Rottman *et al.*
Cognitive Research: Principles and Implications (2023) 8:46
https://doi.org/10.1186/s41235-023-00496-9

Cognitive Research: Principles and Implications

REVIEW ARTICLE

Open Access

Cognitive perspectives on maintaining physicians' medical expertise: I. Reimagining Maintenance of Certification to promote lifelong learning

Benjamin M. Rottman[1,2†], Zachary A. Caddick[1,2†], Timothy J. Nokes-Malach[1,2] and Scott H. Fraundorf[1,2*]

Abstract

Until recently, physicians in the USA who were board-certified in a specialty needed to take a summative test every 6–10 years. However, the 24 Member Boards of the American Board of Medical Specialties are in the process of switching toward much more frequent assessments, which we refer to as *longitudinal assessment*. The goal of longitudinal assessments is to provide formative feedback to physicians to help them learn content they do not know as well as serve an evaluation for board certification. We present five articles collectively covering the science behind this change, the likely outcomes, and some open questions. This initial article introduces the context behind this change. This article also discusses various forms of lifelong learning opportunities that can help physicians stay current, including longitudinal assessment, and the pros and cons of each.

Keywords Longitudinal assessment, Medical boards, Feedback, Continuing medical education

Significance statement

Medical Boards assess whether physicians have the knowledge and skills to practice safely and effectively and whether they are keeping up with current medical developments. Switching from using a summative test every 6–10 years to a longitudinal assessment will impact most of the nearly one million physicians in the USA and could have important consequences for the care they provide to the rest of the population. Thus, it is important to carefully consider the likely consequences of such a switch and to identify the factors that can make longitudinal assessment more successful. Furthermore, given that longitudinal assessment is one out of multiple lifelong learning opportunities for physicians, it is important to consider the holes that it might fill and gaps that remain. We review basic research from cognitive psychology as well as applied research that is relevant to answering these questions. More broadly, the idea of switching from high stakes summative tests to lower-stakes formative testing to provide learners with feedback so that they can effectively regulate their own learning is an important topic for all of education, including outside of medicine. Though this review of basic science, inspired by the switch taking place in medicine, is especially relevant to understanding lifelong learning, much of this research comes from shorter timescales, such as learning that takes place within a semester or academic year, and is therefore relevant to education more broadly.

†Benjamin M. Rottman and Zachary A. Caddick contributed equally to this work.

*Correspondence:
Scott H. Fraundorf
scottfraundorf@gmail.com
[1] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA
[2] Department of Psychology, University of Pittsburgh, Pittsburgh, USA

History of medical boards and goals of the project

Over the past several decades, all 24 Member Boards of the American Board of Medical Specialties (ABMS) began time-limited certificates and required physicians who were initially certified by the Boards, known as *Diplomates,* to take and pass an examination every 6–10 years to maintain their certification.[1] Historically, these examinations took the form of point-in-time multiple-choice question assessments taken by Diplomates at secure testing centers, much like the examinations 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.

Certification assessments started to change in 2014 when the American Board of Anesthesiology began pilot work on their Maintenance of Certification in Anesthesiology ("MOCA Minute") program (Sun et al., 2016). In contrast with traditional point-in-time examinations, MOCA Minute was designed as a proactive "assessment *for* learning" (i.e., a formative assessment) in which Diplomates completed a series of questions in one minute taken longitudinally over the course of the year. Participation was intended to assist Diplomates in keeping up with changes in medicine and to promote learning, retention, and application of 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 the testing effect, 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 at the start of the new millennium into a new paradigm of Maintenance of Certification (MOC), emphasizing continuous professional development. Over the past decade, all 24 ABMS Member Boards agreed to develop and migrate toward continuing certification (CC) programs signaling that training and acquisition of medical practice knowledge and skill begin in medical school, are enhanced during residency, and are maintained throughout a specialist's career.

[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.

Shortly after the introduction of MOCA Minute, other ABMS Member Boards began planning for more frequent, lower-stakes assessments as part of their assessment programs, and, 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 are 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), increased relevance of the assessments to Diplomates' practice, and using an "open-book" format so that the questions focus more on reasoning rather than rote memory. At the same time, the programs are diverse in the frequency of 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.

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 research that reviewed the foundational science 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. We—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 research is intended to provide a theoretical framework for continuing assessment of physicians' clinical knowledge. The framework presents a model of the foundational science, and it 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.

We reviewed the foundational science behind longitudinal assessment and arrived at four critical themes: (1) cognitive skills must be kept current; otherwise, they will decline over time, (2) self-assessment is not always enough to reliably and effectively assess one's own competencies or 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

Table 1 Evidence levels for in-text citations for empirical claims

Evidence level	Type of work
1	Quantitative meta-analysis
2	Narrative review
3	Multiple original experiments/randomized controlled trials (RCTs)
4	Single original experiment/RCT
5	Correlational or quasi-experimental study
6	Opinion paper

relation to assessments. These themes are presented in separate articles that accompany the current one.

In our research, we prioritized empirical findings from basic cognitive science and, where available, complementary medical evidence. Additionally, we identified gaps in knowledge and proposed a number of follow-up studies that would be relevant to longitudinal assessment of medical knowledge. 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 (Table 1). The evidence levels range from 1 to 6, with 1 being the strongest evidence (meta-analyses) and 6 being the weakest evidence (opinion papers).

The rest of this article is organized as follows. First, we provide an overview of how the cognitive and learning sciences can inform longitudinal assessment. Second, we discuss some limitations to the work, in particular about how well basic research can be applied to lifelong learning in medicine. Third, we discuss the role of longitudinal assessment in comparison to other lifelong learning mechanisms, such as continuing medical education, clinical experience, and others.

Cognitive perspectives on longitudinal assessment

Figure 1 presents our learning and assessment model of the role of longitudinal assessment in maintaining the quality of physicians' knowledge and expertise. Arrows denote the causal processes or mechanisms that explain the relationships among the variables. Arrows with solid lines represent positive relations (relationships of increase) and arrows with dotted lines represent negative relations (relationships of decrease). The four boxes are used to place the variables into theoretical groupings discussed in each of our other articles. Below, we summarize and synthesize each of these theoretical groupings before we discuss the cross-cutting theme of feedback.

Cognitive skills must be kept current

As physicians get farther and farther out of residency, three processes happen in parallel (Caddick et al., 2022). 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 (Norman et al., 1989, EL: 5). However, it can also be negative insofar as a physician's clinical experience is also 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 incorrect diagnoses or decisions that deviate from standards of care (Choudhry et al., 2006, EL: 5).

Second, over time, physicians also experience cognitive aging. Research on cognitive aging suggests that, as physicians age, 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. The research we reviewed shows that, on average, older physicians do tend to provide poorer quality of care than younger physicians (Choudhry et al., 2005, EL: 2). 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 consequent changes in pattern recognition, and specialization or changes in clinical practice.

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 (Cabana et al., 1999, EL: 2; Cochrane et al., 2007, EL: 2). 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? As we discuss in Fraundorf et al. (2022a), prior research does support the importance of accurately self-assessing one's own skills and abilities (Metcalfe & Finn, 2008, EL: 3; Ohtani & Hisasaka, 2018, EL: 1; Tullis & Benjamin, 2011, EL: 5). Successful self-assessment includes at least two components. *Resolution* is the ability to identify one's relative strengths and weaknesses, such as a physician's areas of expertise (Eva & Regehr, 2011; Regehr et al., 1996). *Calibration* is the ability to evaluate

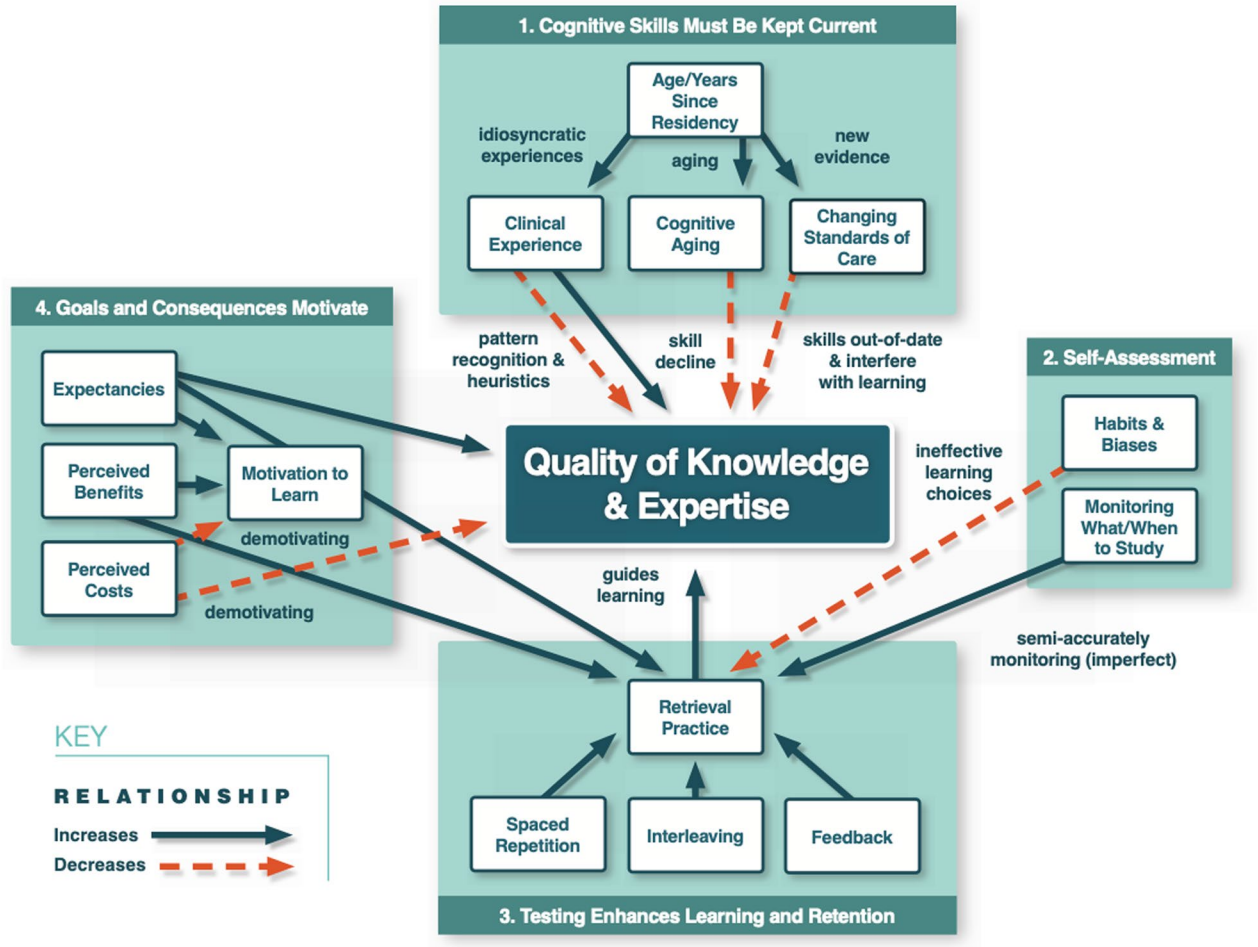

Fig. 1 Synthesis of topics influencing quality of knowledge and expertise

one's overall level of performance, such as whether a physician is overconfident, underconfident, or appropriately confident in their diagnostic and management decisions (Meyer et al., 2013; Podbregar et al., 2001; Zwaan & Hautz, 2019).

However, in self-assessment, individuals do not have direct access to either such component. Instead, they use "informed guesses" which, though somewhat accurate, suffer from systematic biases that are difficult to remove (Koriat, 1995, EL: 5; Koriat, 1997, EL: 5; Schwartz et al., 1997, EL: 2): 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 (Kornell et al., 2011, EL: 3). People also tend to underestimate how much they will forget (Koriat et al., 2004, EL: 3). This implies that physicians may think that they need less continued training to maintain a given level of knowledge than they actually do; indeed, on the whole, physicians tend to be overconfident in their diagnoses (Berner & Graber, 2008, EL: 2).

Further, people tend to avoid many of the learning strategies that are best for long-term retention, such as self-testing, because the sense of difficulty they engender feels—in the moment—like poorer learning (Kirk-Johnson et al., 2019, EL: 3; Yan et al., 2016, EL: 3). Instead, people 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, self-assessment must be supplemented by external sources of assessment, such as continuing certification programs, that can provide a more objective assessment of a physician's knowledge and skills. 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 for testing to serve learning as well as assessment purposes. Although assessments are often viewed as merely tools for decision-making about one's performance level, strong evidence (reviewed in Fraundorf et al., 2022b) indicates that being tested is 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 (Adesope et al., 2017, EL: 1; Rowland, 2014, EL: 1; Yang et al., 2021: EL 1).

Testing is further strengthened when followed by feedback (Rowland, 2014, EL: 1), a phenomenon too often lacking in medical practice itself, and by having tests spaced out over time (Cepeda et al., 2006, EL: 1; Phillips et al., 2019; EL: 2; Pyc & Rawson, 2009, EL: 3). Evidence indicates that greater frequency of testing yields deeper learning (Yang et al., 2021: EL 1). 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 (Brunmair & Richter, 2019, EL: 1), 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 (Ariel & Karpicke, 2017, EL: 4; Einstein et al., 2012, EL: 5; Shaw et al., 2011, EL: 3; Tullis et al., 2013, EL: 4).

Goals and consequences motivate

Testing can also serve as an important motivator (Nokes-Malach et al., 2022). Physicians will be more motivated to study and practice their skills when the perceived benefits of doing so outweigh the perceived costs (Eccles & Wigfield, 2002, 2020; Wigfield & Eccles, 2000; Wigfield et al., 2016). The expectation of specific, challenging assessments can lead people to study longer and more meaningfully (McDaniel et al., 1994, EL: 3; Szpunar et al., 2007, EL: 4); 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 (Bandura, 1997, EL: 2; Honicke & Broadbent, 2016, EL: 1; Pajares, 2008, EL: 2; Schunk & Pajares, 2002, EL: 2).

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 (Harackiewicz & Priniski, 2018, EL: 2; Schiefele et al., 1992, EL: 1). Longitudinal assessment programs would benefit from emphasizing congruence with the physicians' interests (topics and scenarios; Walkington & Bernacki, 2018, EL: 2) and their educational and career goals (e.g., developing expertise and staying current), and by being established as an accurate measure of an important aspect of their knowledge and skills (Guo et al., 2016, EL: 5; Meyer et al., 2019; EL: 5; Putwain et al., 2019, EL: 5; Trautwein et al., 2012, EL: 5).

Decreasing or mitigating the perceived costs of the assessment is also important. More frequent, low-stakes testing may help reduce test anxiety and stereotype threat relative to less frequent, higher-stakes tests (Hinze & Rapp, 2014, EL: 3; Nguyen & Ryan, 2008, EL: 1; Shewach et al., 2019, EL: 1), which in turn can help improve study behaviors and test performance (Ackerman & Heggestad, 1997, EL: 1; Hembree, 1988, EL: 1; Sarason, 1980, EL: 2; von der Embse et al., 2018, EL: 1). 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 (Hidi & Harackiewicz, 2000, EL: 2; Taylor et al., 2014, EL: 1).

A cross-cutting theme: feedback on performance

One cross-cutting theme across this research is the role of feedback. In Caddick et al. (2022), we discuss how accurate and timely feedback is necessary for the development of expertise in any domain. However, the clinical systems 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. In certain cases, feedback could be biased (e.g., a patient avoiding a physician because they were harmed by an error), and the low rates of autopsies in modern medicine have means that errors and misdiagnoses may never be discovered (Shojania et al., 2002, 2003).

In Fraundorf et al. (2022a), we discuss 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. In particular, the poorest performers in a domain are the least accurate in their self-assessments, and they tend to overestimate their knowledge. This overestimation is believed to derive from the same lack of knowledge that caused them to perform poorly in the first place. Poor metacognitive accuracy is particularly problematic in high-stakes environments like medicine if a physician makes incorrect decisions with high confidence. 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 abilities.

We also discuss feedback extensively in our review of the testing effect (Fraundorf et al., 2022b). Though testing improves memory even without feedback of the correct answer, testing with feedback is even more effective. In that work, we also identified a number of open questions regarding precisely how and when to provide feedback.

In Nokes-Malach et al. (2022), we discuss how feedback is critical to several aspects of motivation. Feedback is one important factor in the development of beliefs of self-efficacy. Both positive and negative feedbacks influence one's beliefs of self-efficacy. More generally, longitudinal assessments provide opportunities for individuals both to get multiple pieces of feedback over time and 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 one's goals. For example, to determine 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. Growth mindsets have been hypothesized to be particularly important for situations where one receives negative feedback because mindset influences whether one persists in the face of setbacks. 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.

There has been growing discussion about the lack of feedback in medicine and different ways to begin to implement feedback loops to improve learning and safety (Cifra et al., 2021; Khazen & Schiff, 2021; McGinnis, 2013; National Academies of Sciences, Engineering, & Medicine, 2015; Rosner et al., 2022).

Translating 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. Nevertheless, much of the research comes from basic science studies performed in psychology laboratories and a smaller set from more applied research in various settings, such as classrooms. An even smaller minority was conducted in the context of medicine, and some of these studies involve medical students or nurses in classroom settings rather than practicing clinicians. Thus, a vital question is how well this basic research applies to learning among expert physicians who have years of clinical practice.

This is a challenge in 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—for example—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 laboratory 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. However, there are other dimensions that are potentially more concerning.

One issue is that 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. Another concern has to do with the content. Though some of the studies do involve doctors reasoning about medical topics, many of the studies are about much simpler content that can be taught within the confines of a few hours, or at least within a semester. The raw amount of knowledge that physicians have, in terms of both breadth and depth, is orders of magnitude higher than that in many of 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 but in a setting 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.

Table 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	Y	Y
Feedback	N	Y	Y	S	Y	Y
Spaced	N	Y	S	Y	Y	Y
Self-directed	N	TBD	Y	Y	N	N
Consequences	Y	Y	S	Y	Y	Y
Authentic	N	N	N	Y	Y	Y

Y yes, *N* no, *S* somewhat, *TBD* to be determined

The role of longitudinal assessment in comparison with other lifelong learning mechanisms

The bulk of this paper so far 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 lifelong learning (see Wiese et al., 2022, for a review). 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 2). We first outline six features of learning opportunities that we consider to be important when considering how likely the opportunity would be to lead to learning. Then, we discuss six different learning opportunities for physicians post-residency and for each discuss the learning features that it has and does not have. 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, substantial work (reviewed in Fraundorf et al., 2022b, EL: 2) indicates that testing—*retrieval practice*—can be a powerful learning opportunity in its own right. 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 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 in a cross-cutting section earlier in this article. 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, there is extensive evidence for the benefits of *spaced learning* (Fraundorf et al., 2022b): 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, another cross-cutting topic 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 learners should have *some* degree of control over the topics to be learned; they should not have complete control, nor should they have no control. The exact amounts and type of control are open questions.

Fifth, we have presented people are motivated by *consequences*—by perceived benefits and costs of taking a test and performing well on it (Nokes-Malach et al., 2022). For example, a certain level of arousal is beneficial for learning and performance, but too much is harmful. Furthermore, a sufficiently challenging assessment can facilitate both motivation to learn and 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 paper is whether learning is "authentic" and "naturalistic." Within medical education specifically, and learning sciences more broadly (Barnett & Ceci, 2002, EL: 2; Chen & Klahr, 2008, EL: 2), 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 are 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 et al., 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 et al., 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 (2020b; Rencic et al., 2020a; Schuwirth et al., 2020; Torre et al., 2020). Doing so in a standardized way is obviously a major challenge and currently outside the scope of continuing certification program assessments. Still, the situated cognition model highlights the importance of authentic learning and assessment opportunities.

In Table 2, we classified each cell—whether a particular learning opportunity has a particular feature—as yes, no, or somewhat. However, for many of these cells the answers are more complex, and we discuss them below.

Six lifelong learning opportunities

In this section, we discuss each of the six lifelong learning opportunities in Table 2 and, for each, address each of the six features of learning.

Traditional certification

Traditional certification examinations have a main goal of summative assessment, not learning in and of itself, although studying for the assessment should 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 examination or not, as well as their percentile on the examination, 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 examinations have traditionally been spaced far apart—10 years for many boards—instead of more frequent smaller examinations, they do not capitalize on the benefits of spaced learning. Although some traditional certification examinations allow some degree of customization, such as selecting content specific modules, generally most of the examination is standardized.

Lastly, because traditional certification examinations are largely multiple-choice examinations 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 allows it to serve as a learning opportunity. Another change is that since the assessments happen more frequently, longitudinal assessment capitalizes 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 with each low-stakes assessment 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 count as CME activities. For our present purposes, we focus on CME activities that are self-directed, such as choosing to attend a lecture or 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 count as a CME activity, this is much less common, and audit and feedback have a very different profile in Table 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.

In sum, for our present purposes, 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 of the other learning opportunities in Table 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 even 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.

We describe CME in Table 2 as "somewhat" providing spaced learning. 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. We rate CME as "somewhat" self-directed in Table 2 because many states require CME to remain licensed, and being licensed is a requirement for many jobs and board certification. 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 because a physician retrieves knowledge and practices skills. And because a physician has many experiences with patients, it is clearly spaced out over time. Furthermore, clinical practice is clearly an authentic experience.

However, there are some other features of personal experience that make it a suboptimal learning opportunity. First, as we have already discussed above, the feedback from personal experience is imperfect. Sometimes a mistake will become apparent later, but 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 are 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 rate clinical experience in Table 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; Nokes-Malach et al., 2022). 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 2, we label 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; Cook et al., 2014; Ely et al., 2005; Moja & Kwag, 2015); such decisions are self-directed. The best evidence suggests that higher rates of use of electronic knowledge resources are associated with better knowledge and patient care (Maggio et al., 2019, EL: 1). Still, physicians make the choice of when to look up information, and they often do not seek answers to questions that they have (Ely et al., 1999, EL: 5). Perhaps seeking out answers more frequently could make daily clinical experience more effective as a lifelong learning activity, though of course physicians have limited time in daily encounters to do so.

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 et al., 2016; Musen et al., 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, and suggestions about potential diagnoses. Despite 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, 2010). Due to the ubiquity of CDS generated alerts, there are calls to make alerts and reminders more relevant to avoid alert fatigue (e.g., Embi & Leonard, 2012; Hussain et al., 2019; Kesselheim et al., 2011; Phansalkar et al., 2013).

Despite the challenges of alert fatigue, CDS have the potential to benefit clinicians for several reasons (Chen et al., 2019, EL: 6; Middleton et al., 2016, EL: 6). First, CDS and other forms of technology can help separate tasks that can be done by others in the medical team from those that need to be done by the physician (e.g., Sinsky & Panzer, 2022). For instance, physicians often experience *cognitive load*: Decision-making taxes and sometimes overwhelms the limited capacity of humans to hold information in mind and use it, a capacity that can be further reduced by stress, emotion, and uncertainty (Szulewski et al., 2021; EL: 2). CDS can reduce such load by allowing physicians to *offload* some tasks—that is, to leave them to an external source like the CDS rather than one's own mind (Risko & Gilbert, 2016; EL: 2). Second, CDS may lead to improved patient care even when the presented alert is not learned or remembered by the physician (e.g., a system may recommend the right antibiotic to prescribe, which is beneficial even if the physician does not remember this in the future); indeed, this is the often seen as the primary intended benefit of CDS.

A third possibility, of particular interest given our focus on learning, is that CDS alerts may provide valuable learning opportunities for physicians across many dimensions. Some of these dimensions are directly tied to the fact that they are part of the clinical experience.

