Enhances Learning,* we discussed extensive evidence for the benefits of **spaced learning**: Learning is more effective and efficient when it is spaced out evenly across time than when it occurs in bunches (e.g., cramming right before a test). Given the robust evidence of the benefits of spaced learning, we added it as a desirable criterion here.

Fourth, within the chapter *Testing Enhances Learning,* we argue that **interleaving** concepts that are similar, rather than presenting blocks of content on the same topic, produces better learning and discrimination. Similar to spaced learning, interleaving also has a robust evidence base.

Fifth, another cross-cutting topic that we discussed above in the section *Valuing Longitudinal Assessment vs. Self-Directed Learning* is the degree to which physicians should have control over both the topics that are included and the ways in which they engage in the assessment program. While **self-directed learning** may benefit physician's intrinsic motivation, it is likely that having complete control over learning would lead to ineffective choices. In sum, the evidence suggests that professional learners should have *some* degree of control over which topics they learn; it is clear that they should not have complete control, nor should they have no control. The exact amounts and type of control are open questions worthy of further investigation.

Sixth, the chapter *Goals and Consequences Motivate* reviews evidence about how people are driven by **consequences**--by the perceived benefits and costs of taking a test and performing well on it. For example, a certain level of arousal is beneficial for learning and performance, but too much is harmful. Furthermore, a sufficiently challenging assessment can foster both the motivation to learn and one's ultimate performance, as long as the assessment is not perceived as too difficult. In sum, having some assessments with consequences is beneficial.

Finally, a topic that has not been discussed so far in this report is whether learning is "authentic" or "naturalistic." Within medical education specifically, and learning sciences more broadly, there are concerns that if a learning environment is too artificial, it will do a poor job of preparing learners for the real-world tasks, and that if a testing environment is too artificial, it will do a poor job of predicting real-world performance. A theory from cognitive psychology called *transfer-appropriate processing* proposes

that learning and retention is generally better when the learning environment matches the testing or practice environment (Blaxton, 1989, EL: 3).

Yet, others have observed that some efforts to create learning environments that are highly naturalistic--particularly high-fidelity patient simulators--do not produce learning benefits over low fidelity simulators in skills such as auscultation, surgical motor skills, and critical care and crisis management skills (Norman, Dore, and Grierson, 2012, EL: 2). Others have found that scores on high-fidelity clinical simulations are too imprecise unless impractically large numbers of simulations are used and that multiple choice questions can yield equally high criterion validity in a much shorter amount of time (Swanson, Norcini, Grosso, 1987, EL: 2). The new situated-cognition model of clinical reasoning takes the view of naturalistic or authentic reasoning a step farther. This model (Graber, 2020; Merkebu et al., 2020) stresses that clinical reasoning is not just in the head of the physician but is a much more complex process that involves interactions with the patient and medical team. Thus, advocates for the situated-cognition model have suggested that assessments of clinical reasoning need to go beyond the simple cognitive decision-making that is assessed in multiple choice tests and assess how the physician performs within the complex environment of a medical situation (Rencic, Schuwirth, Gruppen, & Durning, 2020a, 2020b; Schuwirth, Durning, & King, 2020; Torre et al., 2020). Doing so in a standardized way is obviously a major challenge and currently outside the scope of continuing certification programs. Still, the situated cognition model highlights the importance of authentic learning and assessment.

Six Lifelong Learning Opportunities

Table 6-2 classifies each cell as to whether a particular learning opportunity has a feature--as yes, no, or somewhat. For many of these cells the answers are more complex, and are discussed below.

Table 6-2. Comparison of Features for Keeping Cognitive Skills Current across Learning Opportunities

Features of Learning Opportunities	Traditional 10-year Assessment	Longitudinal Assessment	CME	Clinical Experience	Clinical Decision Support Systems	Audit and Feedback
Retrieval Practice	Y	Y	N	Y		
Feedback	N	Y	Y	S	Y	Y
Spaced	N	Y	S	Y		
Self-Directed	N	TBD	Y	Y	N	N
Consequences	Y	Y	S	Y		
Authentic	N	N	N	Y		

Note. Y = Yes, N = No, S = Somewhat, TBD = To Be Determined.