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 that the feedback that they provide is often imperfect. 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, EL: 6; Middleton et al., 2016, EL: 6). Thus, for the foreseeable future, CDS systems can only be viewed as suggestions and imperfect feedback rather than authoritative feedback as would occur in longitudinal assessment or CME. Thus, in Table 2, we list as only "somewhat" present in CDS systems. 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 evaluated the addition of CDS reminders on top of audit and feedback relative to audit and feedback alone, physicians who received the point-of-care reminders were more likely to do the recommended task (e.g., prescribe a drug or vaccine, order a test, perform a screening, encourage smoking cessation) for all 10 clinical conditions tested, suggesting that CDS systems can be an especially effective form of feedback (Coma et al., 2019, EL: 4).

Audit and feedback

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. Though initially done in more cumbersome and time-consuming ways, there are newer automated systems (Tsang et al., 2022). Two meta-analyses found that audit and feedback tends to produce small but often statistically reliable improvements in meeting professional standards (Hysong, 2009, EL: 1; Ivers et al., 2012, EL: 1). 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, most research on audit and feedback does 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 et al., 2010; Grimshaw et al., 2019; Ivers et al., 2014). One suggestion is that providing timely feedback on specific actions is likely to be the most helpful (Tsang et al., 2022).

With regard to our dimensions in Table 2, audit and feedback has a very similar profile compared to CDS, though with some differences, because both are built on top of clinical experience.

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).

In sum, 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 they promote long-term learning, retention, and behavior change could be helpful in this regard.

Summary

Our goal with Table 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 therefore 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 able 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 in that they both provide retrieval practice and consequences, but longitudinal assessment provides a superior learning opportunity through its use of feedback, spaced learning, and potentially being somewhat self-directed.

Conclusions

We report results from a project evaluating a wide breadth of research related to the development and maintenance of expertise in physicians. We provided evidence for four major themes regarding physician expert 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 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. Lastly, we discussed whether other lifelong learning opportunities for physicians meet various psychological considerations that are 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 learning within longitudinal assessment and advance the basic science of learning.

Acknowledgements
We thank Andrew Bazemore, Rebecca S. Lipner, David B. Swanson, and Thomas O'Neill for feedback on earlier drafts of this work, especially for providing information on the history of the boards.

Author contributions
BR and ZC wrote the first draft of the manuscript. TN-M and SF provided feedback. All authors contributed to revising the manuscript.

Funding
This work was funded by a grant from the American Board of Internal Medicine (ABIM), American Board of Medical Specialties (ABMS), and American Board of Family Medicine (ABFM). Individuals from ABIM, ABMS, and ABFM provided feedback on the overall goals of the review and on earlier drafts of the manuscript, but approval of the final manuscript rested with the authors alone.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interest
The authors were not involved with the peer review process of this work.

Received: 1 March 2022 Accepted: 20 June 2023
Published online: 24 July 2023

References

Ackerman, P. L., & Heggestad, E. D. (1997). Intelligence, personality, and interests: Evidence for overlapping traits. *Psychological Bulletin, 121*(2), 219–245jh.
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.
Ariel, R., & Karpicke, J. D. (2017). Improving self-regulated learning with a retrieval practice intervention. *Journal of Experimental Psychology: Applied, 24*(1), 43–56.
Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.
Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637.
Berner, E. S. (2007). *Clinical decision support systems* (Vol. 233). Springer.
Berner, E. S. (2009). Clinical decision support systems: State of the art. *AHRQ Publication, 90069*, 1–26.
Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *The American Journal of Medicine, 121*(5 Suppl), S2–S23.
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.
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.
Brown, P. C., Roediger, H. L., III., & McDaniel, M. A. (2014). *Make it stick: The science of successful learning*. Harvard University Press.
Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin, 145*(11), 1029–1052.
Burden, M., Sarcone, E., Keniston, A., Statland, B., Taub, J. A., Allyn, R. L., & Albert, R. K. (2013). Prospective comparison of curbside versus formal consultations. *Journal of Hospital Medicine, 8*(1), 31–35.
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.
Caddick, Z. A., Fraundorf, S. H., Rottman, B. M., & Nokes-Malach, T. J. (2022). Cognitive perspectives on maintaining physicians' medical expertise: II. Acquiring, maintaining, and updating cognitive skills. Manuscript submitted for publication.
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.
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, T., Jr., 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.
Chen, Z., & Klahr, D. (2008). Remote transfer of scientific reasoning and problem-solving strategies in children. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 36, pp. 419–470). Elsevier.
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.
Cifra, C. L., Sittig, D. F., & Singh, H. (2021). Bridging the feedback gap: a socio-technical approach to informing clinicians of patients' subsequent clinical course and outcomes. *BMJ Quality & Safety, 30*(7), 591–597. https://doi.org/10.1136/bmjqs-2020-012464.
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.
Coma, E., Medina, M., Méndez, L., Hermosilla, E., Iglesias, M., Olmos, C., & Calero, S. (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. *Abstract BMC Medical Informatics and Decision Making*, *19*(1). https://doi.org/10.1186/s12911-019-0976-8.
Cook, D. A., Sorensen, K. J., & Wilkinson, J. M. (2014). 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.
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.
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.
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.
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.
Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L., & Evans, E. R. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ, 319*(7206), 358–361.
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.
Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education, 16*(3), 311–329.
Fraundorf, S. H., Caddick, Z. A., Nokes-Malach, T. J., & Rottman, B. M. (2022b). Cognitive perspectives on maintaining physicians' medical expertise: IV. Best practices and open questions in using testing to enhance learning and retention. Manuscript submitted for publication.
Fraundorf, S. H., Caddick, Z. A., Nokes-Malach, T. J., & Rottman, B. M. (2022a). Cognitive perspectives on maintaining physicians' medical expertise: III. Strengths and weaknesses of self-assessment. Manuscript submitted for publication.
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. *Diagnosis, 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.
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.
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.
Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77.
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.
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.
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., & Oxman, A. D. (2012). Audit and feedback: Effects on professional practice and healthcare outcomes. *Cochrane Database of Systematic Reviews, 2012*(6), 1–227.
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.
Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science, 319*(5865), 966–968.
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.
Khazen, M., & Schiff, G. D. (2021). Feedback on missed and delayed diagnosis: Differential diagnosis of communication dilemmas. *The Joint Commission Journal on Quality and Patient Safety, 47*(2), 71–73. https://doi.org/10.1016/j.jcjq.2020.11.011
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. (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., 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.
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.
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.
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.
McGinnis, J. M. (2013). *Best care at lower cost the path to continuously learning health care in America*. National Academies Press Washington D.C.
Merkebu, J., Battistone, M., McMains, K., McOwen, K., Witkop, C., Konopasky, A., & Durning, S. J. (2020). Situativity: A family of social cognitive theories for understanding clinical reasoning and diagnostic error. *Diagnosis, 7*(3), 169–176.
Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*, 174–179.
Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine, 173*(21), 1952–1958.
Meyer, J., Fleckenstein, J., & Koller, O. (2019). Expectancy value interactions and academic achievement: Differential relationships with achievement measures. *Contemporary Educational Psychology, 58*, 58–74.
Middleton, B., Sittig, D. F., & Wright, A. (2016). Clinical decision support: A 25 year retrospective and a 25 year vision. *Yearbook of Medical Informatics, 25*(1), S103.
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.
Moja, L., Kwag, K. H., Lytras, T., Bertizzolo, L., Brandt, L., Pecoraro, V., & Iorio, A. (2014). Effectiveness of computerized decision support systems linked to electronic health records: A systematic review and meta-analysis. *American Journal of Public Health, 104*(12), e12–e22.
Musen, M. A., Middleton, B., & Greenes, R. A. (2014). Clinical decision-support systems. In *Biomedical informatics* (pp. 643–674). London: Springer.
National Academies of Sciences, Engineering, and Medicine. (2015). *Improving diagnosis in health care*. National Academies Press.
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.
Nokes-Malach, T. J., Fraundorf, S. H., Caddick, Z. A.., & Rottman, B. M. (2022). Cognitive perspectives on maintaining physicians' medical expertise: V. Using an expectancy-value framework to understand the benefits and costs of testing. Manuscript submitted for publication.
Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. *Medical Education, 46*(7), 636–647.

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.
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.
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* (pp. 111–139). Lawrence Erlbaum Associates Publishers.
Phansalkar, S., Van der Sijs, H., Tucker, A. D., Desai, A. A., Bell, D. S., Teich, J. M., & 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.
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*(9), 886–902.
Podbregar, M., Voga, G., Krivec, B., Skale, R., Pareznik, R., & Gabrscek, L. (2001). Should we confirm our clinical diagnostic certainty by autopsies? *Intensive Care Medicine, 27*(11), 1750–1755.
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.
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.
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.
Regehr, G., Hodges, B., Tiberius, R., & Lofchy, J. (1996). Measuring self-assessment skills: An innovative relative ranking model. *Academic Medicine, 71*(10), S52–S54.
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, 7*, 227–240.
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.
Risko, E. F., & Gilbert, S. J. (2016). Cognitive of floading. *Trends in Cognitive Sciences, 20*(9), 676–688.
Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.
Rosner, B. I., Zwaan, L., & Olson, A. P. J. (2022). Imagining the future of diagnostic performance feedback. *Abstract Diagnosis, 10*(1), 31–37. https://doi.org/10.1515/dx-2022-0055.
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.
Sarason, I. G. (Ed.). (1980). *Test anxiety: Theory, research, and applications*. Lawrence Erlbaum Associates.
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* (pp. 183–212). Lawrence Erlbaum Associates Inc.
Schiff, G. D. (2008). Minimizing diagnostic error: The importance of follow-up and feedback. *The American Journal of Medicine, 121*(5), S38–S42.
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* (pp. 15–31). Academic Press.
Schuwirth, L. W., Durning, S. J., & King, S. M. (2020). Assessment of clinical reasoning: Three evolutions of thought. *Diagnosis, 7*(3), 191–196.
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.
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.
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.
Shojania, K. G., Burton, E. C., McDonald, K. M., & Goldman, L. (2002). The autopsy as an outcome and performance measure. AHRQ Publication No. 03-E002. Rockville, MD: Agency for Healthcare Research and Quality.
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., Burton, E. C., McDonald, K. M., & Goldman, L. (2003). Changes in rates of autopsy-detected diagnostic errors over time: A systematic review. *JAMA, 289*(21), 2849–2856.
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.
Sinsky, C. A., & Panzer, J. (2022). The solution shop and the production line—The case for a frameshift for physician practices. *New England Journal of Medicine, 386*(26), 2452–2453.
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.
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.
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.
Szulewski, A., Howes, D., van Merriënboer, J. J. G., & Sweller, J. (2021). From theory to practice: The application of cognitive load theory to the practice of medicine. *Academic Medicine, 96*(1), 24–30.
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.
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.
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.
Tsang, J. Y., Peek, N., Buchan, I., van der Veer, S. N., & Brown, B. (2022). Systematic review and narrative synthesis of computerized audit and feedback systems in healthcare. *Journal of the American Medical Informatics Association, 29*(6), 1106–1119.
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*(3), 429–442.
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.
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. (2018). Personalization of instruction: Design dimensions and implications for cognition. *Journal of Experimental Education, 86*(1), 50–68.
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. *JAMA, 296*(9), 1071. https://doi.org/10.1001/jama.296.9.1071
Wiese, A., Galvin, E., Korotchikova, I., & Bennett, D. (2022). Doctors' attitudes to maintenance of professional competence: A scoping review. *Medical Education, 56*(4), 374–386.
Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81.

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.

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.

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435.

Zwaan, L., & Hautz, W. E. (2019). Bridging the gap between uncertainty, confidence and diagnostic accuracy: Calibration is key. *BMJ Quality & Safety, 28*(5), 352–355.

Publisher's Note

Caddick *et al.*
Cognitive Research: Principles and Implications (2023) 8:47
https://doi.org/10.1186/s41235-023-00497-8

Cognitive Research: Principles and Implications

REVIEW ARTICLE

Open Access

Cognitive perspectives on maintaining physicians' medical expertise: II. Acquiring, maintaining, and updating cognitive skills

Zachary A. Caddick[1,2†], Scott H. Fraundorf[1,2*†], Benjamin M. Rottman[1,2†] and Timothy J. Nokes-Malach[1,2]

Abstract

Over the course of training, physicians develop significant knowledge and expertise. We review dual-process theory, the dominant theory in explaining medical decision making: physicians use both heuristics from accumulated experience (System 1) and logical deduction (System 2). We then discuss how the accumulation of System 1 clinical experience can have both positive effects (e.g., quick and accurate pattern recognition) and negative ones (e.g., gaps and biases in knowledge from physicians' idiosyncratic clinical experience). These idiosyncrasies, biases, and knowledge gaps indicate a need for individuals to engage in appropriate training and study to keep these cognitive skills current lest they decline over time. Indeed, we review converging evidence that physicians further out from training tend to perform worse on tests of medical knowledge and provide poorer patient care. This may reflect a variety of factors, such as specialization of a physician's practice, but is likely to stem at least in part from cognitive factors. Acquired knowledge or skills gained may not always be readily accessible to physicians for a number of reasons, including an absence of study, cognitive changes with age, and the presence of other similar knowledge or skills that compete in what is brought to mind. Lastly, we discuss the cognitive challenges of keeping up with standards of care that continuously evolve over time.

Keywords Diagnosis, Expertise, Dual-process theory, Memory, Retrieval failure, Aging

Significance statement

Physicians' expertise and ability to keep up with changing evidence is central for positive patient health outcomes. Here, we begin our review by first evaluating evidence of how expertise is acquired in the medical domain, considering dominant theories about the important benefits and detriments experience has in evaluating new information. We introduce complimentary evidence related to the role memory plays in learning and, since physicians often have careers that spans several decades, the impact that aging and time since initial certification may have on physicians' clinical performance. Importantly, medicine is an ever-evolving field which leads to changes to the standards of care over time. In light of this reality, we discuss how evidence from the cognitive science literature can inform how best to address the challenge of this complex information environment. Given the wide scope of this endeavor, we conclude with proposals to fill in important gaps in knowledge that may continue to propel the field of medicine and physician assessment forward to ensure the highest standards of care possible.

†Zachary A. Caddick, Scott H. Fraundorf and Benjamin M. Rottman have contributed equally to this work.

*Correspondence:
Scott H. Fraundorf
scottfraundorf@gmail.com
[1] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA
[2] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA

Introduction

A primary goal of continuing certification programs is to ensure that board-certified physicians maintain at least a certain minimum level of expertise. In this article, we examine what principles and findings from cognitive science imply about acquiring, maintaining, and updating medical expertise. We first discuss psychological theories of the acquisition of higher-order cognitive medical skills, diagnosis in particular. Next, we discuss how learned skills can be maintained in the face of forgetting and age-related changes in cognition. Lastly, we discuss the process of updating and acquiring new knowledge as medical practice evolves.

We focus on medical decision making and expertise from the traditional information-processing perspective of cognition that undergirds cognitive psychology. We acknowledge that medical decision making is much more complex in that it occurs situated in a dynamic 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 only test a physician's cognitive abilities independently, not their performance in a clinical environment, we focus on the individual physician's cognitive skills.

To situate the strength of the evidence and claims made, we attach evidence levels (EL) to in-text citations for empirical claims (see Table 1). Evidence levels range from 1 to 6, with 1 being the strongest evidence (meta-analyses) and 6 being the weakest (opinion papers).

This article is part of a collection of five articles in this special issue focused on how physicians maintain medical expertise across their careers. These reviews are narrative reviews, not systematic, because they cover a wide variety of topics, not a single narrow topic.

Acquiring medical expertise

Expertise is marked by the acquisition of large amounts of knowledge, which in turn 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 et al., 2017, EL: 5). Rather, experiences play an important role in the development of expertise. For instance, the amount of *deliberate practice*—activities designed to improve targeted aspects of performance—that an individual has completed predicts their level of expertise (Ericsson et al., 1993, EL: 5). It is likely the *quality* of practice, rather than quantity, that is necessary to develop expertise; the mere number of deliberate practice hours on their own does not adequately explain expert performance (Macnamara et al., 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 et al., 1981, EL: 3).

Table 1 Evidence levels for in-text citations for empirical claims

Evidence level	Type of work
1	Quantitative meta-analysis
2	Narrative review
3	Multiple original experiments/randomized controlled trials (RCTs)
4	Single original experiment/RCT
5	Correlational or quasi-experimental study
6	Opinion paper

Dual-process theories in medical decision making

A common theme in the cognitive psychology literature is the existence of two distinct systems for information processing (for overviews, see Evans, 2008; Kahneman & Frederick, 2002; Sloman, 1996). (The term "systems" in this literature refers to 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. An important difference between the systems 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, at times, produces highly accurate decisions and does so quickly and from little information (Marewski & Gigerenzer, 2012).

Within medicine, the dual-process theory is widely accepted as the dominant paradigm for understanding clinical decision making generally, and especially for diagnosis (Croskerry, 2009a, EL: 2; 2009b, EL: 2; Croskerry et al., 2013a, 2013b, EL: 2; Norman & 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 report on Improving Diagnosis in Healthcare (National Academies of Sciences, Engineering, & Medicine, 2015, Chapter 2, EL: 2).

System 2 is understood as the analytical, hypothetico-deductive reasoning or problem-solving 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. This sort of logical reasoning was studied extensively in the earlier years of research on medical decision making (e.g., Elstein et al., 1978; Barrows et al., 1982; Coderre et al., 2003; Newfeld et al., 1981). However, one of the broad conclusions of this research is that changes in hypothetico-deductive problem solving do not appear to explain the transition from novices to experts (e.g., Boshuizen & Schmidt, 1992; Elstein & Schwarz, 2002; Groen & Patel, 1985; Neufeld et al., 1981; Norman, 2005). Instead, it seems that experts use a variety of different forms of knowledge and representations. This is not to say that hypothetico-deductive problem solving is not used by experts; of course, someone with insufficient medical training cannot effectively engage in this sort of reasoning, and of course experts do slow down, ask for second opinions, consult resources, and engage in other forms of careful analytical thinking. Rather, the point is that experts have additional skills and knowledge, one of which is extensive experience with individual patients.

In contrast to System 2, System 1 is the non-analytic decision making in medicine is the very fast pattern recognition process. For instance, pattern recognition allows a physician to quickly think of hyperthyroidism when seeing a patient who is skinny, tremulous, perspiring a lot, and has 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 to order further diagnostic testing, choosing a treatment, or deciding whether to refer to a specialist. Pattern recognition is believed to rely on the same cognitive processes that people use every day, e.g., for categorizing animals as dogs versus cats or for identifying 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 allow decisions to be reached quickly and efficiently. Pattern recognition through categorization can in fact be viewed as one such heuristic (e.g., Nilsson et al., 2008), though heuristics are broader than just pattern recognition (e.g., National Academies of Sciences, Engineering, & Medicine, 2015, Chapter 2; Whelehan et al., 2020). For example, the *representativeness* heuristic 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 et al., 1978; Eddy, 1982; Rottman, 2017). This is appropriate when the two diseases are roughly equally prevalent, but leads to *base rate neglect* when one is more common.

Heuristics are neither inherently bad nor inherently good. They can help physicians make fast decisions, which can be critical in situations with time pressure. And, simple rule-based heuristics 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 in the base rate neglect example, which can lead to suboptimal decisions.

Despite the fact that the dual-process theory is widely accepted as the dominant model of clinical decision making, there are important debates, open questions, and ambiguities with this model. Some researchers question whether there is a clear distinction between the two systems and whether there are only two systems (e.g., De Neys, 2021, 2022; Evans, 2008). Other research has challenged the assumption that they are always opposed to each other and instead may work together (Cushman & Morris, 2015; see Kool et al., 2018 for a review). Though the two systems are often presented as always coming to different decisions, it is likely that they would often come to the same decision in a given situation (De Neys, 2022), which raises a question of how to distinguish which of the two systems is responsible for a given decision.

Perhaps the most pressing question is how people coordinate between the two systems. For example, does some sort of signal of low confidence in System 1 lead to the engagement of System 2, or are both systems engaged simultaneously with a discrepancy-monitoring system noticing when they are coming to different decisions (De Neys, 2022)? These are not just important questions for psychology but are directly related to medicine. Moulton et al. (2007) proposed that being an expert in medical decision making involves "slowing down when you should." This emphasis on "when you should" raises the questions of when an individual should slow down and how this slowing down works. Moulton et al. (see also Croskerry, 2009b) argue that these questions are exactly what the field needs to address to truly understand medical expertise and how to support and train expertise. Stated another way, instead of focusing on fast and slow thinking processes individually, it is more important to understand how experts know that they need to slow down and think a bit harder in certain situations; unfortunately this question is hard to address and there is little existing research.

The role of experience in medical decision making

Here we take a slightly different approach from the dual-process model. This review focuses on the role that experience (interactions with many patients over time) plays

in expertise. Experience with prior patients is at the core of non-analytic System 1: recognizing the pattern in the current case as similar to prior cases. In contrast, relying on rules, evidence, guidelines, and knowledge of pathophysiological and pharmacology rather than one's own idiosyncratic past experience fits with System 2. We do not claim that prior experiences with individual patients cannot be part of the analytical model of decision making and slow deliberative thought. Indeed, of course it is possible that when making a decision about a current patient that a physician may carefully and analytically make comparisons to individual prior patients; in naturalistic decision making it is not possible to know whether a physician is engaged in 'slow' or 'fast' thinking at a given moment (or that only one process occurred). However, to the extent that extensive experience can produce very fast decisions, it is more likely to be involved in System 1 thinking.

The strength of utilizing experience is that it allows for 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, but they also 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 et al., 1978, EL: 5; Gruppen et al., 1988, EL: 5; Pelaccia et al., 2014, EL: 5). This has led to the provocative question by Norman et al. (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 can quickly pattern-match a target case from a large set of prior cases.

However, there are also downsides to relying heavily on experience. Even though pattern recognition and reliance 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 et al., 1999, EL: 4; for other similar studies, see Brooks et al., 1991, EL: 4; Young et al., 2011, EL: 4).

A few studies that shown a similar role of prior experience in real-world medical decision making. One study investigated how often physicians prescribed warfarin for patients with atrial fibrillation in order to prevent a stroke, which despite the risks, is a standard of practice (Choudhry et al., 2006, EL: 5). When one of the physician's patients who was on warfarin experienced a severe bleeding event that was likely a side effect of the warfarin, the physicians were about 20% less likely to prescribe warfarin for patients with atrial fibrillation for at least a year afterward. Another study found that after delivering a baby and experiencing labor and delivery complications, a physician was a bit more likely to use a vaginal delivery instead of cesarean, or vice versa, depending on the prior delivery method (Singh, 2021). In sum, physicians' decisions can be impacted by recent experiences with other patients.

A particular challenge is that these experiences are idiosyncratic. If a physician works in a specialized clinic, they may see certain types of patients 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 were correlated with their probability of providing it as a potential diagnosis (Rottman et al., 2016, EL: 5). In general, this tendency makes sense from the rational Bayesian perspective of diagnosis, in which general "prior" beliefs about the likelihood of diseases in the population are 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 prevalence beliefs are distorted by one's experience, some diagnoses could be overlooked. Additionally, the appearance of patients with rare diseases or rare side effects from treatments (e.g., the bleeding events discussed above) is governed by chance, so physicians may be influenced by the vicissitudes of daily practice. Thus, 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 (e.g., Ericsson, 2015, EL: 2; Hattie & Timperley, 2007, EL: 2; Kahneman & Klein, 2009, EL: 2), and lack of feedback is believed to

contribute to overconfidence (Kahneman & Klein, 2009, EL: 2). However, the medical system is poor at providing feedback (National Academies of Sciences, Engineering, & Medicine, 2015, EL: 6; Schiff, 2008, EL: 6). An error in diagnosis or treatment may never be discovered, in which case feedback is never received. Additionally, 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. For these reasons, two Institute of Medicine reports (McGinnis et al., 2013; National Academies of Sciences, Engineering, & Medicine, 2015; see also Rosner et al., 2023) have called for healthcare organizations to create better feedback systems.

In summary, accumulating experience with many patients is believed to be a critical aspect of how doctors become experts. However, due to working in specialized practices, chance encounters with certain patients but not others, and imperfect feedback systems, doctors will not always experience (and re-experience) certain types of patients. For this reason, one value of a longitudinal continuing certification program is to provide physicians with a well-rounded set of vignettes with feedback to supplement their real-world experiences. We recommend prioritizing gaps that are likely to occur in a physician's practice as well as those that have a meaningful impact on the quality of care delivered.

Maintaining expertise

Cognitive skills decline over time

Once learned, cognitive skills must be maintained: 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). That is, people forget much of what they see and hear in the initial minutes and hours afterward, somewhat more in the following days and weeks, and comparatively little of what remains in the months and years ahead (see Fig. 1). This power-law function may reflect the rate at which people are likely to stop re-encountering those topics (Anderson & Schooler, 1991, EL: 5).

In terms 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

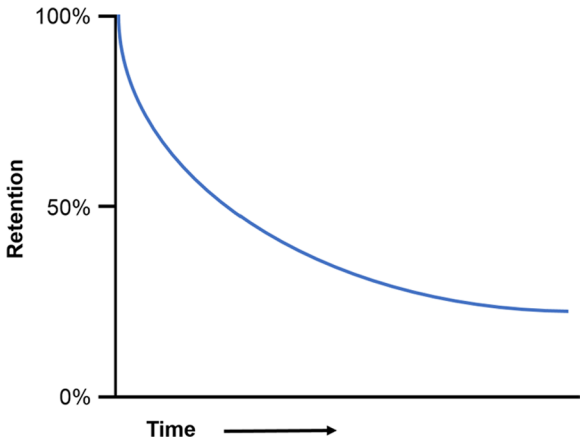

Fig. 1 Prototypical forgetting curve

in training if it is not deliberately practiced. How often important information needs to be practiced is likely to vary across individuals and be influenced by other variables. Nevertheless, the rapid decline in retention after any given study episode suggests that it is beneficial to distribute study over time to alleviate these losses.

Reasons for forgetting

Why do people forget? One intuitive hypothesis might be that we simply run out of mental "storage space" and that prior knowledge is forced out to make room for the new. And, it is indeed 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 learning and knowledge (Drachman, 2005; Landauer, 1986). Laboratory studies have demonstrated that people can readily learn and remember hundreds or even thousands of pictures or sentences even after only seconds of exposure to each (e.g., Shepard, 1967, EL: 4; Standing, 1973, EL: 4; Standing et al., 1970, EL: 4). Indeed, the cortex of the human brain contains approximately 150 trillion (1.5×10^{14}) synapses (Drachman, 2005, EL: 3). That is orders of magnitude more than what the average individual knows: 40,000 (4×10^{4}) words (Brysbaert et al., 2016, EL: 3), 750 (7.5×10^{2}) people (Zheng et al., 2006, EL: 4), or, more generally, approximately 1 billion (10^{9}) bits of information (Landauer, 1986, EL: 2).

Another intuitive hypothesis as to why we forget is that skills and knowledge are lost to *decay*; that is, memories fade and are simply lost over time. This hypothesis receives more support; given the regularities in how memories decline over time (discussed above), it is likely that the passage of time contributes to forgetting

(Wixted, 2004: EL 2; Sadeh et al., 2014: EL 2). However, decay is not likely to be the *whole* story, since not all memories are fated to be lost over time: 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 et al., 1975, EL: 5) even after decades of disuse.

Thus, cognitive psychologists often emphasize *interference* from other, similar information as an additional cause of forgetting. 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*). Thus, categorizing individual exemplars is more difficult (slower) for broader categories (e.g., "cancer" or "plants") than for more narrow ones (e.g., "red-green colorblindness" or "flowers"; Landauer & Freedman, 1968, EL: 3; Landauer & Meyer, 1972, EL: 2; c.f., Collins & Quillian, 1970, EL: 3). 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 et al., 1994, EL: 2; Roediger, 1978, EL: 2). Imagine the process of diagnosing a patient with chest pain. Retrieving *myocardial infarction* in response to the cue *chest pain* reinforces thinking of *myocardial infarction* for future cases, and it also correspondingly weakens the likelihood of considering *aortic dissection* as a diagnosis. Thus, less common concepts and information are particularly vulnerable to interference (Anderson et al., 1994, EL: 3). This suggests it should be particularly important for physicians to practice similar and easily confusable concepts, especially those similar to more common concepts (e.g., common diagnoses) and thereby vulnerable to retrieval-induced forgetting.

It is not always adaptive to retrieve

The phenomenon of retrieval-induced forgetting relates to another key property of human memory: At least in some cases, it is beneficial or adaptive *not* to bring all of one's knowledge to mind (Bjork, 1989, EL: 2; Kuhl et al., 2007, EL: 4; MacLeod, 1998: EL 2; Nørby, 2005, EL: 2; Popov et al., 2019, EL: 3; Wimber et al., 2015, EL: 4). In the case of retrieval-induced forgetting, for instance, it is likely beneficial on the whole to prioritize frequently used facts and concepts over those less frequently used, so as to reduce interference and cognitive demands (Bäuml & Samenieh, 2010, EL 4; Kuhl et al., 2007, EL: 4; Nørby, 2005, EL: 2; Popov et al., 2019, EL: 3; Wimber et al., 2015; EL: 4). For instance, standards of care change (as we discuss below), and it could be beneficial for physicians not to bring to mind outdated standards. Indeed, when explicitly told that some information is obsolete or otherwise should now be forgotten, people can prioritize study and retention of other, to-be-remembered information (for more discussion of the mechanisms of such *directed forgetting*, see MacLeod, 1998; EL: 2; Sahakyan et al., 2013, EL: 2). Forgetting the details of individual episodes or exemplars (e.g., individual patients) can also facilitate learning broader patterns or prototypes (e.g., diagnoses), supporting System 1 pattern recognition system, as discussed above (Nørby, 2005, EL: 2; Posner & Keele, 1968; EL: 3).

Thus, it is unlikely that it would be possible or even desirable to eliminate forgetting completely. Another implication is that is not necessarily advisable for physicians to try to remember every detail of every case, and they are perhaps better served by abstracting more general principles. Lastly, it may be valuable to explicitly highlight when standards of care or other information is out of date so that physicians can leverage directed forgetting to prioritize current, relevant knowledge and skills.

Inaccessible knowledge can often be recovered or relearned

Although people may sometimes be unable to bring to mind the desired 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). A classic example is the *tip-of-the-tongue* phenomenon (e.g., Burke et al., 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. Thus, failure to retrieve an idea at any point in time is not necessarily, or even likely, diagnostic 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 relearned more quickly than 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 helps 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 relates to them or matches how they were initially learned or acquired (Bjork & Richardson-Klavehn, 1989, EL: 2). More frequent longitudinal assessment, rather than point-in-time assessment, could thus serve as a cue to keep this knowledge accessible and/or facilitate relearning.

Effects of age on memory and learning

Beyond the time that has elapsed since medical training, another source of skill decline may be aging. We first cover the basic science of aging and then address studies of aging 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., age 20), 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 et al. 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 et al., 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 et al., 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). Declines in processing speed are relevant to many areas of medicine because physicians often see high volumes of patients in a 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., age 80 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 et al., 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 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 et al., 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., recognizing a familiar set of symptoms as a particular disease) (e.g., Craik, 1986, EL: 2; Hoyer & Verhaeghen, 2005; 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 better remember patients and treatments consistent with their experience.

Several other generalizations regarding memory 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 et al., 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 support each other's memory (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 prioritizing material that is worth more "points" toward a goal (Castel, 2007, EL: 3; Castel et al., 2002, EL: 3; Castel et al., 2007, EL: 3), that a speaker emphasizes intentionally (Fraundorf et al., 2012, EL: 4), that aligns with a motivational bias for positivity (Charles et al., 2003, EL: 3; Mather & Carstensen, 2005, EL: 3; May et al., 2005, EL: 3), or that comes from a more trustworthy source (Rahhal et al., 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, a final generalization is that 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 sum, there is reason to be optimistic that physicians can 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 they may access their knowledge more slowly.

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., Ajmi & Aase, 2021, EL: 2; Eva, 2002, 2003; Durning et al., 2010; Williams, 2006; EL: 2, Council on Medical Education, 2015, EL: 2). Assessing the role of age in a physician's ability to provide high-quality care is quite complicated because a number of factors are so highly correlated that it is usually impossible to distinguish them.

First, it is possible that a physician's abilities decline with age due to memory or processing decline. Second, it is possible that knowledge and skills could decline due to the passage of time out of medical school and residency; this variable is highly confounded with age among physicians insofar as most physicians enter medical school at roughly similar ages. Third, it is possible that as the number of years since residency increases, 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 broader skills that have become less relevant to their practice. In the following paragraphs, we unpack evidence relevant to aging physicians. When we mention correlations with age, we acknowledge that many other factors, explained above, are highly correlated with age. Thus, we are using "age" as a proxy variable, and this is not meant to implicate cognitive aging as the reason for these associations.

Reliable evidence indicates that the quality of healthcare provided decreases with physician age. A systematic review of 62 studies found that 45 studies (73%) reported a decrease in performance for some or all outcomes (Choudhry et al., 2005, EL: 2). Another 13 (21%) found no association. The remaining four (6%) 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 a period of time during which evidence-based medicine and quality-assurance techniques, such as performance evaluation, were becoming adopted. So, it is possible that the apparent age-related declines may instead be driven by the fact that the older physicians were trained prior to this shift and that the newer generation of physicians, who were trained to value evidence-based medicine, may not exhibit declines in quality of care as they age if they stay up to date with the evidence.

Since this systematic review, several notable 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 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 exam 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. (2015, EL: 5) similarly found worse diagnostic performance for clinical vignettes among older physicians. In sum, there is 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 is less certain. One possibility already discussed (see also Eva, 2002, EL: 2), is that performance decline may be caused by negative changes in cognitive processing. Another possibility is that older physicians fail to learn and/or retain changing standards of care. In fact, another study of scores on the ABIM test 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 et al., 2008). However, this study is quite dated, and it is not certain that this finding would still hold among the current cohort of physicians, who participate in different forms of continuing education than physicians 30 years ago. More studies of this nature could help elucidate why knowledge and performance appear to decline with age.

Although performance generally declines with age, there are some cases where it may not. 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 (see also Eva, 2002, EL: 2). Being able to continue to rely on automatic forms of memory accords with the important role of non-analytical medical decision making. Some studies have indeed 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, two studies (Eva et al., 2010, EL: 5; Hobus & Schmidt 1993, EL: 5; see discussion in Eva, 2002) found a positive relation between age/years of experience and diagnostic accuracy. 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, doctors 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. An Institute of Medicine report (McGinnis et al., 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 have systematized the barriers to using current standards of care (Cabana et al., 1999, EL: 2; Cochrane et al., 2007, EL: 2). Though the reviews differ in many ways, there is substantial agreement in terms of the cognitive and attitudinal barriers identified. 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* (believe the treatment or standard would be beneficial) and *agree* with the new standard of care. Fourth, the physician should feel confident they can implement it, termed *self-efficacy*. By contrast, a physician may feel (for example) uncomfortable providing treatment for a condition at the boundary of their scope of practice. Fifth, the physician must overcome *habits* or *inertia*, doing things the same way as they have always been done.

How might a longitudinal assessment program affect these barriers? In Fraundorf et al. (2022), we discuss the overwhelming evidence that repeated testing benefits learning and retention and protects against interference from previously learned practices. Thus, longitudinal assessment would likely improve familiarity, awareness, and knowledge. It is less clear whether longitudinal assessment could impact the attitudinal barriers, such as outcome expectancy and self-efficacy, though it is possible that providing feedback (correct answers, explanations, and citations) might alleviate attitudinal barriers. Aside from longitudinal assessment, continuing medical education (CME) is the primary system currently in place designed to help physicians maintain cognitive skills and gain new skills. In the first article in this

collection (Rottman et al., 2023), we discuss the strength and limitations of CME and compare CME to longitudinal assessment.

Proposed studies and future directions

Providing feedback about strengths and weaknesses and tracking age gaps over time

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. This classification can then be used in three ways.

First, physicians can be provided with feedback about performance on these two different types of questions. This 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 (Day et al., 1988). 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 age differences.

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. 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 involving a new standard, that might indicate that a new standard has not been learned.

Measuring response time during testing

Some prior research has examined the relationships between age, response time, accuracy, and case difficulty (Barrows et al., 1982, EL: 5; Elstein et al., 1978, EL:5; Gruppen et al., 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 this research could benefit from broader case materials and from larger, more representative samples of physicians across many specialties. In fact, given that many items on continuing certification program assessments already incorporate questions about diagnosis and treatment, and response time is easy to record in a computer system, data could be easily obtained to test these relationships.

Identifying out-of-date information

As we discussed above, laboratory evidence indicates that learners have the capability to engage in directed forgetting of material that has been explicitly cued as to-be-forgotten (e.g., because it is out of date or incorrect). There may be opportunities to leverage this capacity in longitudinal assessment; for example, by presenting outdated standards of care and explicitly indicating they are no longer current and should not be retained. We hypothesize that this should lead to better retention of current standards of care than a procedure in which out-of-date material is not explicitly addressed.

Debiasing and clinical reasoning interventions

A more open-ended suggestion is to consider using longitudinal assessment programs as a platform to test debiasing and clinical reasoning interventions. In particular, the dual-process theory presumes that there are two systems of reasoning, one that is faster and one that is slower, and that these systems need to be coordinated (Croskerry, 2009a, EL: 2; 2009b, EL: 2; Norman & 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). Croskerry and colleagues have reviewed potential ways to attempt to teach physicians to avoid common biases (e.g., Croskerry et al., 2013a; b).

Though several debiasing and clinical reasoning interventions have been proposed and tested, there is only mixed evidence whether they work (e.g., Schmidt & Mamede, 2015, EL: 2; see also Isler et al., 2020, for research on debiasing training outside medicine). Still, a number of consenses still express interest in such interventions (Olson et al., 2019; National Academies of Sciences, Engineering, & Medicine, 2015, pp. 4–32 to 4–34; Parodis et al., 2021), and new debiasing interventions are being tested with some promise (Kuhn et al., 2020; Mamede et al., 2020). Longitudinal assessment programs could serve as a testing ground for brief interventions that could be embedded right before or during a question.

In summary, there are a number of potential ways that data collected from longitudinal assessment programs, or from interventions embedded inside the programs, could be used 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.

Examining basic mechanisms of non-analytical reasoning in diagnosis

Longitudinal assessment also provides opportunities to further uncovering the basic mechanisms of non-analytical reasoning in medical decision making. Experiments on the role of non-analytical reasoning in diagnosis can be directly embedded into a longitudinal assessment program and, by doing so, would 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 et al., 1999, EL: 4; Young et al., 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 exist 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 more and larger studies with physicians will help to establish these phenomena within medicine with more certainty.

A longitudinal assessment also provides a remarkable opportunity to design studies embedded into the training 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 (emergency medicine, internal medicine, etc.; Norman et al., 2007).

In sum, 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 increase motivation to follow evidence-based guidelines or be used in debiasing efforts.

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 barriers that a longitudinal assessment program is intended to address are awareness, familiarity, and knowledge about the new standard of care. However, it is possible that 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.

Summary and conclusion

We discussed four topics related to how physicians acquire, maintain, and update cognitive skills. First, we reviewed the dual-process theory of medical expertise, which proposes that medical decision making is a combination of (a) fast intuitive thinking (non-analytical processing) shaped by experience and (b) 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 in which 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 why people forget or fail to retrieve information. Forgetting can sometimes be an adaptive in that it allows people to discard information that is out of date or less important, and forgetting individual details can help people to learn broader patterns. Nevertheless, it is clear that people sometimes fail to bring relevant information to mind. But, this does not necessarily indicate it is permanently lost. Rather, it may become accessible again later, especially with the right 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. The literatures both on interference and on analytical thinking point toward an opportunity for a longitudinal continuing certification program to attempt to fill in potential gaps in experience by testing cases that are somewhat rare but of high importance to patient care.

Third, whereas crystallized intelligence (e.g., medical knowledge) remains intact until age 80, 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; however, another contributor could be a failure to keep up with changing standards of care.

Fourth, we reviewed physician-level barriers to keeping up with changing standards of care, including not being aware of or knowledgeable about a new standard, not believing that the new standard is better, not being confident in how to implement the new standard, and habits. The goal of continuing certification programs has traditionally been to assess whether physicians are keeping up with changing standards. The switch to a longitudinal assessment program presents the opportunity of serving both as assessment and as an educational program to make physicians more knowledgeable about new standards and their application in patient care.

Acknowledgements
We thank Andrew Bazemore, Rebecca S. Lipner, David B. Swanson, and Thomas O'Neill for feedback on earlier drafts of this work.

Author contributions
ZC, SF, and BR wrote the first draft of the manuscript. TN-M provided feedback. All authors contributed to revising the manuscript.

Funding
This work was funded by a grant from the American Board of Internal Medicine (ABIM), American Board of Medical Specialties (ABMS), and American Board of Family Medicine (ABFM). Individuals from ABIM, ABMS, and ABFM provided feedback on the overall goals of the review and on earlier drafts of the manuscript, but approval of the final manuscript rested with the authors alone.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors were not involved with the peer review process of this work.

Received: 1 March 2022 Accepted: 20 June 2023
Published online: 25 July 2023

References

Ajmi, S. C., & Aase, K. (2021). Physicians' clinical experience and its association with healthcare quality: A systematised review. *BMJ Open Quality, 10*(4), e001545.

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

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.

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.

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.

Bäuml, K.-H.T., & Samenieh, A. (2010). The two faces of memory retrieval. *Psychological Science, 21*(6), 793–795.

Bjork, R. A. (1989). Retrieval inhibition as an adaptive mechanism in human memory. In H. L. Roediger & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honour of Endel Tulving* (pp. 309–330). Erlbaum.

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* (pp. 313–344). Lawrence Erlbaum Associates Inc.

Boshuizen, H. P., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediates and novices. *Cognitive Science, 16*(2), 153–184.
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 and 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 and 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.
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.
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.
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.
Coderre, S., Mandin, H. H. P. H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education, 37*(8), 695–703.
Cohen, H., & Lefebvre, C. (Eds.). (2017). *Handbook of categorization in cognitive science* (2nd ed.). Elsevier.
Collins, A. M., & Quillian, M. R. (1970). Does category size affect categorization time? *Journal of Verbal Learning and Verbal Behavior, 9*(4), 432–438.
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). 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. (2013a). Cognitive debiasing 1: Origins of bias and theory of debiasing. *BMJ Quality and Safety, 22*(Suppl 2), ii58–ii64.
Croskerry, P., Singhal, G., & Mamede, S. (2013b). Cognitive debiasing 2: Impediments to and strategies for change. *BMJ Quality and Safety, 22*(Suppl 2), ii65–ii72.
Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Sciences, 112*(45), 13817–13822.
De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on Psychological Science, 16*(6), 1412–1427.
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 inds: 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.
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 University Press.
Elstein, A. S., & Schwarz, A. (2002). Clinical problem solving and diagnostic decision making: Selective review of the cognitive literature. *BMJ, 324*(7339), 729–732.
Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning*. 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.
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.
Groen, G. J., & Patel, V. L. (1985). Medical problem-solving: Some questionable assumptions. *Medical Education, 19*(2), 95–100.

Hargis, M. B., & Castel, A. D. (2018). Younger and older adults' associative memory for medication interactions of varying severity. *Memory, 26*(8), 1151–1158.
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.
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.
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.
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. (2005). Memory aging. In J. Birren & K. Schaie (Eds.), *Handbook of the psychology of aging* (6th ed., pp. 209–232). 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., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and Biases: THe Psychology of Intuitive Judgment, 49*(49–81), 74.
Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526.
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.
Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience, 10*(7), 908–914.
Kuhn, J., van den Berg, P., Mamede, S., Zwaan, L., Diemers, A., Bindels, P., & van Gog, T. (2020). Can we teach reflective reasoning in general-practice training through example-based learning and learning by doing? *Health Professions Education, 6*(4), 506–515.
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.
Landauer, T. K., & Freedman, J. L. (1968). Information retrieval from long-term memory: Category size and recognition time. *Journal of Verbal Learning and Verbal Behavior, 7*(2), 291–295.
Landauer, T. K., & Meyer, D. E. (1972). Category size and semantic-memory retrieval. *Journal of Verbal Learning and Verbal Behavior, 11*(5), 539–549.
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.
Luo, L., & Craik, F. I. (2008). Aging and memory: A cognitive approach. *The Canadian Journal of Psychiatry, 53*, 346–353.
MacLeod, C. (1998). Directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Lawrence Erlbaum Associates Publishers.
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.
Mamede, S., de Carvalho-Filho, M. A., de Faria, R. M. D., Franci, D., Nunes, M. D. P. T., Ribeiro, L. M. C., Biegelmeyer, J., Zwaan, L., & Schmidt, H. G. (2020). 'Immunising' physicians against availability bias in diagnostic reasoning: a randomised controlled experiment. *BMJ Quality and Safety, 29*(7), 550–559.
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.
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.
Neufeld, V., Norman, G., Feightner, J., & Barrows, H. (1981). Clinical problem-solving by medical students: A cross-sectional and longitudinal analysis. *Medical Education, 15*(5), 315–322.
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.
Nørby, S. (2005). Why forget? On the adaptive value of memory loss. *Perspective on Psychological Science, 10*(5), 551–578.
Norman, G. R. (2005). Research in clinical reasoning: Past history and current trends. *Medical Education, 39*(4), 418–427.
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.
Olson, A., Rencic, J., Cosby, K., Rusz, D., Papa, F., Croskerry, P., Zierler, B., Harkless, G., Giuliano, M. A., Schoenbaum, S., & Colford, C. (2019). Competencies for improving diagnosis: An interprofessional framework for education and training in health care. *Diagnosis, 6*(4), 335–341.
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.
Parodis, I., Andersson, L., Durning, S. J., Hege, I., Knez, J., Kononowicz, A. A., Lidskog, M., Petreski, T., Szopa, M., & Edelbring, S. (2021). Clinical reasoning needs to be explicitly addressed in health professions curricula: recommendations from a European consortium. *International Journal of Environmental Research and Public Health, 18*(21), 11202.
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.
Popov, V., Marevic, I., Rummel, J., & Reder, L. M. (2019). Forgetting is a feature, not a bug: Intentionally forgetting some things helps us remember others by freeing up working memory resources. *Psychological Science, 30*(9), 1303–1317.
Posner, M., & I., & Keele, S. K. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353–363.
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 and Cognition, 6*(1), 54–63.
Rosner, B. I., Zwaan, L., & Olson, A. P. (2023). Imagining the future of diagnostic performance feedback. *Diagnosis, 10*(1), 31–37.
Rottman, B. M. (2017). Physician Bayesian updating from personal beliefs about the base rate and likelihood ratio. *Memory and Cognition, 45*(2), 270–280.
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., Caddick, Z. A., Nokes-Malach, T. J., & Fraundorf, S. H. (2023). Cognitive perspectives on maintaining physicians' medical expertise: I. Reimaginging Maintenance of Certification to promote lifelong learning. *Cognitive Research: Principles and Implications* [advance online publication]. https://doi.org/10.1186/s41235-023-00496-9.
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.
Sahakyan, L., Delaney, P. F., Foster, N. L., Abushanab B. (2013). *List-Method Directed Forgetting in Cognitive and Clinical Research*. Elsevier (pp. 131–189)
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.
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.
Singh, M. (2021). Heuristics in the delivery room. *Science, 374*(6565), 324–329.
Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3.
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.
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.
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.
Tullis, J. G., & Benjamin, A. S. (2015). Cueing others' memories. *Memory and 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 and 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.
Wimber, M., Alink, A., Charest, I., Kriegeskorte, N., & Anderson, M. C. (2015). Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature Neuroscience, 18*, 582–589.
Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*, 235–269.
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.
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.
Council on Medical Education. (2015). *Competency and the aging physician* [White paper]. American Medical Association. https://www.cppph.org/wp-content/uploads/2016/02/AMA-Council-on-Medical-Education-Aging-Physician-Report-2015.pdf
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, pp. 139–144).
Ebbinghaus, H. (1885). Ueber das Gedächtnis.
Fraundorf, S. H., Caddick, Z. A., Nokes-Malach, T. J., & Rottman, B. M. (2022). *Cognitive perspectives on maintaining physicians' medical expertise: IV. Best practices and open questions in using testing to enhance learning and retention*. Manuscript submitted for publication.
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, pp. 242–247).
Hobus, P. P. M., & Schmidt, H.G. (1993). In Schmidt, H. G. (Ed.), *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.
Kool, W., Cushman, F. A., & Gershman, S. J. (2018). Competition and cooperation between multiple reinforcement learning systems. *Goal-Directed Decision Making*, 153–178.
Moscovitch, M. (1982). A neuropsychological approach to perception and memory in normal and pathological aging. In *Aging and cognitive processes* (pp. 55–78). Springer.
De Neys, W. (2022). Advancing theorizing about fast-and-slow thinking. *Behavioral and Brain Sciences*, 1–68.
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).