Traditional Certification

Traditional certification assessments have a main goal of summative assessment, not learning in and of itself, although studying for the assessment does induce learning. Consequently, of the learning

features reviewed above, the main one included in traditional assessment is consequences: If a physician fails and does not pass on repeated attempts within the time window, then they lose certification until they successfully pass. Though general feedback is provided about whether a physician passed the assessment or not, as well as their percentile on the assessment, and sometimes feedback on areas of weakness by topic, feedback specifically on individual items is not provided. This type of assessment can still serve as a form of retrieval practice--it could help reinforce knowledge that the physician already has --but because it does not include detailed feedback, it cannot help the physician understand their mistakes and could potentially reinforce their wrong answers.

Because certification assessments have traditionally been spaced far apart--10 years for many boards--instead of more frequent smaller assessments, they do not capitalize on the benefits of spaced learning. Although some traditional certification assessments allow some degree of customization, such as selecting content specific modules, generally most of the assessment is standardized. And, because traditional certification assessments are largely multiple-choice that take place outside of clinical practice, they are not as authentic as some of the other learning opportunities that more directly reflect--or are even embedded within--clinical practice, such as clinical decision support systems and audit and feedback.

Longitudinal Assessment

Longitudinal assessment is designed to capitalize on certain learning opportunities that traditional assessments do not. In particular, whereas traditional assessment does not provide feedback about individual questions, longitudinal assessment does, which will allow it to serve as a learning opportunity. Another change is that, since the assessments will be happening more frequently, longitudinal assessment will capitalize on the advantages of spaced learning.

One topic that each board needs to consider is the extent to which learning will be self-directed; that is, whether physicians will get any choice in topics that they want to be assessed on and learn about. As argued above, we believe that giving physicians some degree of control could have advantages for motivation and for choosing topics that are most relevant to a physician's practice. However, doing so also presents challenges for having a fair assessment of ability because physicians could game the system by choosing to be assessed primarily on their perceived strengths and not their weaknesses.

Proposals for longitudinal assessment do not change the consequences of failing from those of traditional assessments. However, longitudinal assessments will allow physicians to improve each quarter over the cycle at which there is a consequence (typically every 5 years). Longitudinal and traditional assessments are also the same with respect to authenticity in that both are fairly artificial and differ considerably from clinical practice (e.g., short verbal questions rather than the richness of actually interacting with patients).

Continuing Medical Education (CME)

There is a very extensive body of research on the efficacy of CME. Cervero and Gaines (2015, EL: 2) note 39 systematic reviews of evidence about CME over the period of 1977 until 2015, and they

provide a helpful summary of this field. One overall conclusion is that, in general, CME tends to show small to medium effects on physician knowledge and performance.

A challenge in conceptualizing CME is understanding the range of activities that can sometimes count as CME. Group learning meetings (e.g., courses, conferences, lectures, workshops), online education, videos, reading journal articles or textbooks on one's own, point-of-care learning (e.g., reading online references), and audit and feedback sometimes counts as CME activities. For the purposes of this report, we are going to focus on CME activities that are self-directed (e.g., choosing to attend a lecture, choosing to read an article on one's own). One reason for this position is that most CME activities are in fact self-directed in that the physician can choose the topics to be studied--though, for example, a hospital system may require all medical staff to complete certain online coursework that counts as CME. Another reason is that it cleanly separates CME from other learning opportunities, such as audit and feedback; even though audit and feedback sometimes counts as a CME activity, this is much less common, and audit and feedback has a very different profile in Table 6-2 than typical CME activities.

Another challenge for conceptualizing CME is how to view the use of point-of-care information services, such as looking up reference information to guide decision-making about an individual patient. Even though using online point-of-care references now often counts for CME, we include this as part of patient care since the context and goal is tied directly to decision-making for an individual patient; in contrast, most other CME, such as attending a lecture, is in a separate context outside of direct clinical care.

For the purposes of this report, we consider CME to take place outside of direct clinical care and to be self-directed in that the physician chooses the topics they want to learn about, though these are not always true of activities that count for CME credit. Unlike all the other learning opportunities in Table 6-2, CME activities usually do not involve retrieval practice. For example, in a didactic lecture, or when reading an article, the majority of content is simply presented without testing the learner first and then providing feedback. Of course, sometimes presenters may choose to ask the audience questions, but if this is done it usually comprises a fairly small amount of the total content being covered. Correct information is conveyed to the learner, so even though it is not in the form of feedback after being tested, the learner is still exposed to answers about the content.

With regard to spaced learning, Table 6-2 says "somewhat." The reason is that physicians can choose when to engage in CME activities, so it is possible that they complete many CME activities close to the deadline. On the other hand, given the large numbers of hours of CME requirements, presumably they are often completed in bits over longer stretches of time. As explained above, our definition of CME for the purposes of this report is that it is self-directed, though in reality there are sometimes CME activities that are not self-directed. With regards to consequences, Table 6-2 says "somewhat" because many states require CME to remain licensed, and licensure is a requirement for board certification and for many jobs. However, many CME activities do not test knowledge and simply record that the activity was completed. Therefore, the consequences are tied to the minimal standards for completion, not tied to success. Lastly, most CME is not authentic in that learning takes place outside of clinical care.

Clinical Experience

Physicians' daily experiences with patients, and any accompanying efforts to search for information to guide decision-making about the patient, can serve as a valuable opportunity in many respects. Each patient encounter serves as a retrieval practice experience as a physician retrieves knowledge and practices skills. As a physician has many experiences with patients, it is clearly spaced out over time. Furthermore, it is clearly an authentic experience.

However, there are some other features of personal experience that make it a suboptimal learning opportunity. First, as already discussed throughout this report, the feedback from personal experience is imperfect. Though sometimes a mistake will become apparent later on, often a physician will not know about mistakes that they made.

Second, physicians face many different sorts of consequences in daily practice. The most prevalent consequence is patients' health outcomes. Since physicians are motivated to help patients achieve their health goals, medical errors are associated with a number of subsequent psychological consequences for physicians such as a decrease in quality of life, burnout, and depression (West et al., 2006, EL: 5). Other consequences can include legal action for malpractice. However, since many mistakes are not discovered and therefore there are no consequences, we labeled clinical experience in Table 6-2 as only "somewhat" yielding consequences. Furthermore, the *Improving Diagnosis in Health Care* report (National Academies of Sciences, Engineering, & Medicine, 2015, EL: 6) suggests that guilt, shame, and legal action are likely not productive consequences for learning (see also avoidance-based goals in Chapter 5); instead, this report recommended adopting a non-punitive culture and finding ways to close the feedback loop so that errors are more frequently and quickly discovered.

Lastly, in Table 6-2 we labeled clinical experience as self-directed. For each individual patient, the physician decides whether to make a clinical decision immediately or whether to look up information in online resources or consult with colleagues (Burden et al., 2013; Ely, Osheroff, Chambliss, Ebell, & Rosenbaum, 2005; Cook, Sorensen, & Wilkinson, 2014; Moja & Kwag, 2015). Such decisions are self-directed. The best evidence suggests that higher rates of use of electronic knowledge resources is associated with better knowledge and patient care (Maggio, 2019, EL: 1). Still, physicians make the choice of when to look up information, so it is likely that there are some instances in which they should look up information but do not.

Clinical Decision Support Systems

Clinical decision support (CDS) systems, otherwise known as *best practice alerts* (BPAs), *electronic health record alerts*, or *clinical reminder alerts*, are systems built into the electronic medical record that provide health providers with recommendations and alerts about patient care (e.g., Berner, 2007, 2009; Middleton, Sittig, & Wright, 2016; Musen, Middleton, & Greenes, 2014). Among others, they include reminders that a patient should get a flu shot, prescription alerts about drug-drug interactions, alerts that a patient is starting to deteriorate, suggestions about potential diagnoses. Given the prevalence and diversity of CDS, the total number of high-quality studies eligible to be reviewed in meta-analyses are still fairly modest, and researchers have not specified why some alerts work better than others (Moja et al., 2014; Shojania et al., 2009; Shojania et al., 2010. Due to the ubiquitousness of CDS-generated alerts, there are calls to make alerts and reminders more relevant to avoid alert fatigue (e.g., Embi et al., 2012; Hussain et al., 2019; Kesselheim et al., 2011; Phansalkar et al., 2013). Despite the

challenges of alert fatigue, there are also reasons to believe that CDS alerts have the potential to be helpful for clinicians (Chen et al., 2019, EL: 6; Middleton, Sittig, & Wright, 2016, EL: 6). Here, we are particularly interested in the possibility that CDS alerts may be able to provide valuable learning opportunities for physicians.