Publisher's Note

Fraundorf *et al.*
Cognitive Research: Principles and Implications (2023) 8:58
https://doi.org/10.1186/s41235-023-00511-z

Cognitive Research: Principles and Implications

REVIEW ARTICLE

Open Access

Cognitive perspectives on maintaining physicians' medical expertise: III. Strengths and weaknesses of self-assessment

Scott H. Fraundorf[1,2*], Zachary A. Caddick[1,2], Timothy J. Nokes-Malach[1,2] and Benjamin M. Rottman[1,2]

Abstract

Is self-assessment enough to keep physicians' cognitive skills—such as diagnosis, treatment, basic biological knowledge, and communicative skills—current? We review the cognitive strengths and weaknesses of self-assessment in the context of maintaining medical expertise. Cognitive science supports the importance of accurately self-assessing one's own skills and abilities, and we review several ways such accuracy can be quantified. However, our review also indicates a broad challenge in self-assessment is that individuals do not have direct access to the strength or quality of their knowledge and instead must infer this from heuristic strategies. These heuristics are reasonably accurate in many circumstances, but they also suffer from systematic biases. 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 is the Dunning–Kruger effect: the poorest performers in a domain are also the least accurate in self-assessment. Further, explicit instruction is not always sufficient to remove these biases. We discuss what these findings imply about when physicians' self-assessment can be useful and when it may be valuable to supplement with outside sources.

Keywords Medical expertise, Metacognition, Self-assessment

Significance statement

Providing high-quality care requires practicing physicians to assess their own knowledge and skills: when judging whether a tentative diagnosis is appropriate, when deciding whether they need to refer a patient to a specialist, or when selecting what skills and materials to study and practice. The present review captures both the strengths and weaknesses of self-assessment, especially as it could be applied to the context of maintaining and updating medical expertise. We show that self-assessment can be reasonably accurate, and we discuss how this could be leveraged in maintaining physicians' medical expertise. However, we also highlight some systematic biases and errors in self-assessment, which point to a need for additional, external sources of feedback and guidance.

Introduction

Physicians' ability to accurately self-assess their knowledge is likely to be critical to multiple aspects of acquiring and retaining expert performance over time, such as deciding what material to study (and how long to study that material) for continuing certification program assessments, deciding among CME options, and deciding whether to look up additional information for making a decision about an individual patient. Self-assessing knowledge is also critical for deciding whether to refer a patient to a sub-specialist versus treating a patient oneself.

*Correspondence:
Scott H. Fraundorf
scottfraundorf@gmail.com
[1] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA
[2] Department of Psychology, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA

Table 1 Evidence levels for in-text citations for empirical claims

Evidence level	Type of work
1	Quantitative meta-analysis
2	Narrative review
3	Multiple original experiments/randomized controlled trials (RCTs)
4	Single original experiment/RCT
5	Correlational or quasi-experimental study
6	Opinion paper

Here, we review what cognitive science suggests about the nature of self-assessment: what it is, why it is important, and how it can be measured. We consider both ways in which self-assessment is accurate as well as its systematic biases and weakness, and we describe theoretical perspectives that account for both. We discuss what may be needed to improve self-assessment before highlighting open questions and proposing relevant future studies.

This article is part of a collection of five articles in this special issue focused on how physicians maintain medical expertise across their careers. We take the approach of a narrative review, not systematic, because it covers a wide variety of topics. To situate the strength of the evidence and claims made, we attach evidence levels (EL) to in-text citations for empirical claims (See Table 1). Evidence levels range from 1 to 6, with 1 being the strongest evidence (meta-analyses) and 6 being the weakest (opinion papers).

What is self-assessment?

The notion of *self-assessment* has been criticized in the literature on medical expertise for being poorly defined (Eva & Regehr, 2005). It is true that self-assessment is a multifaceted construct and can refer to related but distinct processes. We thus begin by introducing the framework of Nelson and Narens (1990, EL: 2), which has been extremely influential within cognitive psychology. This framework identifies two processes relevant to self-assessment. First, people 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, people 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 (Metcalfe & Finn, 2008, EL: 3; Metcalfe, 2009, EL: 2; Thiede et al., 2003, EL: 4). Across domains and participant groups, learners often 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; Miller, 2005, 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 more overall knowledge than learners who spend on their time on easy material (Tullis & Benjamin, 2011, EL: 5; c.f., Nelson & Leonesio, 1988, EL: 5). More broadly, good awareness of one's own thinking (i.e., metacognition) predicts academic success even when controlling for general intelligence (Ohtani & Hisasaka, 2018, EL: 1).

A key implication for the retention 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 what to review in everyday practice. Indeed, physician overconfidence has been linked to diagnostic errors (Berner & Graber, 2008, EL: 2).

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 self-assessments are accurate, then higher confidence should predict a higher probability of correct responding, and lower confidence a lower probability.

Methodologists (e.g.,Juslin et al., 1996; Lichtenstein & Fischhoff, 1977; Murphy, 1973; Nelson, 1996; Nelson & Dunlosky, 1991; Schraw, 2009; Yates, 1982) have delineated how monitoring accuracy can be assessed in terms

of both calibration and resolution. *Calibration* (or *absolute accuracy*) is how well a learner can predict their overall level of performance. For example, if I predict that I will get a B average in my classes this term, do I earn a B average (good calibration), or do I earn an A or C average (poorer calibration)? Calibration identifies whether learners are overconfident, underconfident, or appropriately confident in their skills. Good calibration would be demonstrated if, for instance, a physician who estimated that their initial diagnoses were incorrect 10% of the time was indeed incorrect 10% of the time (rather than more or less). 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 calibration requires learners to provide judgments on a scale that can be directly compared to objective criterion performance. For example, to measure calibration on tests where objective accuracy is measured 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, or smaller intervals such as "0%", "25%", "50%", "75%", or "100%"). This would represent a change for many assessments of medical expertise, where confidence is often assessed using more subjective terms, such as "somewhat confident" or "very confident." Unfortunately, such ratings do not permit a true assessment of whether a learner is overconfident or underconfident because there is no objective definition of what it means to be "somewhat confident." However, there would be several potential advantages to collecting confidence judgments in a format that can assess calibration—most critically, the ability to give physicians feedback on whether they are overconfident or underconfident, as well as asking novel research questions, such as how calibration varies across performance outcomes.

A second type of monitoring accuracy is *resolution* (or *relative accuracy*), which is how well a learner can identify their relative strengths and weaknesses, such as their areas of expertise, or the particular patients for whom their judgments are more or less likely to be corrected. For example, if I think I am more knowledgeable about diabetes than thyroid problems, is that true (good resolution), or am I in fact better with the thyroid than diabetes (bad resolution)? In self-assessing medical expertise, good resolution would be demonstrated physicians expressed more confidence in the specific situations were indeed better at. This kind of monitoring is important when physicians decide which patients need further consideration and when they choose which topics to study for continuing certification program assessments or which CME activities to participate in.

Researchers have debated which form of monitoring is most important for physicians. Some (Omron et al., 2018; Zwaan & Hautz, 2019) have argued that a particular problem for physicians is poor calibration—specifically, overconfidence. Physicians may be overconfident in their skills because even when they make an error (e.g., misdiagnose a patient or provide incorrect treatment), they often does not get adequate feedback about this because the patient may recover anyway, go to another treatment center, or die (see Rottman et al., 2023, for further discussion). Indeed, meta-analysis and review suggest overconfidence is widespread and physicians' self-monitoring is poorly calibrated (Berner & Graber, 2008, EL: 2; Gordon, 1991: EL 1). On the other hand, Eva, Regehr, and colleagues have argued (Regehr et al., 1996; Eva & Regehr, 2005, 2007, 2011) that, in practice, physicians rarely need to assess their overall level of performance or functioning; rather, it is more important to identify the specific cases for which physicians need to slow down and devote more care, a capacity that seems to align with resolution. Our view is that it is likely both calibration ("do I know enough about hypertension?") and resolution ("do I know more about hypertension or diabetes?") would be valuable for physicians, but it is clear more work in this space is needed, especially to directly compare these two capabilities. Indeed, one reason for the lack of clarity on this point may be that not all work has recognized that there are separate measures of metacognitive monitoring that quantify different things.

Metacognitive monitoring can be reasonably accurate

Confidence predicts accuracy

Can learners monitor their learning per both standards discussed above? 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 calibration and discrimination. 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, generally speaking, the more confident someone is in their memory, the more likely it is to be accurate (e.g., Banks, 2000, EL: 5; Benjamin et al., 2009, EL: 5; Egan, 1958, EL: 5; Tweed et al., 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

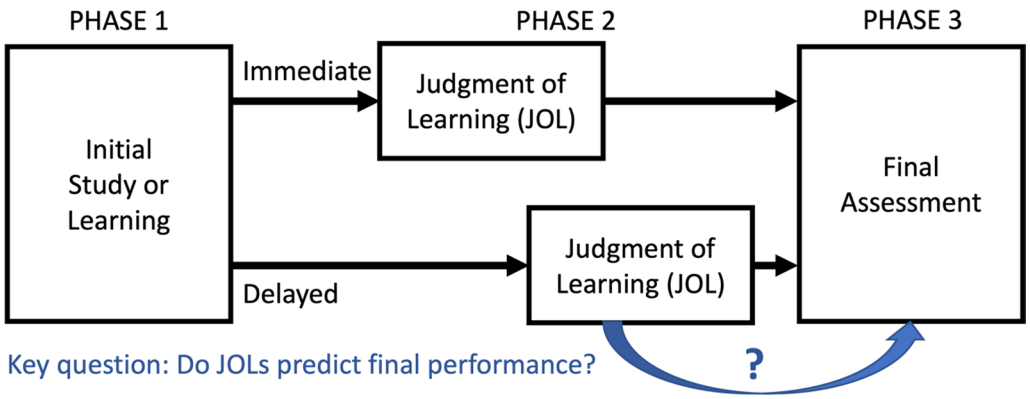

Fig. 1 Schematic design of the typical judgment-of-learning (JOL) study procedure with immediate JOLs (top row) and delayed JOLs (bottom row)

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), as well as particular *categories* of knowledge (e.g., science vs. history, or ankle problems vs. knee problems; Eva & Regehr, 2007, EL: 4). 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, access, and use 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 (Fig. 1). 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), which is 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 depicted in a photograph of a bird. Lastly, the learner takes some form of test or assessment on the material. When JOLs are made at a delay after initial learning, they can strongly predict later performance (Nelson & Dunlosky, 1991, EL: 5); meta-analysis indicates a 0.75 correlation between delayed JOLs and later performance (Rhodes & Tauber, 2011, EL: 1). However, when JOLs are made immediately after learning, their predictive power is somewhat reduced (a correlation of 0.42; Rhodes & Tauber, 2011, EL: 1), for reasons we discuss later.

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 physicians' self-assessments do predict their objective performance, but only weakly to moderately. (Note, however, that these measures did not always distinguish calibration from discrimination.) Indeed, the ability to accurately judge whether one knows something can be challenging in the health sciences: Learners' accuracy in self-assessing their knowledge about healthcare varies widely, but on average is fairly poor (Gordon, 1991, EL: 2), especially for clinical performance as compared to factual knowledge. Where calibration diverges from the ideal, it is often in the direction of physicians being overconfident in their diagnoses, decision-making, and assessments (Berner & Graber, 2008, EL: 2; Gordon, 1991: EL 1).

Thus, depending on one's perspective, the glass of self-assessment is 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, 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. Physicians may be able to choose and practice the particular topics that they struggle with (assuming that the early assessments are fairly low-stakes). 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 pediatric problems or about adult ankles, feet, elbows, shoulders, or spines.

Would such learner control of which materials to study be helpful? Laboratory studies find that learner control of which materials to study is superior to allocating study time equally or based on normative difficulty (Koriat et al., 2006, EL: 3; Mazzoni & Cornoldi, 1993, Experiment 3, EL: 4; Tullis & Benjamin, 2011, EL: 3). However, a meta-analysis of classroom studies (Karich et al., 2014, EL: 1) found weak to nonexistent 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.

People can control reporting in multiple ways

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 and how 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. Indeed, Ward et al. (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 their cognitive skills. Here, we evaluate in turn each of several response strategies: declining to respond, adjusting the grain size of a response, looking up information, and seeking help from others.

Koriat and Goldsmith (1994, EL: 3) developed a two-phase laboratory procedure to test whether people can accurately self-assess whether to respond at all. 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 have the option to withhold responses; payment for participation is structured such that participants lose 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 people 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 et al., 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 chose to skip items that they would have answered incorrectly.

A less drastic adjustment than withholding a response entirely is to provide an estimate or judgment at a different *grain size*. 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). People can also self-assess the appropriate grain size to some degree. The two-phase procedure described above yields similar evidence for effective metacognition 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 et al., 2005, EL: 5; Goldsmith et al., 2002, EL: 5; Koriat et al., 2008, EL: 5; Neisser, 1988; Yaniv & Foster, 1997, EL: 5).

Two other ways that people can adjust their responses are to withhold a response until they can consult an external resource (e.g., the internet; Ferguson et al., 2015, EL: 3) or another person for help. Here, people's behavior may align less closely with their metacognitive monitoring; although people are broadly more likely to consult external aids when less confident (Cotler et al., 1970, EL: 5; Nelson & Fyfe, 2019; EL: 5; Undorf et al., 2021, EL: 3), they sometimes do not seek help even when low in confidence (Undorf et al., 2021, EL: 3). One reason for this may be that seeking external help incurs additional costs, such as requiring more time or—in the case of asking another person—social judgment from one's peers or supervisors (Halabi & Nadler, 2017, EL: 2; Karabenick & Gonida, 2018, EL: 3; Nadler, 1991, EL: 3; Nadler, 2017, EL: 3; Nadler & Chernyak-Hai, 2014, EL: 3, but see Miranda Lery Santos et al., 2020, EL: 4, for null effects of the time taken to request help). Such negative consequences of help-seeking may be particularly strong for individuals from socially disadvantaged groups, for whom help-seeking may be viewed as reinforcing negative stereotypes of inability or dependence (Halabi et al., 2016, EL: 5; Halabi & Nadler, 2017, EL: 2; Nadler, 2017, EL: 3; Nadler & Chernyak-Hai, 2014, EL: 3). However, these conclusions stem from studies with varied forms of help or external resources, and there is a need to study help-seeking behavior with the specific kinds of resources most apt to be used by physicians (e.g., UpToDate).

Nevertheless, the broad need to self-assess when to report versus when to "look it up" 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 to train this metacognitive skill. In the proposed studies below, we describe one method that might be used for such an assessment.

Metacognitive monitoring is subject to systematic biases

Although monitoring can be reasonably accurate in some cases, as we discuss above, research has also documented several important errors and biases in self-assessment. We review several key biases before turning to theoretical accounts that can explain them.

Learners underestimate both learning and forgetting

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 et al., 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 *underconfidence-with-practice effect*; Koriat, 2008b, EL: 3; Koriat et al., 2002; c.f., Serra & Dunlosky, 2005, EL: 3). Even when people do expect their skills to improve, they rely too greatly on their initial experiences in forming expectations: People who are initially the most adept at a task tend to forecast their skills will improve the most (the *performance heuristic*; Critcher & Rosenzweig, 2014, EL: 3), even though in fact such people have the least room to improve.

The tendency for people to treat their present state of skill or 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 may be inflated in the absence of external feedback. Second, they may conversely 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 the observation 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 resolution, but *delayed JOLs* made sometime after later initial learning (e.g., during a second, later study session) predict memory quite accurately. This difference can be explained in terms of the ease-of-processing heuristic we discuss below (Begg et al., 1989, EL: 5). Immediately after studying, knowledge is still active in the learner's immediate working (or short-term) memory[1] 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 et al., 1998, EL: 3). By comparison, what comes to mind sometime after training is much more diagnostic of long-term retention (Begg et al., 1989, EL: 5). An implication for long-term retention is that self-assessments are best performed separately from learning or feedback; 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 long-term expertise.

Learners sometimes evaluate information sources based on superficial fluency

Learners 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 et al., 2013, EL: 3; see also Fiechter et al., 2018, EL: 3).

This bias suggests that fluency of use is important to consider in designing any continuing certification platform. There may be some cases in which *dis*fluency is desirable insofar as it can engender more analytic, "System 2" thinking (e.g., Alter, 2013, EL: 2; Alter et al., 2007, EL: 3; Alter et al., 2013, EL: 6; Diemand-Yauman et al., 2011, EL: 3; Keysar et al., 2012, EL: 3), although this claim has also been disputed (Meyer et al., 2015, EL: 1; Thompson et al., 2013, EL: 3; Yue et al., 2013, EL: 3). However, that may be less relevant to a longitudinal assessment, which is intended for assessment and learning, rather than optimizing in-the-moment decision-making. Thus, all other things being equal, fluency is likely to help create physician buy-in for continuing certification: Physicians

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

will likely perceive that they are learning more if the system presents a fluent, easy-to-use experience.

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 block practice and view this as superior to interleaving (Kirk-Johnson et al., 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). This apparent metacognitive error has been attributed to the fact that blocked practice creates a sense of fluency in the moment even though it is less effective for long-term learning and retention (Kirk-Johnson et al., 2019, EL: 3; Yan et al., 2016, EL: 3).

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 a fill-in-the-blank prompt) 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 thus that learners often mistake the initial effort required by effective study strategies (Schmidt & 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 they are explicitly directed.

Accessing external knowledge may be misperceived as having knowledge

Modern information technology allows physicians—and others—to quickly access external sources of information (e.g., via UpToDate.com). But, several studies have found that accessing information from the internet or other external sources (e.g., books) can create the illusion of internally possessing that knowledge (Eliseev & Marsh, 2023, EL: 3; Fisher et al., 2015, EL: 3; Hamilton & Yao, 2018, EL: 3; Pieschl, 2021, EL: 4; Siler et al., 2022, EL: 3; Ward, 2021, EL: 3), though this finding has not always been replicated (Ferguson et al., 2015, EL: 4). 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 still accurately reflect later performance. Indeed, as we have discussed above, knowing when to consult external resources is an important metacognitive skill, and—-in an era of easily accessible information technology—it may be important to know how and where to locate external information than to memorize it oneself (Marsh & Rajaram, 2019, EL: 2; Sparrow et al., 2011, EL: 3). 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 et al., 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 at studying in the face of other, competing activities (Kurzban et al., 2013, EL: 6). However, it may also reflect errors in self-monitoring insofar as learners do not always recognize when learning can be increased by continuing to study (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, external assessment may potentially be useful for inducing additional, beneficial practice beyond what learners would naturally engage in.

Poor performers overestimate their performance

Another important bias that has been identified in the calibration of metacognitive monitoring is the *Dunning–Krueger effect* (Fig. 2): People with low skill often greatly overestimate their performance (Dunning et al., 2003, EL: 5; Kruger & Dunning, 1999, EL: 4). That is, those who perform poorly in a domain are often unaware they are doing poorly; they are "unskilled and unaware." (By contrast, high performers if anything *underrate* their

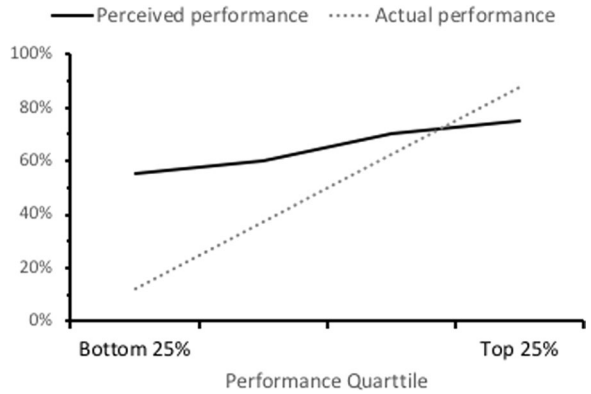

Fig. 2 Prototypical Dunning–Kruger effect (not representing data from any specific study)

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—most critically for our purposes—medicine (Berner & Graber, 2008, EL: 2; Davis et al., 2006, EL: 2; Hodges et al., 2001: EL 5; Parker et al., 2004, EL: 5; Sears et al., 2014, EL: 2).

What causes the Dunning–Kruger effect? In most domains, the knowledge required for effective metacognitive monitoring 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 would use to solve the problem; a student who has learned the wrong rules will both produce the wrong answer *and* be unable to tell that answer is wrong. Low skill thus results in a "double curse" of both inaccurate performance and inaccurate self-assessment. An implication for medical expertise is that physicians low in domain knowledge may be unaware of this fact and unable to correctly self-assess their lack of expertise.

Other factors can influence what learners choose to study

Choices in 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 (Ariel et al., 2009, EL: 3; Ariel et al., 2011, EL: 3; Macaluso et al., 2022, EL: 4).

Thus, while there are advantages to customization, leaving the areas of physicians' continuing study wholly up to physicians (e.g., for CME courses or for a continuing certification program) may be insufficient because physicians in some cases may defer to what they find interesting or what they routinely do rather than where they may need the most continuing education.

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 comes from circumstances—such as very difficult questions for which the most common response is incorrect—that reverse the confidence-accuracy relationship, so 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 et al., 1997, EL: 2; Koriat, 1997, EL: 5): Learners make an "informed guess" about their skill and knowledge based on various heuristics that are often, but not always, correct (Benjamin et al., 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 do often bring to mind more information about material they know well, but it is not guaranteed to be correct.

The inferential nature of metamnemonic monitoring implies that not *all* self-assessment will be accurate and that physicians may benefit from external feedback on their accuracy. Further, while heuristic strategies are often accurate—which is likely why they exist in the first place—there are edge cases where they fail to produce optimal outcomes, which could explain some of the biases discussed above.