CDS is promising as a learning opportunity across many dimensions. Some of these dimensions are directly tied to the fact that they are part of the clinical experience, which is why some of the cells are merged with the Clinical Experience column. CDS systems involve retrieval practice with feedback. Consider a physician prescribing a medicine and receiving an alert about a potential drug-drug interaction. This can be viewed as a type of retrieval practice in the sense that, when entering the prescription, a physician tests their knowledge of whether it is appropriate for this given patient and their other prescriptions. If the alert raises an important drug-drug interaction that the physician did not remember or consider, this could be a useful learning opportunity. Or, they may have already considered this interaction but decided to prescribe it anyways, in which case it still is reinforcing correct knowledge. CDS alerts are spaced in the sense that they occur frequently during patient care, and authentic in that they are embedded in patient care. And, they are not self-directed in that physicians usually cannot turn them on or off. CDS alerts typically do not have any consequences attached to them, aside from the consequence of the patient's health outcomes intrinsic to clinical practice.

One weakness with CDS systems, in terms of providing learning opportunities, is their tendency to provide imperfect feedback. Physicians often override alerts and ignore or reject the suggestion--often for good reasons, such as the alert being generated by incomplete or incorrect patient data, logic that does not perfectly fit the patient, or others (Van Der Sijs et al., 2006; Middleton, Sitting, & Wright, 2016). Thus, for the foreseeable future, CDS systems can only be viewed as suggestions and imperfect feedback rather than authoritative feedback like in longitudinal assessment or CME. This is why feedback is listed in Table 6-2 as only "somewhat" present. Still, it is likely that this sort of feedback can be useful as a learning opportunity (Goodnough et al., 2014, EL: 5; Chen et al., 2015, EL: 5). Indeed, in a large-cluster randomized study that compared the addition of CDS reminders on top of audit and feedback vs. audit and feedback alone for 10 clinical conditions, the physicians who received the point-of-care reminders were more likely to do take the recommended action (e.g., prescribe a drug or vaccine, order a test, perform a screening, encourage smoking cessation) for all 10 conditions, suggesting that CDS systems can be an especially effective form of feedback (Coma et al., 2019, EL: 4).

Feedback based on Audit Results

Audit and feedback is a quality improvement technique in which an individual's performance is measured and compared to a desired professional standard, and then the individual is given feedback about their performance. Two meta-analyses found that audit and feedback tends to produce small but often statistically significant improvements in meeting professional standards. The improvement seems to be larger for healthcare professionals starting out at lower levels of performance and when specific suggestions for improvement are provided (Hysong, 2009, EL: 1; Ivers et al., 2012, EL: 1). However, current views of audit and feedback do not explain why audit and feedback sometimes works better than other times, nor how to design the best audit and feedback systems for particular situations (Gardner, Whittington, McAteer, Eccles, & Michie, 2010; Ivers et al., 2014; Grimshaw et al., 2019).

With regards to our dimensions in Table 6-2, audit and feedback has a very similar profile compared to CDS, though with some differences. Furthermore, because audit and feedback is built on top of clinical experience, many of the cells in Table 6-2 are merged with clinical experience as well as with CDS systems. Audit and feedback involves retrieval practice in the sense that physicians test their knowledge and skills daily in clinical work. Feedback is a core component of audit and feedback; one difference compared to CDS reminders and alerts is that the feedback is delayed and grouped together (e.g., given every month) rather than at the point of service. Learning is spaced over time naturally in clinical practice. Learning is not self-directed in that it is usually the organization, not the individual physician, that decides to implement an audit and feedback program, and usually there is not a way to opt out. For consequences, similar to CDS, audit and feedback typically does not have any consequences aside from the patient's health outcomes, which is intrinsic to clinical practice. A few studies have investigated the role of adding financial incentives on top of audit and feedback, with mixed results (Ivers et al., 2012, EL: 4).