In particular, one heuristic that may explain many of the biases reviewed above is what Kornell et al., (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 et al., 1989, EL: 3; Oppenheimer, 2008, EL: 2; see also the closely related heuristic of *easily learned, early remembered*: Koriat, 2008b, EL: 4). Researchers have argued for the prevalence of this heuristic in learners' judgments on the basis 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), 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 may disregard features that *do* matter for retention but that do not enhance immediate fluency (Sungkhasettee et al., 2011, EL: 3), such as the planned number of future study opportunities (Kornell et al., 2011, EL: 3). Not all of these effects necessarily reflect implicit effects of fluency; in some cases, they might reflect learners' explicit beliefs that, for instance, text printed in large type is indeed more memorable (Besken et al., 2019, EL: 3; Mueller et al., 2014, EL: 3; Undorf & Zimdahl, 2019, EL: 3), so it remains an important ongoing debate the extent to which biases stem from an ease-of-processing heuristic versus learners' genuine beliefs (correct or incorrect) about what variables influence learning. Nevertheless, processing fluency has been observed to influence JOLs even in cases where verbalizable beliefs do not have such an influence (Undorf et al., 2017, EL: 3; Yang et al., 2018, EL: 3); indeed, in at least some cases, fluency has been shown to directly mediate effects on JOLs (Undorf et al., 2017, EL: 3; Yang et al., 2018, EL: 3). Therefore, the ease-of-processing heuristic appears to account for at least some, though not all, biases in metacognitive monitoring.

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 (Benjamin et al., 1998, EL: 3; 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. The ease-of-processing heuristic can also explain why information from external sources, like the internet, can be mistaken for personal knowledge: The ability to rapidly access knowledge online can create a feeling of cognitive ease that learners may mistake for genuine understanding. Indeed, 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).

The ease-of-processing heuristic is likely to have implications in clinical settings. As we have reviewed elsewhere (Caddick et al., 2023), physicians are often quite successful in their clinical decision-making. But 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 et al., 2013, EL: 5; Gruppen et al., 1988, EL: 5; Pelaccia et al., 2011, EL: 5), they might not always appropriately judge a wrong answer that also arrives quickly and easily.

Explicit instruction does not remove self-assessment biases

We have 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 these non-analytic judgments would be to simply warn physicians that 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, verbalizable beliefs about which circumstances favor learning and performance, which 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, 2020, 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 et al., 2014a, 2014b, EL: 5), of college students (Hartwig & Dunlosky, 2012, EL: 5; Karpicke et al., 2009, EL: 5; McCabe, 2011, EL: 5; Morehead et al., 2016, EL: 5), and even of college instructors (Morehead et al., 2016, EL: 5). For example, most people describe self-testing only 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; McCabe, 2011, EL: 5; Morehead et al., 2016, EL: 5; Yan et al., 2014a, 2014b, EL: 5); thus, they are unlikely to spontaneously make use of the testing effect. Why do people have such mistaken beliefs about effective learning? One reason may be that they were simply never taught otherwise: About two-thirds of the U.S. population report they never received formal instruction on how best to learn (Yan et al., 2014a, 2014b, EL: 5).

Second, even when learners *do* hold accurate analytic beliefs (e.g., they believe that testing potentiates long-term retrieval), those beliefs are not always activated and *used* in self-assessment. For instance, although presumably all adults understand to some degree that information is forgotten over time, people asked to predict how much they will 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). Similarly, 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., Besken & Mulligan, 2014, EL: 3; Yan et al., 2016, EL: 3).

A key implication for the maintenance of cognitive skills 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. Instead, external prompts for practice and self-assessment may be critical.

Proposed studies and future directions

Response scale for confidence judgments

Currently, physicians' confidence judgments are collected on different scales across various longitudinal assessments. It would be useful to explore the optimal means of assessing confidence. 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 calibration (e.g., overconfidence vs. underconfidence) to be collected and provided as feedback. It would also be useful to determine how many different intervals or categories of confidence can be differentiated by learners—can physicians meaningfully distinguish between, for instance, being "very confident" versus "extremely confident"? This issue is important because, given imprecision in how people translate internal confidence into external ratings (Benjamin et al., 2009: EL 2), a scale with too many categories may in fact decrease the accuracy of confidence ratings (Benjamin et al., 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).

Autonomy and learning outcomes

Given that people can self-assess their knowledge and skills with reasonable accuracy in many situations, it may be of interest to allow physicians some control over the topics they study. We suggest it would be valuable to investigate how greater autonomy in choosing to-be-learned material affects physicians' learning outcomes. Individuals could be randomized to groups with varying degrees of control over the learned content (e.g., 25% control of content vs. 75% control), before both groups' knowledge is tested at a later date. Learning gains could be compared across methods for the chosen material, unchosen material, and overall.

Perhaps it would also make sense to allow physicians to specify which sorts of material they want to study for which reason. For example, they could separately rate which areas are most important for their practice and how confident they are in their knowledge of each area, and the test could then focus on topics that are relevant but for which the physician has lower confidence. Additionally, motivational measures could be assessed to see if increased autonomy leads to increased intrinsic motivation (see Nokes-Malach et al., 2022, for further discussion).

Though there are reasons to hypothesize that autonomy can lead to improved learning—both by increasing motivation and by capitalizing on physicians' knowledge of their areas of weakness—this is not a certainty. In fact, one study on continuing medical education found that quality of care improved only for CME topics that physicians did *not* prefer to learn about, rather than the ones they did (Sibley et al., 1982, EL: 4). In sum, it is important to study if and how autonomy or self-direction over study topics can improve learning; there are reasons to think that it may help, but also reasons to think that it may not.

Physician customization and psychometric quality

Longitudinal assessment has two purposes. First, a longitudinal assessment serves as a summative assessment that Diplomates must pass to maintain their certification. It is critical to establish and maintain the quality of the summative aspects that will be used to make pass–fail decisions. The pass–fail decision is often the hurdle that prevents some Diplomates from remaining certified,

and in those instances, the test publisher will need firm evidence to justify that decision. In particular, making defensible pass–fail decisions is simplified if there is a high degree of standardization so that all examinees attempting to maintain their certification are responsible for similar content mastery reflecting the certificates they hold.

Second, a longitudinal assessment should also provide formative feedback to help Diplomates continue to improve the breadth, depth, and currency of their medical knowledge throughout their career (an "assessment *for* learning"). At times, this second purpose may be at odds with the first. Consider customization that allow each participant to tailor the assessment (in whole or part) to the areas in which they need or wish to improve. This customization may help provide better formative feedback and give Diplomates a greater sense of relevance to their practice. However, customization can sometimes degrade the fit between the measurement and the intended meaning of the certificate.

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 by customization. A few relevant questions include: Is the precision of the participants' scores sufficient to make defensible pass–fail decisions? Are the number of questions scored for summative purposes sufficient to represent the specialty or subspecialty? If questions are being repeated for spaced repetition, are the scores degraded by the lack of independence?

Self-assessment versus self-monitoring

Eva and Regehr (2011, EL: 3) propose a distinction between *self-assessment* at the global level (e.g., "How good a physician am I?") versus *self-monitoring* of specific topic areas (e.g., "How much do I know about hypertension?"). In laboratory studies, they found that college students 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 a longitudinal assessment focuses on. At what level of granularity must these confidence ratings be collected to be accurate? We suggest comparing self-assessment accuracy across different levels of granularity. For instance, physicians 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 accurate self-assessment can be obtained by querying physicians at a more general level or only at the item level.

Objective versus comparative self-assessment

Another dimension on which self-assessments vary 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 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 et al., 2003, EL: 3) and, perhaps as a result, are more responsive to objective than comparative feedback (Moore & Klein, 2008, EL: 3). Nevertheless, it would be useful to collect physicians' self-assessments in both objective and comparative terms to determine which yields more accurate self-assessment.

Do physicians know when to look it up?

In their practice, physicians have the option of deferring judgment to look up information or refer a patient to a specialist. It may thus be useful to evaluate how accurately physicians can judge when they should consult external resources. This could be tested by adapting the Koriat and Goldsmith (1994) procedure discussed above. In a first encounter with each test item, physicians could be given an option to withhold a response; then, in a second pass through each item, physicians would be required to respond. If physicians can correctly identify when they have insufficient knowledge to answer a question on their own, second-pass accuracy should be lower on the questions where physicians withheld an initial response compared to questions where they volunteered one. Further, given potential differences in when people withhold answers entirely versus request help (Undorf et al., 2021, EL: 3), it would be useful to study when physicians choose to consult an external resource and whether these behaviors indeed improve their accuracy.

Determine how to create learner buy-in

Learners' self-assessment of the potential benefits of a longitudinal assessment system is unlikely to be wholly accurate given the biases in self-assessment described above. Further, merely instructing people on desirable learning strategies—such as simply *telling* them that they will learn more from longitudinal assessment—is generally insufficient enough to change beliefs or behavior (McDaniel & Einstein, 2020: EL 2; Yan et al., 2016: EL 3). To guide learners to truly recognize the value of longitudinal assessment and create the most buy-in, more rigorous intervention may be needed to promote accurate

self-assessment (Gordon, 1992: EL 2), such as presenting differentiated feedback on performance under different learning conditions (Benjamin, 2003, EL: 3; Tullis et al., 2013, EL: 3; Yan et al., 2016, EL: 3).

Summary and conclusion

Metacognitive control of learning consists of two processes: (a) monitoring of one's own knowledge and abilities and (b) control of learning and performance strategies. Prior research supports that accurately self-assessing (monitoring) one's own abilities and knowledge is important to guiding (controlling) one's learning and maintaining one's expertise. For instance, self-assessment is associated with the quality of learning strategies an individual employs and consequently their learning outcomes.

Learners do not appear to have direct access to the strength of their skills or knowledge and instead have only an "informed guess." These "informed guesses," although partly accurate, are subject to systematic biases. For example, information or skills that feel 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 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 and retain target information. The tendency to judge learning based on in-the-moment fluency can also lead to choosing suboptimal learning strategies because those strategies feel more fluent at the time of study.

Another notable bias in the self-assessment literature is the Dunning–Kruger effect, the robust finding—including among physicians—that the poorest performers are the least accurate in their self-assessments and tend to overestimate their actual ability. Conversely, the top performers tend to underestimate their ability, though this bias is not as severe.

Although some preliminary evidence suggests that experiencing different learning conditions with feedback might improve self-assessment accuracy, merely instructing learners about the existence of these biases is not enough to remediate them. Instead, externally guided learning for physicians—including in a longitudinal assessment program—is likely to be critical to retaining and updating cognitive skills.

Acknowledgements
We thank Andrew Bazemore, Rebecca S. Lipner, David B. Swanson, and Thomas O'Neill for feedback on earlier drafts of this work.

Author contributions
S.F. wrote the first draft of the manuscript. Z.C., T. N.-M., and B.R. provided feedback. All authors contributed to revising the manuscript.

Funding
This work was funded by a grant from the American Board of Internal Medicine (ABIM), American Board of Medical Specialties (ABMS), and American Board of Family Medicine (ABFM). Individuals from ABIM, ABMS, and ABFM provided feedback on the overall goals of the review and on earlier drafts of the manuscript, but approval of the final manuscript rested with the authors alone.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors were not involved with the peer-review process of this work.

Received: 1 March 2022 Accepted: 9 August 2023
Published online: 30 August 2023

References

Alter, A. L., Oppenheimer, D. M., & Epley, N. (2013). Disfluency prompts analytic thinking—But not always greater accuracy: Response to Thompson et al. (2013). *Cognition*, *128*, 252–255.
Alter, A. L. (2013). The benefits of cognitive disfluency. *Current Directions in Psychological Science, 22*(6), 437–442.
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.
Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569–576.
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.
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. (2003). Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition, 31*(2), 297–305.
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.
Besken, M., Solmaz, E. F., Karaca, M., & Atılgan. (2019). Not all perceptual difficulties lower memory predictions: Testing the perceptual fluency hypothesis with rotated and inverted object images. *Memory & Cognition, 47*, 906–922.
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.
Caddick, Z. A., Fraundorf, S. H., Rottman, B. M., & Nokes-Malach, T. J. (2023). Cognitive perspectives on maintaining physicians' medical expertise: II. Acquiring, maintaining, and updating cognitive skills. *Cognitive Research: Principles & Implications*, *8*, 47.
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.
Cotler, S., Quilty, R. F., & Palmer, R. J. (1970). Measurement of appropriate and unnecessary help-seeking dependent behavior. *Journal of Consulting and Clinical Psychology, 35*(3), 324–327.
Critcher, C. R., & Rosenzweig, E. A. (2014). The performance heuristic: A misguided reliance on past success when predicting prospects for improvement. *Journal of Experimental Psychology: General, 143*(2), 480–485.
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.
Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): Effects of disfluency on educational outcomes. *Cognition, 118*(1), 111–115.
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.
Eliseev, E. D., & Marsh, E. J. (2023). Understanding why searching the internet inflates confidence in explanatory ability [advanced online publication]. *Applied Cognitive Psychology*, 1–10.
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. (2007). Knowing when to look it up: A new conception of self-assessment ability. *Academic Medicine, 82*(10), S81–S84.
Eva, K. W., & Regehr, G. (2011). Exploring the divergence between self-assessment and self-monitoring. *Advances in Health Sciences Education, 16*(3), 311–329.
Ferguson, A. M., McLean, D., & Risko, E. F. (2015). Answers at your fingertips: Access to the Internet influences willingness to answer questions. *Consciousness and Cognition, 37*, 91–102.
Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7*(2), 117–140.
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.
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.
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).
Halabi, S., Dovidio, J. F., & Nadler, A. (2016). Help that hurts? Perceptions of intergroup assistance. *International Journal of Intercultural Relations, 53*, 65–71.
Halabi, S., & Nadler, A. (2017). The intergroup status as helping relations model: Giving, seeking and receiving help as tools to maintain or challenge social inequality. In E. van Leeuwen & H. Zagefka (Eds.), *Intergroup helping* (pp. 205–221). Springer International Publishing.
Hamilton, K. A., & Yao, M. Z. (2018). Blurring boundaries: Effects of device features on metacognitive evaluations. *Computers in Human Behavior, 89*, 213–230.
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.

Karabenick, S. A., & Gonida, E. N. (2018). Academic help seeking as a self-regulated learning strategy: Current issues, future directions. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (pp. 421–433). Routledge/Taylor & Francis Group.

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, H. L., III. (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.

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science, 23*(6), 661–668.

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

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., 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.

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.

Macaluso, J. A., Beuford, R., & Fraundorf, S. H. (2022). Familiar strategies feel fluent: The role of study strategy familiarity in the misinterpreted-effort model of self-regulated learning. *Journal of Intelligence, 10*(4), 83.

Marsh, E. L., & Rajaram, S. (2019). The digital expansion of the mind: Implications of internet usage for memory and cognition. *Journal of Applied Research in Memory and Cognition, 8*(1), 1–14.

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. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*(3), 159–163.

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

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.

Meyer, A., Frederick, S., Burnham, T. C., Guevara Pinto, J. D., Boyer, T. W., Ball, L. J., Pennycook, G., Ackerman, R., Thompson, V., & Schuldt, J. P. (2015). Disfluent fonts don't help people solve math problems. *Journal of Experimental Psychology: General, 144*(2), e16.

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.

Miranda Lery Santos, M., Tricot, A., & Bonnefon, J.-F. (2020). Do learners declining to seek help conform to rational principles? *Thinking & Reasoning, 26*(1), 87–117.

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.
Nadler, A. (1991). Help-seeking behavior: Psychological costs and instrumental benefits. In M. S. Clark (Ed.), *Prosocial behavior* (pp. 290–311). Sage Publications Inc.
Nadler, A. (2017). The human essence in helping relations: Belongingness, independence, and status. In M. van Zomeren & J. F. Dovidio (Eds.), *The Oxford handbook of the human essence* (pp. 123–134). Oxford University Press.
Nadler, A., & Chernyak-Hai, L. (2014). Helping them stay where they are: Status effects on dependency/autonomy-oriented helping. *Journal of Personality and Social Psychology, 106*(1), 58–72.
Neisser, U. (1988). Five kinds of self-knowledge. *Philosophical Psychology, 1*(1), 35–59.
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.
Nelson, L. J., & Fyfe, E. R. (2019). Metacognitive monitoring and help-seeking decisions on mathematical equivalence problems. *Metacognition and Learning, 14*, 167–187.
Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another item, not of the absolute performance on an individual item: Comment on Schraw (1995). *Applied Cognitive Psychology, 10*, 257–260.
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.
Nokes-Malach, T. J., Fraundorf, S. H., Caddick, Z. A., & Rottman, B. M. (2022). *Cognitive perspectives on maintaining physicians' medical expertise: V. Using an expectancy-value framework to understand the benefits and costs of testing*. Manuscript submitted for publication.
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.
Omron, R., Kotwal, S., Garibaldi, B. T., & Newman-Toker, D. E. (2018). The diagnostic performance feedback "calibration gap": Why clinical experience alone is not enough to prevent diagnostic errors. *AEM Education and Training, 2*(4), 339–342.
Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241.
Parker, R. W., Alford, C., & Passmore, C. (2004). Can family medicine residents predict their performance on the in-training examination? *Residency Education, 36*(10), 705–709.
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.
Pieschl, S. (2021). Will using the Internet to answer knowledge questions increase users' overestimation of their own ability or performance? *Media Psychology, 24*(1), 109–135.
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, H. L., III., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.
Rottman, B. M., Caddick, Z. A., Nokes-Malach, T. J., & Fraundorf, S. H. (2023). Cognitive perspectives on maintaining physicians' medical expertise: I. Reimagining maintenance of certification to promote lifelong learning. *Cognitive Research: Principles & Implications*, *8*, 46.
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.
Sibley, J. C., Sackett, D. L., Neufeld, V., Gerrard, B., Rudnick, K. V., & Fraser, W. (1982). A randomized trial of continuing medical education. *New England Journal of Medicine, 306*(9), 511–515.
Siler, J., Hamilton, K. A., & Benjamin, A. S. (2022). Did you look that up? How retrieving from smartphones affects memory for source. *Applied Cognitive Psychology, 36*(4), 738–747.
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.
Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333*(6043), 776–778.
Stone, S. M., & Storm, B. C. (2019). Search fluency as a misleading measure of memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 47*(1), 53–64.
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.
Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73.
Thompson, V. A., Prowse Turner, J. A., Pennycock, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition, 128*, 237–251.
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.
Undorf, M., Livneh, I., & Ackerman, R. (2021). Metacognitive control processes in question answering: Help seeking and withholding answers. *Metacognition and Learning, 16*, 431–458.
Undorf, M., & Zimdahl, M. F. (2019). Metamemory and memory for a wide range of font sizes: What is the contribution of perceptual fluency? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(1), 97–109.
Undorf, M., Zimdahl, M. F., & Bernstein, D. M. (2017). Perceptual fluency contributes to effects of stimulus size on judgments of learning. *Journal of Memory and Language, 92*, 293–304.
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, A. F. (2021). People mistake the internet's knowledge for their own. *Proceedings of the National Academy of Sciences, 118*(43), e2105061118.
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. (2014a). 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. (2014b). Why does guessing incorrectly enhance, rather than impair, retention? *Memory & Cognition, 42*, 1373–1383.
Yang, C., Huang, T.S.-T., & Shanks, D. R. (2018). Perceptual fluency affects judgments of learning: The font size effect. *Journal of Memory and Language, 99*, 99–110.
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.
Yue, C. L., Castel, A. D., & Bjork, R. A. (2013). When disfluency is—and is not—a desirable difficulty: The influence of typeface clarity on metacognitive judgments and memory. *Memory & Cognition, 41*, 229–241.
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.
Zwaan, L., & Hautz, W. E. (2019). Bridging the gap between uncertainty, confidence, and diagnostic accuracy: Calibration is key. *BMJ Quality & Safety, 28*, 352–355.

Publisher's Note

Cognitive Research: Principles and Implications

REVIEW ARTICLE

Open Access

Cognitive perspectives on maintaining physicians' medical expertise: IV. Best practices and open questions in using testing to enhance learning and retention

Scott H. Fraundorf[1,2*], Zachary A. Caddick[1,2], Timothy J. Nokes-Malach[1,2] and Benjamin M. Rottman[1,2]

Abstract

Although tests and assessments—such as those used to maintain a physician's Board certification—are often viewed merely as tools for decision-making about one's performance level, strong evidence now indicates that the experience of being tested is 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, known as the testing effect. We review meta-analytic evidence for the learning benefits of testing, including in the domain of medicine, and discuss theoretical accounts of its mechanism(s). We also review key moderators—including the timing, frequency, order, and format of testing and the content of feedback—and what they indicate about how to most effectively use testing for learning. We also identify open questions for the optimal use of testing, such as the timing of feedback and the sequencing of complex knowledge domains. Lastly, we consider how to facilitate adoption of this powerful study strategy by physicians and other learners.

Keywords Medical expertise, Testing effect, Feedback, Interleaving

Significance statement

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. Indeed, physicians already complete some form of periodic testing in the form of longitudinal assessment for continuing certification. We present evidence that this call is justified insofar as there is robust evidence that the experience of testing can itself be a way to enhance learning and retention. We also discuss what cognitive research implies about how to optimally leverage testing, including longitudinal assessment, as a learning device. Lastly, we discuss how the use case of longitudinal assessment highlights open empirical and theoretical questions regarding the testing effect.

Introduction

Physicians and other healthcare professionals are tasked with acquiring and maintaining multiple forms of knowledge and cognitive skills, including diagnosis, treatment and management, clinical procedures, interpersonal skills, and basic biological and anatomical knowledge. In recent years, there has been a growing call for a greater reliance upon testing as a studying and learning tool in the health professions (Brown, 2017, EL: 6; Cilliers, 2015, EL: 6; Chesluk et al., 2019; Fung et al., 2019, EL: 6; Griffith et al., 2017; EL: 6; Kulasegaram & Rangachari, 2018, EL: 3; Piza et al., 2019; EL: 5; Rapp et al., 2014, EL: 6; Richmond et al. 2019, EL: 6).

*Correspondence:
Scott H. Fraundorf
scottfraundorf@gmail.com
[1] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA
[2] Department of Psychology, University of Pittsburgh, Pittsburgh, USA

Table 1 Evidence levels for in-text citations for empirical claims

Evidence level	Type of work
1	Quantitative meta-analysis
2	Narrative review
3	Multiple original experiments/randomized controlled trials (RCTs)
4	Single original experiment/RCT
5	Correlational or quasi-experimental study
6	Opinion paper

These calls typically promote 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 et al., 2019: EL 2) provides evidence that spaced repetition enhances practicing clinicians' acquisition of knowledge and their clinical behaviors.

Such spaced repetition could be incorporated into the longitudinal assessment programs used in many medical professions. For instance, physicians certified by one of the American Board of Medical Specialties (ABMS) must periodically pass an examination to maintain their certification. Historically, these exams have taken the form of a point-in-time, multiple-choice assessment every six to ten years. More recently, all 24 Boards have announced programs that involve a shift toward more frequent, lower-stakes assessments and test formats that focus on reasoning rather than rote memorization (for further review, see Rottman et al., 2022). One of the primary motivations for this switch is so that these more frequent lower-stakes tests can serve as learning opportunities, rather than just assessment; unlike the older tests, the new longitudinal assessments provide physicians with feedback to promote learning.