Audit and feedback is similar to CDS alerts on the dimensions that we covered. Both have many desirable features of learning opportunities, though both require additional research into how to make them most clinically effective and least disruptive. Analyzing them from a perspective of how well they promote long-term learning, retention, and behavior change could be helpful in this regard.

Summary of Table 6-2

In summary, our goal with Table 6-2 is not to classify certain learning opportunities as better or worse, but to show how they are different in terms of important dimensions of learning and have different strengths and weaknesses. For example, though there are a number of weaknesses with CME in terms of learning, a strength is that it allows for a very high degree of self-directed learning. A physician who has identified an area of weakness may be willing to devote a lot of time to learning that topic. Collectively, these varied learning opportunities fill different sorts of knowledge gaps. That said, it seems to us that longitudinal assessment fills a similar role as traditional certification assessments in that they both provide retrieval practice and consequences, but longitudinal assessment provides a superior learning opportunity through its use of feedback and spaced learning, and potentially have the learner focus some of the content to their specific practice.

Translatability of Basic Research to Lifelong Learning in Medicine

We have endeavored to report what we view as the best and most relevant evidence out of a much larger body of research. Nevertheless, much of the research comes from basic science studies performed in psychology labs, and a smaller set from more applied research in various settings, such as classrooms. An even smaller minority was conducted in the context of the health professions, and some of these studies involve medical students or nurses in classroom settings rather than practicing physicians. Thus, a vital question is how well this basic research applies to learning among physicians with years of clinical practice.

This is a challenge along multiple dimensions. One dimension is simply that there are major demographic differences in that physicians are older. Though, in theory, this could make a difference, and though we cite evidence of age-related declines in working memory, we do not have specific

reasons to believe that age-related changes interact with evidence such as retrieval practice or spacing. Another potential concern is the setting; perhaps lab and classroom settings are different from a standardized test setting. Again, we do not see theoretical reasons to be concerned that the setting would make a major difference.

Continuing certification involves learning over decades--one's entire working life--whereas almost all the studies cited, except for the few on continuing certification, involve much shorter time frames. Though some of the studies do involve physicians' reasoning about medical topics, many of the studies are about much simpler content that can be learned within the confines of a few hours or days. The degree of knowledge that physicians possess, both in terms of breadth and depth, is much greater than what is typically examined in these studies. Another concern, highly related to the previous points, is that most of this basic research was conducted not with experts but with novices; that is, people learning about material that does not tap into extensive knowledge systems that they have developed over many years. Despite these current limitations, we view these gaps in the literature as exciting opportunities to study basic science phenomena in settings of critical societal importance. For this reason, we believe that many of the studies we proposed would be of interest both to basic science researchers to advance theoretical understanding and to the ABMS member boards for their practical value.

Conclusion

In this report, we evaluated a wide breadth of research related to the development and maintenance of expertise in physicians. We provided evidence for four major themes regarding physician performance: 1) cognitive skills need to be kept current, 2) self-assessment is not enough, 3) testing enhances learning and retention, and 4) goals and consequences motivate. We then created a learning model detailing our understanding of how these complementary themes interact and how they contribute to a physician's knowledge and expertise as related to patient care. We identified a number of practical recommendations for longitudinal assessment programs in medicine. We also identified relevant gaps in research and proposed a number of studies to address these gaps that would provide utility to medical boards. Lastly, we discussed whether other lifelong learning opportunities for physicians meet various psychological considerations believed to benefit learning. Going forward, there is considerable potential for the cognitive and learning sciences to collaborate with medical boards to conduct studies of longitudinal assessment programs that both test ways to improve longitudinal assessment and to advance the basic science of learning.

References

Berner, E. S. (2007). Clinical decision support systems (Vol. 233). New York: Springer Science+ Business Media, LLC.

Berner, E. S. (2009). Clinical decision support systems: state of the art. AHRQ publication, 90069, 1-26.

Blaxton, T. A. (1989). Investigating dissociations among memory measures: Support for a transfer-appropriate processing framework. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15(4), 657.

Burden, M., Sarcone, E., Keniston, A., Statland, B., Taub, J. A., Allyn, R. L., Reid, M.B., Cervantes, L., Frank, M.G., Scaletta, N., Fung, P., Chadaga S.R., Mastalerz, K., Maller, N., Mascolo, M., Zoucha, J., Campbell, J., Maher, M.P., Stella, S.A., & Albert, R. K. (2013). Prospective comparison of curbside versus formal consultations. Journal of hospital medicine, 8(1), 31-35.

Cervero, R. M., & Gaines, J. K. (2015). The impact of CME on physician performance and patient health outcomes: An updated synthesis of systematic reviews. *Journal of Continuing Education in the Health Professions*, *35*(2), 131-138.

Chen, H., Butler, E., Guo, Y., George Jr, T., Modave, F., Gurka, M., & Bian, J. (2019). Facilitation or hindrance: physicians' perception on best practice alerts (BPA) usage in an electronic health record system. Health Communication, 34(9), 942-948.

Chen, J. H., Fang, D. Z., Tim Goodnough, L., Evans, K. H., Lee Porter, M., & Shieh, L. (2015). Why providers transfuse blood products outside recommended guidelines in spite of integrated electronic best practice alerts. Journal of hospital medicine, 10(1), 1-7.

Cook, D. A., Sorensen, K. J., & Wilkinson, J. M. (2014, May). Value and process of curbside consultations in clinical practice: a grounded theory study. In Mayo Clinic Proceedings (Vol. 89, No. 5, pp. 602-614). Elsevier.

Coma, E., Medina, M., Méndez, L. *et al.* (2019).Effectiveness of electronic point-of-care reminders versus monthly feedback to improve adherence to 10 clinical recommendations in primary care: a cluster randomized clinical trial. *BMC Med Inform Decis Mak* 19**,** 245.

Embi, P. J., & Leonard, A. C. (2012). Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. Journal of the American Medical Informatics Association, 19(e1), e145-e148.

Ely, J. W., Osheroff, J. A., Chambliss, M. L., Ebell, M. H., & Rosenbaum, M. E. (2005). Answering physicians' clinical questions: obstacles and potential solutions. Journal of the American Medical Informatics Association, 12(2), 217-224.

Gardner, B., Whittington, C., McAteer, J., Eccles, M. P., & Michie, S. (2010). Using theory to synthesise evidence from behaviour change interventions: the example of audit and feedback. Social science & medicine, 70(10), 1618-1625.

Goodnough, L. T., Shieh, L., Hadhazy, E., Cheng, N., Khari, P., & Maggio, P. (2014). Improved blood utilization using real-time clinical decision support. Transfusion, 54(5), 1358-1365.

Graber, M. L. (2020). Progress understanding diagnosis and diagnostic errors: thoughts at year 10. *7*(3), 151-159.

Grimshaw, J. M., Ivers, N., Linklater, S., Foy, R., Francis, J. J., Gude, W. T., & Hysong, S. J. (2019). Reinvigorating stagnant science: implementation laboratories and a meta-laboratory to efficiently advance the science of audit and feedback. BMJ quality & safety, 28(5), 416-423.

Hussain, M. I., Reynolds, T. L., & Zheng, K. (2019). Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. Journal of the American Medical Informatics Association, 26(10), 1141-1149.

Hysong, S. (2009). Meta-Analysis: Audit and Feedback Features Impact Effectiveness on Care Quality. Medical Care, 47(3), 356–363. https://doi.org/10.1097/MLR.0b013e3181893f6b

Ivers, N., Jamtvedt, G., Flottorp, S., Young, J. M., Odgaard-Jensen, J., French, S. D., O'Brien, M.A., Johansen, M., Grimshaw, J., & Oxman, A. D. (2012). Audit and feedback: effects on professional practice and healthcare outcomes. Cochrane database of systematic reviews, (6).

Ivers, N. M., Sales, A., Colquhoun, H., Michie, S., Foy, R., Francis, J. J., & Grimshaw, J. M. (2014). No more 'business as usual' with audit and feedback interventions: towards an agenda for a reinvigorated intervention. Implementation Science, 9(1), 14.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, *64*(6), 515-526.

Kesselheim, A. S., Cresswell, K., Phansalkar, S., Bates, D. W., & Sheikh, A. (2011). Clinical decision support systems could be modified to reduce 'alert fatigue' while still minimizing the risk of litigation. Health affairs, 30(12), 2310-2317.