In this paper, we examine how such testing can be used to enhance learning and retention of medical expertise. We review the extensive literature on the cognitive benefits of testing on learning and retention. We describe the overall phenomenon as well as how it may be moderated by a number of variables—a key one being feedback—and that may thus constitute best practices for using testing. We consider theoretical explanations for the cognitive mechanisms that underlie the benefits of testing as well whether learners can be trained to employ this helpful learning strategy on their own. Lastly, we consider open questions and future directions in test-enhanced learning. We focus on these principles as they pertain to physicians, as part of a broader collection of five articles in this special issue focused on how physicians maintain medical expertise across their careers, but many of the principles we discuss would also be applicable to maintaining expertise among other healthcare professionals, such as nurses, dentists, or therapists.

This work takes the approach of a narrative review, not systematic, because it covers a wide variety of topics. To situate the strength of the evidence and claims made, we attach evidence levels (EL) to in-text citations for empirical claims (see Table 1). Evidence levels range from 1 to 6, with 1 being the strongest evidence (meta-analyses) and 6 being the weakest (opinion papers).

Overview and basic design

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). The basic testing-effect experiment compares, at a minimum, two groups to which individuals are randomly assigned: a restudy group and a testing group (e.g., Carpenter et al., 2008, EL: 4; Karpicke & Roediger, 2008, EL: 4; Roediger & Karpicke, 2006a, EL: 3, 2006b, EL: 3). The restudy group initially studies information and then has an additional study opportunity later. The testing group initially studies

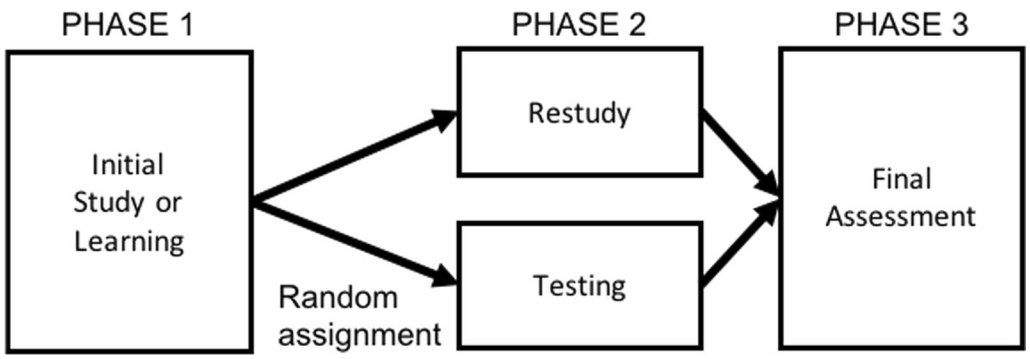

Fig. 1 Schematic design of the typical testing-effect study procedure

information and, instead of restudying 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 Fig. 1). Critically, by comparing testing to restudying for the same period of time, this design controls for the total time that each group spends engaging with the subject matter; as a result, any differences that emerge are driven by testing itself and not by mere re-exposure to information.

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 testing over restudy for long-term retention. 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*. The testing effect 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 et al., 2011, EL: 3; McDermott et al., 2014, EL: 3), mathematics and statistics (Hopkins et al., 2016, EL: 4; Kang et al., 2011a, 2011b, EL: 4; Lyle & Crawford, 2011, EL: 4), geography and maps (Carpenter & Pashler, 2007, EL: 4; Rohrer et al., 2010, EL: 3), psychology (McDaniel et al., 2007, EL: 4; Wiklund-Hörnqvist et al., 2014, EL: 4), and history (Agarwal et al., 2012, EL: 4; Carpenter et al., 2009, EL: 4; McDermott et al., 2014, EL: 3; Nungester & Duchastel, 1982, EL: 4; Roediger et al., 2011, EL: 3). Most critically for our purposes, several experiments have shown benefits of retrieval practice for learning among medical students (LaDisa & Biesboer, 2017: EL 5; Raupach et al., 2016: EL 3) and medical residents (Larsen et al., 2009, EL: 4).

How beneficial is testing? Rowland's (2014, EL: 1) meta-analysis 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 found an even larger Hedges' g of 0.70. Further, retrieval practice better enhances long-term retention and comprehension than some other popular educational techniques, such as concept mapping (Karpicke & Blunt, 2011, EL: 3). A general conclusion, then, is that being tested is likely to be an effective way of enhancing physicians' long-term retention of medical expertise.

Moderators

Researchers have also varied the parameters of the basic testing-effect design presented in Fig. 1 to explore potential moderators of the testing effect, which we now review.

Retention interval and "cramming"

One important characteristic of any learning task is the *retention interval*—the time between initial learning (e.g., reading a document or taking a practice test) and the final assessment. The benefits of testing for retention remain even when assessed 8 to 24 months later (Agarwal et al., 2012, EL: 4; Kerfoot, 2009, EL: 4). In fact, the benefits of testing relative to restudy are intensified with a longer retention interval, 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 et al., 2003, EL: 3; Yeo & Fazio, 2019, EL: 3). For example, Rowland's (2014, EL: 1) meta-analysis found that the difference between testing and restudy was larger when the retention interval was longer than a day (Hedges' $g = 0.69$) than when the retention interval was less than a day (Hedges' $g = 0.41$). Put another way, testing is particularly beneficial when material must be retained for a long time; although cognitive skills decline on the whole over a longer retention interval (Rubin & Wenzel, 1996, EL: 2; Wixted, 2004, EL: 3), this decline is *smaller* with testing relative to restudy.

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

How much testing: frequency, length, 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 suggests 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 et al., 1990, EL: 4; Strong, 1912, EL: 4). However, others have argued that the list-length effect disappears when various confounders are carefully controlled (Dennis & Humphreys, 2001, EL: 3; Dennis et al., 2008, EL: 3; Kinnell & Dennis, 2011, EL: 3), and, at any rate, the *total* amount learned is greater with longer lists (Murayama et al., 2016, EL: 3;

Ward, 2002; EL: 4). In sum, there does not appear to be any *cognitive* reason to avoid longer tests, and this decision can instead be made based on time and motivational constraints.

A related question concerns *how many times* learners should be tested on the same material. The literature suggests that the benefits of multiple tests are nuanced. On the one hand, adding a second test—or even more—does enhance retention above and beyond the first (Roediger & Karpicke, 2006a, 2006b, 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 et al., 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). On the other hand, 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 et al., 2002, EL: 3; Rawson & Dunlosky, 2011, EL: 3; Vaughn & Rawson, 2011, EL: 3; Yang et al., 2021: EL 1), such that additional tests yield diminishing returns. In sum, there is some moderate benefit to continuing to occasionally practice even learned concepts, but many benefits from retrieval practice could be realized with just one test.

Timing of tests: spaced learning

When should learners be tested? Cognitive scientists have extensively studied the broader question of when to schedule learning, whether in the form of restudying or testing. As we discussed above, practicing twice is somewhat better than practicing once (Madigan, 1969, EL: 4). Critically, a second learning session is particularly 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 (Cepeda et al., 2006, EL: 1; Crowder, 1976, EL: 3; Madigan, 1969, EL: 4; c.f., Timer et al., 2020, EL: 4). This effect has been referenced with varying terminology in the literature, including *the spacing effect*, *spaced education*, *spaced training*, and *distributed practice* (Versteeg et al., 2019, EL: 2). 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 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 et al., 2014, EL: 3).

Further, even when using spaced learning, spacing study episodes with longer gaps (*lags*) is generally better than spaced learning with relatively short gaps, which has been termed the *lag effect* (Cepeda et al., 2006, EL: 1; Crowder, 1976, EL: 3; Madigan, 1969, EL: 4; Melton, 1967, EL: 4). 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). However, extremely long lags may be harmful (Cepeda et al., 2009, EL: 3; Cepeda et al., 2008, EL: 3). The optimal lag is likely to depend 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, spacing practice over a long span of time—such as through longitudinal assessment—is likely to result in the most enduring medical knowledge.

Interventions in the health sciences have sometimes combined the testing effect and spaced learning by having learners answer test questions periodically over time, a practice often termed *spaced repetition*. Systematic review (Phillips et al., 2019: EL 2) indicates that spaced repetition enhances healthcare professionals' acquisition of knowledge and their clinical behaviors (as measured both via self-report and objective records). Further, spaced repetition activities generally meet with acceptance and uptake; in the studies reviewed by Phillips et al., 87% of participants in spaced-repetition interventions indicate they would participate in future spaced-repetition activities, and completion rates were high. Not all of the studies of healthcare professionals reviewed by Phillips et al. (2019) involved physicians (e.g., some involved nurses), and only some used randomized controlled trials with experimental designs, indicating a need for more high-quality studies specifically with physicians. Nevertheless, Phillips et al. (2019) concluded that spaced repetition is "one of the few evidence-based pedagogies that

can increase knowledge, promote retention of knowledge […] and positively affect clinical practice" (p. 899).

Test format and type of knowledge

Physicians are tasked with acquiring and maintaining several different types of knowledge, such as basic factual knowledge, diagnosis and classification, medical procedures and clinical behaviors. Could testing enhance retention of each of these?

In general, testing indeed appears to be effective across many testing formats and types of knowledge. 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, or 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, or answering a question asked by a patient; Adesope et al., 2017, EL: 1; c.f., Hinze & Wiley, 2011, EL: 4; Rowland, 2014, EL: 1), 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). Adesope et al., (2017, EL: 1) formally examined test format in their meta-analysis and found a significant benefit of testing over restudy for all test formats. For this reason, the specific format of a test item is likely of less importance than the presentational quality of the question (e.g., clarity, readability, and veracity of text).

Some controversy has existed as to whether testing benefits more complex knowledge types and tasks, such as problem-solving (c.f., Karpicke & Aue, 2015, EL: 6; Leahy et al., 2015, EL: 4; Rawson, 2015, EL: 2; van Gog & Kester, 2012, EL: 4; van Gog et al., 2015, EL: 4; van Gog & Sweller, 2015, EL: 3). However, meta-analytic evidence suggests testing does benefit complex problem-solving tasks and other types of high-level conceptual knowledge (Yang et al., 2021: EL 1), and several studies have found benefits of testing specifically for clinical behaviors and skills (Kromann et al., 2009, EL: 4; Larsen et al., 2009, EL: 4; Raupach et al., 2016, EL: 3). Another finding 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 et al., 2010, EL: 4; Siler & Benjamin, 2019, EL: 3), somewhat analogous to diagnosing or classifying patients.

In sum, 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.

Ordering of practice material

Given that the content to longitudinal assessments generally includes multiple concepts and items, a natural question is whether there are better or worse ways to order such material. The optimal ordering of learning material has frequently studied in cognitive psychology, although not always in the specific context of the testing effect. Cognitive psychologists who have studied this issue more broadly 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; Firth et al., 2021, EL: 1) suggests that, for most materials, interleaving practice results in superior learning than blocked practice, with a medium effect size.

Of course, various intermediate schedules are also possible, such as beginning with blocked practice and then transitioning to interleaved (Yan et al., 2017, EL: 5). Preliminary evidence suggests that an intermediate degree of interleaving is optimal in more complex domains, such as when topics are arranged 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 et al., 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 two potentially confusable patient presentations (e.g., shortness of breath could reflect heart problems or lung problems) requires understanding 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 could both be used, but especially why there is a reason to choose one over the other. Thus, one recommendation is that, when there is a concern that two diagnoses or two treatments may be confused (i.e., they may be subject to interference; Caddick et al., 2022), it would likely be beneficial

to interleave those concepts together on the *same* assessment rather than blocked into different assessments. One caveat is that much of the work on the interleaving benefit has not been specific to retrieval practice (e.g., in some studies, learners merely viewed the exemplars without being tested) and it would be useful to confirm the benefits of interleaving specifically in the context of retrieval practice.

One other line of work has explored item sequencing specifically in the context of test items and their difficulty. This work is grounded in the more general principle of the *peak-end rule*: people tend to judge experiences primarily as a function of (a) the affective peak (i.e., the strongest positive or negative experience) and (b) their ending experience (Diener et al., 2001, EL: 3; Do et al., 2008, EL: 3; Kahneman et al., 1993, EL: 4). In line with this, laboratory studies have shown that adding easier items, which are likely to engender a positive experience of success, to the end of a test increases learners' willingness to engage in future testing, even when the additional items extend the overall length of the test (Cho, 2021, EL: 4; Finn & Miele, 2016, EL: 3; Finn & Miele, 2021, EL: 3; O'Day, 2022, EL: 3). Consequently, we suggest there may be potential value in ending each longitudinal assessment "on a high note" with a few relatively easy items that are likely to encourage continued participation in the program.

Transfer to untested material

Evidence suggests that retrieval practice can support *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). It is generally rare—if not impossible—to observe *far transfer*, where training or practice in one domain also confers benefits to other, wholly unrelated domains (Sala & Gobet, 2017, EL: 1). However, the learning benefits of the testing effect do appear to transfer to more closely related material (*near transfer*). For example, Kang et al., (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 restudy 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 et al., (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 exam 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*. Similarly, the benefits of being tested on part of a science text can sometimes generalize to other, related facts from the text (Chan, 2010 EL: 4; Chan et al., 2006, EL: 4), though this has not been observed in all studies (Pan & Rickard, 2018, EL: 1; Woolridge et al., 2014, EL: 3).

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 populations, 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. An implication for longitudinal assessment of medical expertise is that 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 of particular interest is how the testing effect may be modulated by prior knowledge of the tested domain. Several studies have examined whether the degree of learners' prior knowledge correlates with the magnitude of the testing effect, with mixed results: One study found that retrieval practice has a compensatory effect such that it is more beneficial for learners with low existing topic knowledge (Cogliano et al., 2019, EL: 5), another conversely found that retrieval practice is more beneficial for learners with *high* knowledge (Carpenter et al., 2016, EL: 5), and others found testing equally effective regardless of prior topic knowledge (Glaser & Richter, 2022, EL: 5; Xiaofeng et al., 2016, EL: 5), although all of these studies are limited by their correlational nature. More recently, in an experimental study, Buchin and Mulligan (2023, EL: 4) manipulated learners'

topic knowledge by having them study an academic topic across multiple days of training before introducing a retrieval-practice manipulation; this study found that the testing effect equally benefited high-knowledge and low-knowledge learners.

Other work has examined how the relevance of testing may be modulated by more general academic aptitude or cognitive abilities. The boost provided by testing may be especially helpful for students who would otherwise struggle: A larger testing effect has sometimes been observed for learners lower in the ability to hold information in active memory (*working memory capacity*; Agarwal et al., 2017, EL: 5), in reading comprehension (Callender & McDaniel, 2007, EL: 5), or in general intelligence (Brewer & Unsworth, 2012, EL: 5). Because working memory typically declines with age (Park et al., 2002, EL: 5), this may make testing particularly important for older physicians. However, other studies have found testing benefits to be equal regardless of working memory or general intelligence (Bertilsson et al., 2021, EL: 5; Jonsson et al., 2021, EL: 5; Pan et al., 2015, EL: 5; Wiklund-Hörnqvist et al., 2014, EL: 5).

In general, then, there does not seem to be consistent evidence that retrieval practice benefits only a select group of learners, either in terms of prior knowledge or general cognitive ability. Instead, Jonsson et al. (2021) conclude that retrieval practice is "a learning method for all." This means that physicians are likely to be among those who benefit from the testing effect, and moreso that testing could help physicians across a range of backgrounds and knowledge.

Feedback after testing

When learners are tested—either during 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 so perhaps a more didactic approach that prevents learners from making mistakes would be better (e.g., *errorless learning*; for further discussion, Metcalfe, 2017, EL: 3; Middleton & Schwartz, 2012, EL: 2).

Evidence indicates that the benefits of testing for long-term learning do indeed depend in part 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 et al., 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 pretest memory strength for individual pieces of information. In contrast, restudy conditions provide a memory boost for all items reviewed, but it is a weaker boost than received for correctly recalled items in the test condition.

However, 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). Indeed, as long as feedback is given, errors generated by learners in practice testing do not impair long-term performance (Butler et al., 2008, EL: 3; Huelser & Metcalfe, 2012, EL: 3; Kang et al., 2011a, 2011b, EL: 3; Kornell et al., 2015, EL: 3; Kornell et al., 2009, EL: 3; Kornell & Metcalfe, 2014, EL: 4; Metcalfe, 2017, EL: 3; Metcalfe & Kornell, 2007, EL: 4; Richland et al., 2009 EL: 3; c.f., Knight et al., 2012, EL: 3, for more mixed results). In fact, testing with feedback is so powerful that an unsuccessful retrieval attempt followed by feedback is more beneficial than simply reading the correct information without attempting retrieval (Kornell et al., 2009, EL: 4; Hays et al., 2013, EL: 4; Richland et al., 2009 EL: 4). Thus, the concern that errors during learning undermine long-term knowledge is unfounded so long as feedback is given.

Further, because corrective feedback allows people to learn even from difficult tests, feedback allows learners to be presented with more challenging and demanding tests (e.g., short answer rather than multiple choice) that lead to better learning (Kang et al., 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. 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.

How should feedback be given?

The form of 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 et al., 1991, EL: 1; Fazio et al., 2010, EL: 3; Metcalfe, 2017, EL: 3; Moreno, 2004, EL: 3; Pashler et al., 2005, EL: 4; Whyte et al., 1995, EL: 4) although this may be qualified by the learner's knowledge level (Hausmann et al., 2013, EL: 5).

Some studies have also examined 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 yield gains over providing the correct answer alone (Bangert et al., 1991, EL: 1; Corral & Carpenter, 2020, EL: 4; Kulhavy et al., 1985, EL: 4; Mandernach, 2005, EL: 4; Smits et al., 2008, EL: 4; Whyte et al., 1995, EL: 4, but see Butler et al., 2013, EL: 3, for somewhat more mixed results). 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 et al., 2018, EL: 3), but, to date, there is not much research on this approach. In sum, there is evidence that feedback should include the correct answer, but further explanation beyond that may be unnecessary.

Another relevant feature of feedback is its reliability and validity. Gnepp et al., (2020, EL: 3) found that individuals may be skeptical of negative feedback when the feedback provider's accuracy or credentials are in question. This study examined workplace feedback from a manager, and it likely differs from the relative objectivity offered by an automated system providing feedback about errors. Still, it suggests there may be value to citing information sources in feedback to add authority and objectivity.

When should feedback be given?

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 (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., 2011a, 2011b, EL: 4; Metcalfe et al., 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 et al., 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 et al., 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 because 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 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 et al., 2015, EL: 4; Metcalfe & Miele, 2014; EL: 4), then incorporate the corrected information into their next retrieval attempt.

Why does feedback help?

Why is feedback so effective at ameliorating errors? One possible mechanism, of course, is that feedback simply presents another opportunity to encounter correct information. This is supported by the fact that, as we reviewed above, verification feedback alone is not particularly helpful; the correct answer must be provided (Bangert et al., 1991, EL: 1).

Another important factor may be that, when an error is committed with high confidence, the resulting negative feedback can be especially memorable (the *hypercorrection effect*; Butler et al., 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 et al., 2013, EL: 5; Metcalfe, 2017, EL: 3; Metcalfe & Finn, 2011, EL: 5; Sitzman et al., 2015, EL: 5). The importance of such hypercorrective feedback accords with multiple theoretical perspectives in 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 et al., 2011, EL: 3). Thus, feedback seems particularly effective at 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 related phenomenon, converse to the hypercorrection effect, is that if the learner *is* correct, but has low confidence (e.g., a "lucky guess"), feedback increases the probability that this correct response will be retained later (Agarwal et al., 2012, EL: 4; Butler et al., 2008, EL: 3; Fazio et al., 2010, EL: 3; c.f., Pashler et al., 2005, EL: 4). Thus, we recommend providing feedback for correct

as well as incorrect responses. Although feedback may be redundant when a learner is highly confident in their response *and* correct, it is unlikely to negatively affect learning (Hays et al., 2010, EL: 4; Karpicke & Roediger, 2008, EL: 4).

Finally, feedback can perhaps serve as a cue to forget or inhibit incorrect information.[1] In general, when people are explicitly told that some information is incorrect, obsolete, or otherwise should now be forgotten, they can favor retention of other, to-be-remembered information (the phenomenon of *directed forgetting*; MacLeod, 1998; EL: 2; Sahakyan et al., 2013, EL: 2). Feedback that one is incorrect or has performed poorly may be a cue to initiate this directed forgetting process on erroneous knowledge.

Training people to use retrieval practice

Most research on the testing effect has focused on testing administered by educators and professional organizations. However, learners can also choose to test themselves as a learning strategy. Unfortunately, research indicates that, on the whole, learners use this strategy only rarely; students often prefer less efficacious strategies, like re-reading (Karpicke et al., 2009, EL: 5; Kirk-Johnson et al., 2019, EL: 5), including learners in the health sciences (Coker et al., 2018; EL: 5; Jouhari et al., 2016, EL: 5; Piza et al., 2019, EL: 5). Further, even those who *do* employ testing might do it for other reasons—for instance, to assess what they have learned from other study activities rather than as a learning activity in its own right (Hartwig & Dunlosky, 2012, EL: 5; Kornell & Son, 2009: EL 5).

Nevertheless, some learners *do* use testing to study, and they appear to reap learning benefits from it. In laboratory studies, learners who choose to employ more testing show better retention (Karpicke, 2009, EL: 5). Outside of the laboratory, college students who report using more retrieval practice in their own self-regulated learning have higher GPA (Hartwig & Dunlosky, 2012, EL: 5). This conclusion also extends to medical students: Students who employ more practice testing perform better in the first year of medical study (Baatar et al., 2017, EL: 5; West & Sadoski, 2011, EL: 5) and on medical licensing examinations (Burk-Rafel et al., 2017, EL: 5; Deng et al., 2015, EL: 5); West and Sadoski (2011, EL: 5) and Burk-Rafel et al., (2017, EL: 5) both found that the retrieval practice in self-directed study *better* predicts performance than more general academic measures, such as MCAT scores and undergraduate GPA. Although these studies are correlational, when combined with the experimental evidence for the testing effect discussed above, the role of retrieval practice in these students' learning is likely causal. In sum, the literature suggests that many learners, including medical students, do not often leverage retrieval practice, but those who do benefit in their knowledge and academic performance.

Why don't more learners engage in these useful study behaviors? First, they may be aware of the benefits of testing but do not implement it because of the required time and effort and other costs (see also Nokes-Malach et al., 2022). For example, 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% engage in it. Second, students may not have been taught beneficial learning strategies to begin with: Piza et al., (2019, EL: 5) found that the majority of the health profession faculty they surveyed held misconceptions about evidence-based study practices.

As a result, some researchers have examined whether learners can be taught to use testing approaches for learning. Some evidence suggests that individuals who have more formal education in cognitive psychology (McCabe, 2011, EL: 5) or who are assigned practice that allows them to experience the testing effect (Ariel & Karpicke, 2017, EL: 4; Einstein et al., 2012, EL: 5; Tullis et al., 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 increased both college students' intention to apply retrieval practice and their resulting exam performance (Stanger-Hall et al., 2011, EL: 4). And after implementing a supplemental spaced-repetition learning system with attendees at a continuing medical education conference, Shaw et al., (2011, EL: 3) found that 97% of participants stated interest in participating in the system again in the future. An implication for longitudinal assessment of medical expertise, then, is that if physicians are guided to experience the learning benefits of self-testing, they may also adopt more effective study and learning procedures even beyond the assessment itself.

Mechanisms

Understanding *how* and *why* retrieval practice works is important for applying it across situations: A strong theoretical account of the testing effect generates predictions about when and where it can be used, 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 why retrieval practice works can facilitate outreach to learners and educators.

The testing effect is consistent with several broad principles of human cognition. The benefits of practicing retrieval can be seen as an instance of

[1] We thank an anonymous reviewer for suggesting this possibility.

transfer-appropriate processing: The activities that make for the most effective learning are generally those that match the way the material will be used later (Roediger & Blaxton, 1987, EL: 3; Roediger & Butler, 2011, EL: 3). 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. It follows from this principle that the best way to potentiate later retrieval is to practice retrieval itself, rather than to reread or perform other activities less closely related to retrieval. Supporting this account, Adesope et al., (2017, EL: 1; c.f., Rowland, 2014, EL: 1) found evidence in their meta-analysis that similarity of initial and final test moderates the testing effect. When practice tests and final tests use identical test formats, a somewhat larger testing effect occurs (Hedges' $g = 0.63$), compared to when practice tests and final tests differed in format (Hedges' $g = 0.53$).

However, the 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 more effortful or less accurate, 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: 2).

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 et al., 2009, EL: 5; Kornell & Bjork, 2007, EL: 5; Kornell & Son, 2009, EL: 5; Yan et al., 2014, EL: 5). Intuitively, learners may view practicing retrieval as a way to identify what one does and does not know, but not as way to potentiate learning in and of itself. An analogy is 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" metaphor is inconsistent with another broad-standing principle of memory (Karpicke, 2012, EL: 6): Retrieval is in fact a potent *modifier* of memory (Anderson et al., 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 it as well (Roediger & Karpicke, 2006b; Spellman & Bjork, 1992).

More recently, researchers have investigated the cognitive mechanisms of testing in particular. One reason that testing may benefit retention is that it increases 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 targets). More generally, retrieval practice may lead learners to *elaborate* on the target material by 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; Lehman et al., 2014, EL: 4; Peterson & Mulligan, 2013, EL: 3). For example, the life cycle of the malaria parasite comprises multiple stages, including *sporozoites* and *merozoites*, which learners can easily confuse; however, practice 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, retrieval as a modifier of memory) and can take effect through general cognitive mechanisms (elaboration, distinctiveness, 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 testing and retrieval practice will be *most* beneficial when it closely resembles the desired outcome. For instance, retrieval practice with basic factual knowledge alone is less likely to have an impact on clinical behaviors. Rather, assessments will contribute more to learning if they better match the environments physicians encounter in their practice—for instance, by incorporating simulated diagnosis or treatment scenarios.

Future directions

Although there is robust evidence for the testing effect in general and for several key moderators, we highlight three open questions particularly relevant to the optimal use of using testing in the context of longitudinal assessment of physicians' medical expertise.

Use and degree of interleaving

Although meta-analytic evidence indicates learning benefits from interleaving concepts, these studies have employed a variety of learning activities, not only retrieval practice (but see Dobson, 2011, EL: 4, for an example employing retrieval practice). It would be valuable to confirm that the learning benefits of interleaving obtain specifically in the case of testing. Further, most classroom and laboratory studies comparing interleaved and blocked schedules have used a relatively small number of categories or concepts (e.g., four different types of mathematical solids). However, continuing certification program assessments contain many more concepts. 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 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 hypothesize this latter schedule would yield better long-term learning.

Type of explanation in feedback

Assessments often provide explanations of the correct answer when providing feedback; however, we reviewed evidence that such explanations do not necessarily benefit the learner beyond simply receiving the correct response. A study that manipulates the type of explanations provided during feedback may offer insight into how to improve feedback. One possible design would be to compare later learning outcomes given (a) feedback that uses concrete examples to illustrate a point in addition to providing a technical explanation of how the item should be answered versus (b) only an explanation, but no illustrative example.

Presence of citations during feedback

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.

Summary and conclusion

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 enhances subsequent retention and results in better learning than restudy, concept mapping, and many other educational techniques.

Here, we considered the relevance of this testing effect for the retention of medical expertise in light of the fact that medical professionals often take periodic tests or assessments as part of their career. For instance, to maintain certification by one of the Member Boards of the American Board of Medical Specialties, physicians in the USA must participate in periodic Maintenance of Certification assessments. However, the principles we have outlined would apply to other professions within the health science, such as nurses or dentists, as well.

The robust evidence for the testing effect implies that such longitudinal assessments can be learning opportunities ("assessment for learning") as well as summative assessments of a physician's cognitive skills ("assessment of learning"). 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 medical practice. Fortunately, the reviewed literature indicates that being on some tested information can indeed also improve retention for different, but related information.

The testing effect also generalizes across types of knowledge and tests. A variety of test 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). Further, 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, clinical skills) as well.

Lastly, although there are some conflicting findings, the testing effect broadly seems to generalize across learners with a range of prior knowledge or cognitive abilities.

The benefits of testing can be further strengthened by leveraging several important moderator variables. First, the positive effects of testing can be reinforced by increasing the retention interval length. Although it is challenging to determine exactly when a subsequent test should occur, given that clinicians 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. Second, placing gaps between testing sessions themselves maximizes learning outcomes (i.e., spaced repetition). Having tests distributed over time, versus in a contiguous block, should be a key feature to any longitudinal assessment program. Third, switching topics from item to item (interleaving) is likely to be more beneficial than many questions about one topic in a row (blocking). Interleaving may be especially beneficial for easily confused topics, so we suggest using interleaving to bolster cognitive skills and knowledge for targeted areas within medicine (e.g., when two distinct conditions share similar symptoms). Fourth, multiple tests can further boost learning beyond the baseline benefits of a single test, though with diminishing returns. Fifth, ending an assessment "on a high note" with a few relatively easy problems may increase learners' willingness to engage in future testing by capitalizing on the peak-end rule.

One particularly important moderator is feedback, which enhances the learning benefits of testing and is recommended for any longitudinal assessment framework. Feedback can allay concerns over errors generated during a test, and it is especially important when learners respond wrongly (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. Further, because corrective feedback allows learners to learn from even difficult tests, learners can be presented with more challenging and demanding tests. In providing feedback to a learner, explanations for correct/incorrect responses have not been reliably shown to aid learning beyond simply providing the correct answers; however, the use of examples during feedback may be useful and is worth further investigation. Citations for sources of information and reference materials may also be beneficial. There remain open questions about *when* learners should receive feedback; we found evidence that delayed feedback may be superior to immediate feedback, but due to sparse evidence in applied domains, we believe this should be tested within medicine.

A final benefit to a longitudinal assessment program is that guiding practitioners to experience the learning benefits of testing, and highlighting these benefits, may lead them to adopt more effective study and learning habits on their own.

Despite robust evidence for the testing effect in general, relatively limited work has examined the efficacy of retrieval practice in physicians, and more rigorous scientific work is needed. The few studies that have been done often involved designs that limit causal attribution (e.g., cross-sectional, self-report, or correlational methods), although a few well-controlled studies do exist (e.g., Larsen et al., 2009). Further, only a relatively small subset of studies in the medical domain have 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. Nevertheless, despite these valid limitations, basic-science approaches provide a plethora of evidence that testing should benefit cognitive skills in the domain of medicine. By practicing retrieving information from our memory, we strengthen our memories and increase our knowledge.

Acknowledgements
We thank Andrew Bazemore, Rebecca S. Lipner, David B. Swanson, and Thomas O'Neill for feedback on earlier drafts of this work.

Author contributions
SF wrote the first draft of the manuscript. ZC, TN-M, and BR provided feedback. All authors contributed to revising the manuscript.

Funding
This work was funded by a Grant from the American Board of Internal Medicine (ABIM), American Board of Medical Specialties (ABMS), and American Board of Family Medicine (ABFM). Individuals from ABIM, ABMS, and ABFM provided feedback on the overall goals of the review and on earlier drafts of the manuscript, but approval of the final manuscript rested with the authors alone.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors were not involved with the peer review process of this work.

Received: 1 March 2022 Accepted: 26 July 2023
Published online: 08 August 2023

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., 47*, 1083.

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, H. L., III. (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*. Washington University in St. Louis.

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

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.

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

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.

Bertilsson, F., Stenlund, T., Wiklund-Hörnqvist, C., & Jonsson, B. (2021). Retrieval practice: Beneficial for all students or moderated by individual differences? *Psychology Learning & Teaching, 20*(1), 21–39.

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

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. https://doi.org/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.

Buchin, Z. L., & Mulligan, N. W. (2023). Retrieval-based learning and prior knowledge. *Journal of Educational Psychology, 115*(1), 22–35.

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. https://doi.org/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, H. L., III. (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.

Caddick, Z. A., Fraundorf, S. H., Rottman, B. M., & Nokes-Malach, T. J. (2022). *Cognitive perspectives on maintaining physicians' medical expertise: II. Acquiring, maintaining, and updating cognitive skills*. Manuscript submitted for publication.

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, H. L., III. (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.

Cho, K. W. (2021). A hack for learning math: Starting and ending on high notes to create a more pleasurable learning experience. *Educational Psychology Review, 41*(9), 1082–1096.

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.

Cogliano, M., Kardash, C. M., & Bernacki, M. L. (2019). The effects of retrieval practice and prior topic knowledge on test performance and confidence judgments. *Contemporary Educational Psychology, 56*, 117–129.

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*. 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.

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. https://doi.org/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.

Diener, E., Wirtz, D., & Oishi, S. (2001). End effects of rated life quality: The James Dean effect. *Psychological Science, 12*(2), 124–128.

Do, A., Rupert, A. V., & Wolford, G. (2008). Evaluations of pleasurable experiences: The peak-end rule. *Psychonomic Bulletin & Review, 15*(1), 96–98.

Dobson, J. L. (2011). Effect of selected "desirable difficulty" learning strategies on the retention of physiology information. *Advances in Physiology Education, 35*(4), 378–383.

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). Über das Gedächtnis.

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., & Miele, D. (2016). Hitting a high note on math tests: Remembered success influences test preferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(1), 17–48.

Finn, B., & Miele, D. (2021). Boundary conditions of the remembered success effect. *Journal of Applied Research in Memory and Cognition, 10*(4), 621–641.

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

Firth, J., Rivers, I., & Boyle, J. (2021). A systematic review of interleaving as a concept learning strategy. *Review of Education, 9*(2), 642–684.

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.

Glaser, J., & Richter, T. (2022). The testing effect in the lecture hall: Does it depend on learner prerequisites? *Psychological Learning and Teaching, 22*, 159.

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). An evaluation of the effectiveness of just-in-time hints. In *International conference on artificial intelligence in education* (pp. 791–794). Springer.

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.

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

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.

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

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. https://doi.org/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.

Jonsson, B., Wiklund-Hörnqvist, C., Stenlund, T., Andersson, M., & Nyberg, L. (2021). A learning method for all: The testing effect is independent of cognitive ability. *Journal of Educational Psychology, 113*(5), 972–985.

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.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*(6), 401–405.

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

Kang, S. H., McDermott, K. B., & Roediger, H. L., III. (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. (2011b). 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, H. L., III. (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, H. L., III. (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. https://doi.org/10.1126/science.1152408
Keith, N., & Frese, M. (2008). Effectiveness of error management training: A meta-analysis. *Journal of Applied Psychology, 93*, 59–69. https://doi.org/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.
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.
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. https://doi.org/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* (pp. 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. https://doi.org/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. https://doi.org/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, H. L., III. (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.
MacLeod, C. (1998). Directed forgetting. In J. M. Golding & C. M. MacLeod (Eds.), *Intentional forgetting: Interdisciplinary approaches* (pp. 1–57). Lawrence Erlbaum Associates Publishers.
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, H. L., III. (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, H. L., III., & 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.
Nokes-Malach, T. J., Fraundorf, S. H., Caddick, Z. A., & Rottman, B. M. (2022). *Cognitive perspectives on maintaining physicians' medical expertise: V. Using an expectancy-value framework to understand the benefits and costs of testing.* Manuscript submitted for publication.
Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology, 74*(1), 18–22.
O'Day, G. M. (2022). *Ending on a high note: A simple technique for encouraging students to practice retrieval*. Purdue University.
Ohrt, D. D., & Gronlund, S. D. (1999). List-length effect and continuous memory: Confounds and solutions. In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model* (pp. 105–125). Lawrence Erlbaum Associates Publishers.
Pan, S. C., Pashler, H., Potter, Z. E., & Rickard, T. C. (2015). Testing enhances learning across a range of episodic memory abilities. *Journal of Memory and Language, 83*, 53–61.
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., & 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., & 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*, k663.
Roediger, H. L., III., 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, 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, H. L., III., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*(1), 20–27.
Roediger, H. L., III., & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*(3), 249–255.
Roediger, H. L., III., & 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.
Rottman, B. M., Caddick, Z. A., Nokes-Malach, T. J., & Fraundorf, S. H. (2022). *Cognitive perspectives on maintaining physicians' medical expertise: I. Reimagining maintenance of certification to promote lifelong learning*. Manuscript under review.
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.
Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*(4), 734–760.
Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Exploration in the microstructure of cognition* (pp. 216–271). Cambridge, MA: MIT Press.
Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition, 11*(6), 641–650.
Sahakyan, L., Delaney, P. F., Foster, N. L., & Abushanab, B. (2013). List-method directed forgetting in cognitive and clinical research: A theoretical and methodological review. In B. H. Ross (Ed.), *The psychology of learning and motivation* (pp. 131–189). Elsevier.
Sala, G., & Gobet, F. (2017). Does far transfer exist? Negative evidence from chess, music, and working memory training. *Current Directions in Psychological Science, 26*(6), 515–520.
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*, (pp. 702–708), Cambridge

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, 48*, 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. 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, E. K., Jr. (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. (2020). 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*. 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., & Sadoski, M. (2011). Do study strategies predict academic performance in medical school? *Medical Education, 45*(7), 696–703. https://doi.org/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, H. L., III. (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.
Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55*, 235–269.
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.
Xiaofeng, M., Xiao-e, Y., Yanru, L., & AiBao, Z. (2016). Prior knowledge level dissociates effects of retrieval practice and elaboration. *Learning and Individual Differences, 51*, 210–214.
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, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435.
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.

Publisher's Note

Cognitive Research: Principles and Implications

REVIEW ARTICLE

Open Access

Cognitive perspectives on maintaining physicians' medical expertise: V. Using a motivational framework to understand the benefits and costs of testing

Timothy J. Nokes-Malach[1,2], Scott H. Fraundorf[1,2*], Zachary A. Caddick[1,2] and Benjamin M. Rottman[1,2]

Abstract

We apply a motivational perspective to understand the implications of physicians' longitudinal assessment. We review the literature on situated expectancy-value theory, achievement goals, mindsets, anxiety, and stereotype threat in relation to testing and assessment. This review suggests several motivational benefits of testing as well as some potential challenges and costs posed by high-stakes, standardized tests. 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. Attention to instructional framing, test purposes and values, and longitudinal assessment frameworks provide vehicles to further enhance motivational benefits and reduce potential costs of assessment.

Keywords Motivation, Learning, Assessment, Expectancy value, Achievement goals, Mindsets, Stereotype threat, Test anxiety

Significance

Physicians in the USA are required to take continuing education assessments at various points throughout their careers. The medical boards that administer those assessments are considering changes in their structure and implementation, including a more longitudinal assessment model. Understanding the role that motivation can play for learners in both preparing for and taking continued education assessments can inform the assessments' design, purpose, and the policies for giving them. We take a motivational perspective on the potential benefits and costs of testing and the implications of longitudinal assessment. We review prior research on motivation for learning from cognitive, social, and educational psychology, including studies from both laboratory and classroom settings. This analysis reveals that perceived test difficulty and expectations of success, instruction framing and feedback, alignment to the values of the learner, and creating multiple lower-stakes assessment opportunities are critical issues to consider when redesigning and implementing continued educational assessments to enhance motivation, learning, and performance.

Introduction

Physicians in the USA are required to take continuing education assessments at various points throughout their careers. Currently, many of these assessments take place every 5 to 10 years and can be viewed as summative assessments; however, many specialty boards are considering transitioning to shorter, more longitudinal

*Correspondence:
Scott H. Fraundorf
scottfraundorf@gmail.com
[1] Learning Research and Development Center, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA
[2] Department of Psychology, University of Pittsburgh, 3420 Forbes Ave., Pittsburgh, PA 15260, USA

Table 1 Evidence levels for in-text citations for empirical claims

Evidence level	Type of work
1	Quantitative meta-analysis
2	Narrative review
3	Multiple original experiments/randomized controlled trials (RCTs)
4	Single original experiment (RCT)
5	Correlational or quasi-experimental study
6	Opinion paper

assessments that can also serve as learning opportunities. This change presents an opportunity to consider some of the factors that could enhance the learning value of these assessments or otherwise make them more motivating for physicians.

In particular, for physicians to be motivated to participate in longitudinal assessments and other learning opportunities, they must view participation as having relatively high value and low costs. One general approach that can help us understand the role of value and costs in learning is the expectancy-value theory from social and educational psychology (Eccles & Wigfield, 2002, 2020; Wigfield & Eccles, 2000; Wigfield et al., 2016). This theory is widely used to explain, understand, and predict human motivation in learning and in academic performance. Expectancy-value theory posits that learners' pursuit of 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 (their *expectancies*), as seen in Eq. 1 as follows:

$$\text{Motivation to Learn} = \text{Expectancies} * (\text{Benefits} - \text{Cost}) \quad (1)$$

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 (e.g., new knowledge, feedback on knowledge and skills, continued certification), when the costs of doing so are relatively minor (e.g., reasonable time and effort required), and when there is an expectation of success.

In this paper—part of a collection of five articles in this special issue focused on how physicians maintain medical expertise across their careers—we describe a motivational framework that builds on the expectancy-value theory that also connects to several related motivational theories and ideas from research on achievement goals, mindsets, stereotype threat, and test anxiety. We adopt the approach of a narrative review, not systematic, because we cover a wide variety of topics. To situate the strength of the evidence and claims made, we attach evidence levels (EL) to in-text citations for empirical claims (see Table 1). Evidence levels range from 1 to 6, with 1 being the strongest evidence (meta-analyses) and 6 being the weakest (opinion papers).

Understanding motivation with situated expectancy-value theory

The most recent articulation of the expectancy-value theory has been called the *situated expectancy-value theory* (Eccles & Wigfield, 2020). This version of the theory accounts for both long-term trajectories in the development of expectancies and values as well as the short-term psychological processes engaged in task choice and performance. This version of the theory strongly emphasizes the situated nature of the factors that affect motivation, which include not only the particular features of the immediate situation and task but also the culture(s) an individual resides in, personal characteristics (e.g., gender, race / ethnicity, socioeconomic status, etc.), and past personal experiences related to the achievement activity. Viewing features of the situation as integral to the motivational processes and outcomes mirrors similar approaches in the cognitive and learning sciences that have also focused on the importance of the situation for cognitive activity and for learning and transfer outcomes (Greeno & MMAP, 1997, 1998; Lave, 1988; Lave & Wegner, 1991). We believe that the situated perspective is especially relevant to our present goal of understanding motivation in a particular situation and context—in this case, physicians' engagement with continuing certification assessments. There are several different features of this context that are important to motivation, including both the particular prior educational and assessment experiences of physicians in the USA (e.g., attending medical school, certification tests) as well as the stereotypes associated with the profession (e.g., about who can be a physician or what resources are required to succeed).

We expand certain components of the situated theory of expectancy-value to accommodate relevant cognitive research. For example, as we elaborate upon below, previous work has focused on *expectancies of success*, but here we broaden the scope of the term and include other expectancies, such as *expectancies of being tested*, which also plays an important role in the cognitive literature on memory and learning. We highlight where there are variations—conceptual broadening or narrowing—from the situated expectancy-value concepts and features.

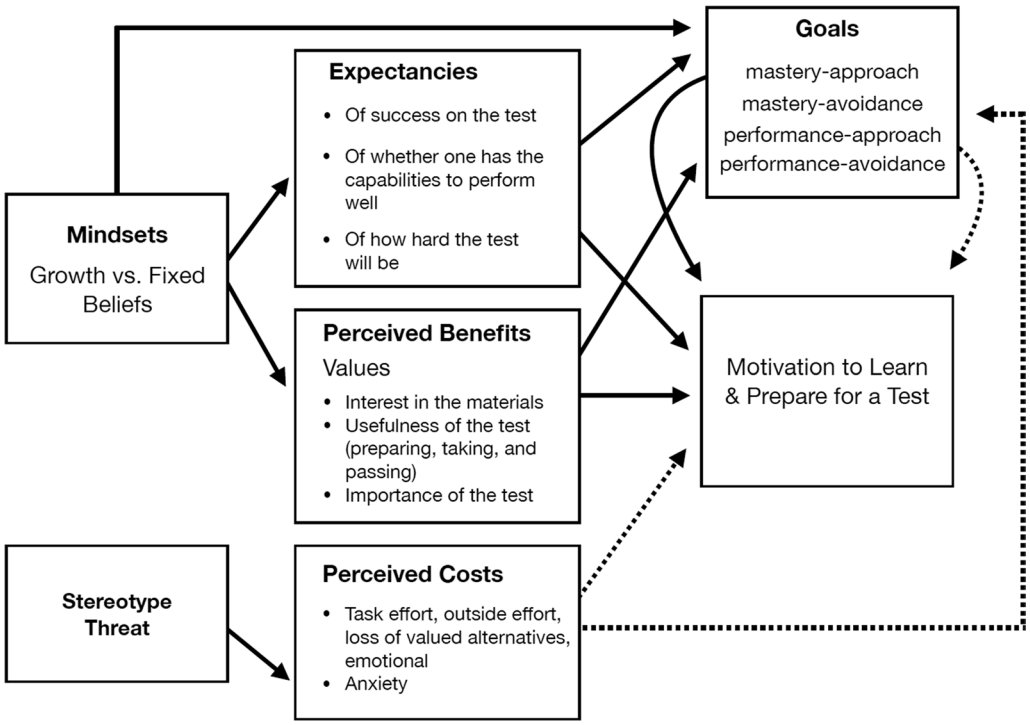

Fig. 1 Situated expectancy-value model of the motivation to learn and interrelations to mindsets, stereotype threat, and achievement goals. The model is adapted from Eccles, Wigfield, and colleagues (Eccles & Wigfield, 2020; Wigfield & Eccles, 2000). Positive relations are denoted by the solid arrows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

Further, we review and include additional motivational theories and ideas that we believe are of particular relevance to the context of physicians maintaining their expertise and completing continuing assessment. These include achievement goals (given that medical training and performance contexts often have a strong focus on mastery and performance), mindsets and stereotype threat (given varying beliefs about the profession and who can succeed), and test anxiety (given that we are focusing on continued assessment that often takes the form of high-stakes standardized tests).

Another reason to bring these different motivational ideas together into a single framework is to help further connect relevant motivational theories and ideas whose relations are often not explicitly discussed in the literature but whose features and processes often overlap and relate to one another. There have been many calls over the years to integrate these different frameworks or to compare and contrast them (e.g., Hattie et al., 2020). Although our goal here is to review those that we view as particularly relevant to physicians and the maintenance of certification, by expanding the situated expectancy-value framework to relevant cognitive research on learning and testing, we connect these strands of prior research back[1] together in the hope that future theorizing will further integrate cognitive and motivational theories of learning and performance.

In Fig. 1, we present a model in which we bring together each of the motivational factors of interest, their hypothesized interrelations, and the motivation to learn. We begin by reviewing research on the effects of one's expectancy for passing the test and of self-efficacy beliefs on engagement and learning outcomes. We then review how learners perceive the benefits of testing and explore the hypothesis that physicians will experience stronger motivation and learning to the degree that the assessment aligns with and confers value to them. We then consider related research on mindsets, which in the context of situated expectancy-value theory, can be viewed as ability beliefs that can influence expectancies and values. We also incorporate achievement goals as both a factor affected by expectancies and values as well as a mediator of the motivation to learn. Next, we discuss the potential perceived costs of testing, such as text

[1] Work on learning and motivation at the beginning of the twentieth century by Thorndike, Hull, and Tolman, among others, closely connected the processes of motivation and learning together within singular theories (see Klein, 2019, for an overview). We hope that future work continues to bring research on cognition and learning and motivation and learning back together in integrated ways.

anxiety and stereotype threat (i.e., a situation in which one is concerned about potentially confirming a negative stereotype related to an aspect of their identity), as well as approaches to mitigate those costs. We end with a discussion of directions for future work in the areas of motivation and the development of medical expertise.