Maggio, L. A., Aakre, C. A., Del Fiol, G., Shellum, J., & Cook, D. A. (2019). Impact of Electronic Knowledge Resources on Clinical and Learning Outcomes: Systematic Review and Meta-Analysis. Journal of medical Internet research, 21(7), e13315.

Merkebu, J., Battistone, M., McMains, K., McOwen, K., Witkop, C., Konopasky, A., Torre, D., Holmboe, E. & Durning, S. J. (2020). Situativity: a family of social cognitive theories for understanding clinical reasoning and diagnostic error. *Diagnosis, 7*(3), 169-176.

Middleton, B., Sittig, D. F., & Wright, A. (2016). Clinical decision support: a 25 year retrospective and a 25 year vision. Yearbook of medical informatics, (Suppl 1), S103.

Moja, L., Kwag, K. H., Lytras, T., Bertizzolo, L., Brandt, L., Pecoraro, V., Rigon G., Vaona, A., Ruggiero, F., Mangia , M., Iorio, A., Kunnamo, Il, & Bonovas, S. (2014). Effectiveness of computerized decision support systems linked to electronic health records: a systematic review and meta-analysis. American Journal of Public hHealth, 104(12), e12-e22.

Moja, L., & Kwag, K. H. (2015). Point of care information services: a platform for self-directed continuing medical education for front line decision makers. Postgraduate medical journal, 91(1072), 83-91.

Musen, M. A., Middleton, B., & Greenes, R. A. (2014). Clinical decision-support systems. In Biomedical informatics (pp. 643-674). Springer, London.

National Academies of Sciences, Engineering, and Medicine. (2015). *Improving diagnosis in health care*. National Academies Press.

Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. Medical education, 46(7), 636-647.

Phansalkar, S., Van der Sijs, H., Tucker, A. D., Desai, A. A., Bell, D. S., Teich, J. M., Middleton, B., & Bates, D. W. (2013). Drug—drug interactions that should be non-interruptive in order to reduce alert

fatigue in electronic health records. Journal of the American Medical Informatics Association, 20(3), 489-493.

Rencic, J., Schuwirth, L. W., Gruppen, L. D., & Durning, S. J. (2020a). A situated cognition model for clinical reasoning performance assessment: a narrative review. Diagnosis, 1(ahead-of-print).

Rencic, J., Schuwirth, L. W., Gruppen, L. D., & Durning, S. J. (2020b). Clinical reasoning performance assessment: using situated cognition theory as a conceptual framework. Diagnosis, 7(3), 241-249.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*(6), 1432–1463.

Schiff, G. D. (2008). Minimizing diagnostic error: the importance of follow-up and feedback. The American journal of medicine, 121(5), S38-S42.

Schuwirth, L. W., Durning, S. J., & King, S. M. (2020). Assessment of clinical reasoning: three evolutions of thought. Diagnosis, 7(3), 191-196.

Shojania, K. G., Jennings, A., Mayhew, A., Ramsay, C. R., Eccles, M. P., & Grimshaw, J. (2009). The effects of on-screen, point of care computer reminders on processes and outcomes of care. Cochrane Database of Systematic Reviews, (3).

Shojania, K. G., Jennings, A., Mayhew, A., Ramsay, C. R., Eccles, M. P., & Grimshaw, J. (2010). Effect of point-of-care computer reminders on physician behaviour: a systematic review. *CMAJ*, 182(5), E216-E225.

Swanson, D. B., Norcini, J. J., & Grosso, L. J. (1987). Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education*, *12*(3), 220-246.

Torre, D., Durning, S. J., Rencic, J., Lang, V., Holmboe, E., & Daniel, M. (2020). Widening the lens on teaching and assessing clinical reasoning: from “in the head” to “out in the world”. *Diagnosis*, *7*(3), 181-190.

Van Der Sijs, H., Aarts, J., Vulto, A., & Berg, M. (2006). Overriding of drug safety alerts in computerized physician order entry. *Journal of the American Medical Informatics Association*, 13(2), 138-147.

Yang C, Luo L, Vadillo MA, Yu R, Shanks DR. Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. (2021). *Psychological Bulletin*. 2021 Mar. DOI: 10.1037/bul0000309.

University of Pittsburgh

American Board of Internal Medicine®

American Board of Medical Specialties

Higher standards. Better care.®

American Board of Family Medicine