Expectancies

Expectations of success affect how one studies and performs

A learner's beliefs about the likelihood of success on a given task has important consequences for their learning activities. Such *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; Bandura, 1997). In the situated expectancy-value model, a learner's beliefs about the likelihood of success on a given task affects their motivation to learn. For example, if a learner is given a task that they perceive themselves as unlikely to succeed on (e.g., the test is extremely difficult or insufficient time is given), then they will be less likely to engage in that activity or to prepare adequately because they expect 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 and failure may be defined differently for different physicians. Prior work has shown that expectancy beliefs have a large impact on academic performance (Meece et al., 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 et al., 2006, EL: 5; Simpkins et al., 2006, EL: 5).

Conversely, high self-efficacy is associated with more productive learning behaviors (Bouffard-Bouchard et al., 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 to persist in trying to learn even in the face of difficulties or challenges (Bandura, 1997, EL: 2; Schunk & Parajes, 2002, EL: 2). These beliefs predict student retention and academic performance in school settings (Honicke & Broadbent, 2016, EL: 1) even after controlling for prior knowledge (Bailey et al., 2017, EL: 5; Kalender et al., 2020, EL: 5). A number of factors have been hypothesized to influence the 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).

This research on self-efficacy has several implications for continuing assessment of medical expertise. First, an assessment must strike a balance of difficulty in that it is perceived as challenging enough to motivate constructive study activities to prepare for that test, but not so difficult 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. 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, that provides a further opportunity to develop self-efficacy by providing multiple pieces of feedback over time. Each piece 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 move up or down depending on performance). Providing repeated performance feedback creates an opportunity to determine whether such feedback leads to more accurate self-assessment.

Expectations of test difficulty affect engagement and performance

While the expectancies component of the situated expectancy-value model focuses on expectancies for success on a task, as we discuss above, we also consider here how the expectation that one will be tested in the future can itself create the opportunity to engage in productive study and learning activities. Although just knowing about the existence of an upcoming assessment does not necessarily promote learning (Hyde & Jenkins, 1973, EL: 4; Postman, 1964, EL: 3), an expectation of being tested in a particular way or on certain types of material can lead to better learning (McDaniel et al., 1994, EL: 3; Szpunar et al., 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 et al., 1983,

EL: 4; Hall et al., 1976, EL: 3; Leonard & Whitten, 1983, EL: 3; Maisto et al., 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). Learners expecting a more difficult test can 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).

The results reviewed above suggest that physicians learn and retain more when they expect to be tested on the knowledge and skills they are developing. This is especially the case when the perceived difficulty of the test is difficult enough[2] to engender deeper, more effective preparation and when the environment guides physicians to effective study behaviors and activities to capitalize on their increased motivation.

Perceived benefits: what is the value of the test to the learner?

In the situated expectancy-value model, the perceived value of a given task or assessment plays a critical role in motivation to prepare for and engage with it. Value is hypothesized to consist of three distinct components, each of which we review in turn (Wigfield et al., 2016).

Intrinsic task value

Intrinsic task value is interest in a task or assessment for its own sake. Theories of interest typically discuss two different kinds: situational and individual (Hidi & Harackiewicz, 2000; Hidi & Renninger, 2006; Krapp, 1999; Schiefele, 1991). *Situational interest* is hypothesized to be a momentary experience that is driven by environmental factors (e.g., a loud noise) and correlated with both cognitive (e.g., attention) and affective (e.g., surprise) factors (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 about the particular topic or task (Renninger, 2000, EL: 2). One perspective on how intrinsic task value emerges is given by discrepancy theory, which posits that learners are motivated to increase a valued competency when they perceive a *discrepancy* between their current skill and a given goal or standard (Fox & Miner, 1999; see also regulatory focus theory, Higgins, 1997, 2012). Laboratory studies have confirmed that people are sensitive to gaps between perceived and desired knowledge (Dunlosky & Hertzog, 1998, EL: 5; Son & Metcalfe, 2000, EL: 3; Tullis & Benjamin, 2010, 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 et al., 1995, EL: 5). It is also associated with better grades in school (Harackiewicz et al., 2008, EL: 5; Schiefele et al., 1992, EL: 1). Intrinsic task value may also be linked to achievement goals for a particular task, as we elaborate upon below. In the domain of medicine, medical students' interests and perceived competence have been shown to predict important career choices, such as medical specialty decisions (Williams et al., 1997, EL: 5). Other research has shown that experimental interventions to increase intrinsic task value 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). This finding aligns with discrepancy theory in that learners need to become aware of (i.e., perceive) discrepancies between actual and desired knowledge to become highly motivated.

Interest and task performance can also be increased by personalizing content (Bernacki & Walkington, 2018, EL: 4; Walkington & Bernacki, 2018, EL: 2). This research implies that the more the content of the test (e.g., topics and patient scenarios) matches the interests of the physician, the more motivated they will be to learn and keep current. It would be desirable to collect information on physicians' medical interests and personal practice to match those interests, or perhaps to allow physicians some opportunities to select relevant topic areas or problem scenarios to be tested on. Another technique to potentially help select material in a longitudinal spaced-repetition paradigm would be to collect ratings of relevance and use them to prioritize content which information is re-presented.

Utility value

Another aspect of value is called *utility value*, or the degree to which preparing for and taking the test is useful for accomplishing some valued outcome; that is, as a means to an end (Eccles, 2009; Wigfield & Eccles, 2002).

[2] Of course, we discuss above, if the test is expected to be *too* difficult and outcome expectations are consequently low, learners may instead disengage from the task entirely.

Often, these valued outcomes are broader personal, educational, or professional goals. Correlational research has shown that utility value is positively associated with engagement, learning, and performance outcomes, such as higher grades (Harackiewicz et al., 2016, EL: 2; Harackiewicz et al., 2014, EL: 2). Further, intervention studies have shown that, when utility value increases, so does academic performance (Harackiewicz & Priniski, 2018, EL: 2; Harackiewicz et al., 2016a, 2016b, EL: 4, Hulleman et al., 2010, EL: 4).

Utility value is relevant to longitudinal assessment of medical expertise in at least two ways. The first concerns the usefulness of preparing to take the assessment. That is, does a physician view preparing for the assessment as a helpful activity that contributes 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 (e.g., acquiring critical new knowledge), the more motivated they will be to study. The second concerns the value ascribed to the test itself. That is, does a physician view the test as useful to achieving broader educational goals (e.g., staying current) and professional goals (e.g., staying employed, being promoted)?

Interviews with physicians preparing for and taking high stakes tests show a range of perceptions of how relevant and related the content is to their current practice (Chesluk et al., 2019a, 2019b, EL: 5). The work we reviewed above implies that such variation in perceptions is likely to affect physicians' motivation to learn. If physicians see the activity of studying between longitudinal assessment sessions as relevant to their broader professional goals, they will be more motivated and more deeply engaged with the material. Alternatively, if they view the assessment as unrelated and disconnected, they may engage only superficially. Fortunately, 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 class or discipline to them (Harackiewicz & Priniski, 2018, EL: 2). Feedback within an assessment can also promote utility value. Some longitudinal assessment platforms require the participant to rate each question's relevance to their medical practice. Periodically providing feedback (e.g., as summary feedback between assessment sessions) regarding questions that a learner missed but that they also rated as relevant to their practice may provide additional motivation for them to review those concepts.

Attainment value

The third component of value is *attainment value*, or the importance 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 measures (e.g., relevant medical knowledge and skills), how accurately it measures 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 performance outcomes provide "data" that can be used to confirm or deny aspects of one's identity (La Guardia, 2009; Ryan & Deci, 2020), including three core needs of autonomy, relatedness, and—most critical to our purposes—competence. If one perceives the assessment as measuring critical medical competence 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 on it, it could call into question either one's view of oneself as an expert, knowledgeable physician, or the validity and accuracy of the test.

Attainment value has been shown to be positively related to engagement (Putwain et al., 2019, EL: 5), effort (Guo et al., 2016, EL: 5), self-concept (Arens et al., 2019, EL: 5), and academic achievement (Trautwein et al., 2012, EL: 5; Meyer et al., 2019; EL: 5). Laboratory research on memory and learning also supports the relevance of attainment value. 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 guides priorities for learning and retention (Castel et al., 2002, EL: 3; Castel et al., 2011, EL: 4; Castel et al., 2013, EL: 3; Hennessee et al., 2018, EL: 3; McGillivray & Castel, 2017, EL: 3).

This work implies 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 time and effort they will invest in performing well on the test. Further, if the assessment provides feedback relevant to physicians' self-concepts and identities (e.g., their identity as a skilled medical doctor), they will show higher investment in developing their skills 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 by apportioning more questions to these topics within the assessment (as is often already done).

Growth mindsets promote 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 used to describe a set of beliefs that can impact one's expectations, meaning-making, and behaviors (Dweck & Yeager, 2019, EL: 2).

One of the most powerful mindsets that has been investigated is people's beliefs about intelligence. Within the situated expectancy-value model of Eccles and Wigfield (2020), mindsets regarding intelligence are captured as part of the self-concept of one's abilities that can influence expectancies and values. Carol Dweck and her colleagues have been some of the leading researchers on mindsets about intelligence and 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, which has been called a *fixed mindset*. Intelligence mindsets and ability beliefs have been hypothesized to affect a learner's expectancies, values, and achievement goals, which in turn affect the motivation to learn (Fig. 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 et al., 2007, EL: 4&5).

A growth mindset predicts positive academic achievement (Costa & Faria, 2018, EL: 1; Blackwell et al., 2007, EL: 4; Gunderson et al., 2013, EL: 5, Henderson & Dweck, 1990, EL: 2, Paunesku et al., 2015, EL: 4; cf. Li & Bates, 2019, EL: 4). Growth and fixed mindsets also relate to students' self-reported interest (Haimovitz et al., 2011, EL: 5), effort (Blackwell et al., 2007, EL: 5; Miele et al., 2011, EL: 5; Miele & Molden, 2010, EL: 3), and learning goals (Blackwell et al., 2007, EL: 5; Haimovitz et al., 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 et al., 2007, EL: 5). There is some evidence that the link between growth mindset and academic achievement is causal: Interventions designed to promote growth mindsets, with messages that portray intelligence as malleable with experience and training, lead to positive changes in motivational and achievement outcomes (Blackwell et al., 2007, Expt. 2, EL: 4; Mueller & Dweck, 1998, EL: 4; Yeager, et al., 2016, EL: 4; c.f. Li & Bates, 2019).

In sum, mindsets about intelligence can have powerful downstream effects on motivational and learning outcomes and can directly impact expectancies, values, and goals. Thus, physicians who believe their intelligence and skills are malleable may be 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 toward more regular spaced testing, which provides the opportunity to improve over time.

Achievement goals and the benefits of pursuing mastery

Achievement goals are the reasons why people engage in study and test activities. Achievement goals are sometimes described and investigated separately from expectancy-value theory, but sometimes are included in an overarching model (Plante et al., 2013). In the situated expectancy-value model of Eccles and Wigfield (2020), long- and short-term goals are described as factors that can influence expectancies for success and subjective task values. Others have hypothesized the converse: that expectancies and values affect the adoption of achievement goals (Elliot, 1999; Greene et al., 2004). Some empirical work supports this second view by showing that expectancies and values have both direct effects on motivation for learning *and* indirect effects through achievement goals (Plante et al., 2013, EL: 5). We incorporate this second view into our model depicted in Fig. 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 (Elliot & McGregor, 2001; Elliot & Murayama, 2008). We thus view achievement

goals as particularly relevant for the context of developing and maintaining medical expertise given the focus on mastery and performance in training and assessment.

Although there has been little work specifically examining achievement goals in the context of practicing physicians, many laboratory experiments and classroom studies have examined these four achievement goals in relation to engagement, learning, and performance outcomes. This literature has consistently linked performance-avoidance goals 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 et al., 1999, EL: 5). In contrast, mastery-approach goals have been consistently associated with positive outcomes, such as self-reported interest and engagement (Elliott & Dweck, 1988, EL: 4; Elliot et al., 1999, EL: 5; Harackiewicz et al., 2002a, 2002b, EL: 5; Harackiewicz et al., 2008, EL: 5) and learning and transfer (Belenky & Nokes-Malach, 2012, 2013, EL: 5). The fact that mastery-approach goals have been related to knowledge transfer is promising for learning in medical education contexts in that it may help physicians acquire sought-after skills critical to adaptive medical expertise (Mylopoulos et al., 2018, EL: 2).

Performance-approach goals correspond to a more intermediate level of performance; they have been related to some positive outcomes, such as better grades and exam performance (Harackiewicz et al., 2002a, 2002b, EL: 2; Linnenbrink-Garcia et al., 2008, EL: 2), but also some negative outcomes, such as less effective self-reported study behaviors (i.e., rote memorization) (Midgley et al., 1996, EL: 5; Senko et al., 2011, EL: 2).

Lastly, 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), and negative outcomes, such as self-reported test anxiety (Elliot & McGregor, 2001, EL: 5).

How do learners come to adopt one type of achievement goal or another? A number of factors can influence achievement goals (Ames, 1992, EL: 2). Prior experimental and classroom work has shown 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 the purpose for a given task is either to "develop their ability or skill and learn from mistakes" or conversely "to compare ability to others and to determine whether they are better or worse than others" can impact learning outcomes and task engagement (Bereby-Meyer & Kaplan, 2005, EL: 3; Elliot & Harackiewicz, 1996, EL: 3). 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.

Perceived costs of testing

General aspects of psychological cost

In the situated expectancy-value framework, an important subcomponent of the motivation to learn is the perceived cost of the study activity or test. This component of the model has historically received less attention than expectancy and other aspects of value[3] (i.e., intrinsic, utility, attainment); however, more recently, several efforts have been made to develop measurement tools that capture important aspects of cost (Conley, 2012; Flake et al., 2015; Trautwein et al., 2012) and to better understand its role in the expectancy-value framework (Barron & Hulleman, 2015; Eccles & Wigfield, 2020). Four aspects of cost have been identified (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), 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 or family time (Galla et al., 2015: EL 5; Kurzban et al., 2013). 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 high-stakes test would increase the perceived emotional cost of the test, as we discuss in further detail below. In interviews with physicians about

[3] In the situated expectancy-value framework, cost is considered a subcomponent of value. Here, we represent cost separately from value in Fig. 1 to highlight its role and to both build on the prior work on costs as well as broaden the definition to consider other types of costs (e.g., financial costs of taking the assessment).

how they prepared for and took continuing certification examinations, the lack of time available because of outside effort involved in studying and the loss of valued alternatives emerged as important themes (Chesluk et al., 2019a, 2019b, EL: 5).

In the situated expectancy-value model, the more of these perceived costs, the less likely one is to be motivated to learn and prepare for the test. Prior work has shown that perceptions of cost predict additional variation in motivation and performance outcomes above and beyond expectancies and values (Jiang et al., 2018; EL 5; Perez et al., 2014; EL 5). In principle, then, the more that perceived costs can be reduced, the stronger an individual's motivation to learn. A recent intervention that aimed to reduce cost in an introductory physics course by focusing on the normalization and temporary nature of effort costs has shown some promising results in reducing subsequent perceived costs and increased class performance (Rosenzweig et al., 2020, EL: 4; c.f. Rosenzweig et al., 2022, EL: 4). This work suggests that one way to mitigate the perceived costs of testing would be to discuss those costs in advance in an effort to normalize them.

Another aspect of continuing certification programs that may impact one'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 tests on perceived costs within the situated expectancy-value framework, but this could be a useful direction for future work.

Test anxiety

As we discuss above, one form of cost in the situated expectancy-value is emotional cost, which includes anxiety. Indeed, there is a substantial literature specifically on test anxiety, which we review here because it is relevant to the context of continued assessment. *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 a negative evaluation in relation to an evaluative test. Several mechanistic 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).

Although some arousal may be good, many individuals approach standardized tests with levels of anxiety that are high enough to impair performance (von der Embse, 2018, EL: 1). Famously, the Yerkes-Dodson law of arousal and performance states that a moderate level of arousal leads to optimal 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 that interfere 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 et al., 2017, EL: 2; Mandler & Sarason, 1952, EL: 4; Sarason, 1980, EL: 2). Test anxiety, in particular, 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).

One reason that high levels of anxiety may be harmful to test-taking is that anxiety 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' $g=-0.33$). 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 testing for learning were diminished if there was significant performance pressure during episodes of memory retrieval. These findings illustrate the importance of reducing pressure during testing in order to maximize learning outcomes. One method that Hinze and Rapp found could reduce pressure was to weigh earlier retrieval-practice questions less than later questions so that learners who do poorly early on can identify areas of weakness and improve upon them. Other work has found that practicing for an assessment using retrieval practice can reduce test anxiety (Agarwal et al., 2014, EL: 3). Thus, the development of a longitudinal assessment focused on learning benefits may reduce the perceived pressure of a one-time, high-stakes test.

Stereotype threat

In the situated expectancy-value framework, stereotypes are described as operating in a broader cultural milieu

that is theorized to impact subsequent self-perceptions and task expectancies and values. Here, we elaborate on these effects and connect them to the broader literature on stereotype threat, another phenomenon that can be triggered in high-stakes testing.

Stereotype threat refers to the diminished performance that can occur when reminded of a negative stereotype in a domain in which one otherwise identifies and has high competence (Steele, 1997). It 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 et al., 2020, EL: 5; Nguyen & Ryan, 2008, EL: 1; Rahhal et al., 2001: EL: 3; Steele & Aronson, 1995, EL: 5; Stone et al., 1999, EL: 5).

Stereotype threat is thought to occur because actors of a stereotyped group may exhibit poorer performance as a consequence of not wanting to reinforce the stereotype. Multiple mediating mechanisms have been proposed for this underperformance, including the depletion of working memory resources 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 et al., 2016, EL: 2). Shewach et al., (2019, EL: 1) found that when motivational incentives, such as a monetary reward, are present, stereotype threat is much less pronounced than when they are absent (Cohen's *d*s 0.14 vs. 0.41, respectively), suggesting that motivation plays an important role in stereotype threat.

Stereotype threat is likely to be applicable to physicians in a continuing assessment given that this context meets the criteria of a domain with which the learners identifies and has high competence (Steele, 1997). Physicians are likely to be highly identified with the medical domain and view themselves as having high competence given their extensive education and training; further, the context of continuing assessment may be viewed as providing results that bear on their evaluation of that competence. However, there are also negative stereotypes associated with particular medical subdisciplines regarding social identities of gender (Fassiotto et al., 2018: EL: 5; Myers et al., 2020: EL: 4) and race and ethnicity (Bullock et al., 2020: EL: 5).

Given the evidence for stereotype threat in testing, a few recommendations for longitudinal assessments may be beneficial. First, stereotype threat is more likely to occur when a test is more difficult (Nguyen & Ryan, 2008, EL: 1; Shewach et al., 2019, EL: 1). Thus, 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 et al., 2016; EL: 2). Second, in situations where demographic information must be collected, this should occur *after* any testing or at some other time not directly before the assessment to reduce potentially activating negative stereotypes related to the individuals' demographics. 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.

Relations between features of the model

The situated expectancy-value framework theorizes that relations between expectancies and values have multiplicative effects on motivation and performance. That is, these components are not just additive but interact with one another. Some prior empirical work has found evidence to support this view in that interactive effects of measures of expectancy and value predicted additional variance in motivation and performance above and beyond expectancy and value alone (Nagengast et al., 2011: EL: 4; Trautwein et al., 2012: EL 4). However, much remains to be discovered as to how expectancy and value interact with one another and with cost.

Proposed studies and future directions

Measuring motivation

There is a lack of research on the effects of motivation in longitudinal testing scenarios, including continuing certification programs. One reason for this may be that, to measure motivation in this context, one must determine not only *what* components of motivation to measure but *how* to measure them.

There are many potential methods to measure aspects of motivation, including self-report surveys, interviews, behaviors (e.g., choice, time spent, errors, etc.), and physiology, among others. We suggest beginning with self-report surveys because researchers have developed validated measures for many of the constructs that we have discussed that can be used in a variety of educational contexts (e.g., self-efficacy, Bandura, 2006; Fencl & Scheel, 2005; interest and value, Linnenbrink-Garcia et al., 2010; Pintrich et al., 1993; cost, Flake et al., 2015; achievement goals, Elliot & Murayama, 2008; mindsets, Dweck, 1999, 2006; test anxiety, Putwain et al., 2021). Self-report surveys are also relatively easy to implement in a longitudinal assessment context. A first step, then, would be to adapt the items to the continuing certification context and conduct validation studies with this new domain and population.

Given such validated instruments, there are many research questions that could then be investigated concerning the role of motivation in longitudinal assessments. We suggest that one place to start would be to assess whether the motivational components reviewed in this paper—expectancies, perceived values and costs,

achievement goals, and mindsets—predict performance in the assessments. Measuring these components of motivation 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. They could also be extremely informative in assessment design decisions and potential interventions. For example, if self-reported utility value strongly predicts performance in the continuing certification context, then interventions (e.g., a brief writing task or instructional framing) could be tested to increase perceptions of utility value, as we elaborate upon below.

In this context, motivation would be relevant both as a process as an outcome. Whether a particular motivational factor, such as utility value, predicts performance in the assessment may shine light onto the specific motivational processes at play when one is taking the assessment. In addition, if it is determined that certain motivational components are particularly important in this context, then measures of motivation could also serve as outcomes or dependent measures for other interventions and assessment changes.

The role of financial cost in a situated expectancy-value framework

As we noted above, there is little or no work examining how the financial cost of taking an assessment may affect motivation, though from the learner's perspective, these may be an important consideration. Future work should examine how financial costs relate to other cost perceptions and how they affect motivation and performance outcomes.

Interventions to increase motivation and performance

Throughout our review, we identified multiple components of our motivational framework for which past work has developed successful interventions. Such interventions can enhance aspects of value, facilitate productive goals and mindsets, and mitigate potential costs. Here, we describe five that may be particularly well suited to the current context of longitudinal assessments.

Interventions that introduce choice and personalization can increase intrinsic task value and performance (Walkington & Bernacki, 2018, EL: 2; Patall et al., 2008, EL: 1). Thus, 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 require any choice within the assessment system itself; the assessment system could automatically assign personalized test content based on an initial survey that the physician would take about their clinical practice.

Second, it might also be possible for interventions to increase the assessment's perceived utility value. For example, we would predict that motivation to learn could be increased by a 10-min exercise in which individuals write about how the preparation and assessment is relevant to their educational and professional goals.

Third, building on interventions in the broader achievement goal literature (Elliot & Harackiewiz, 1996, EL: 3; Elliot & Dweck, 1988, EL: 4; Graham & Golan, 1991, EL: 3), an assessment's instructions and structure could help individuals adopt a mastery achievement goal. For example, instructions could emphasize understanding and improvement. The shift toward longitudinal assessment also allows for a focus on *intrapersonal comparison*; that is, focus on how a physician can improve relative to their past performance rather than other physicians.

Fourth, revising instructions may also provide an avenue to reduce some aspects of perceived costs. Instructions that normalize aspects of the time and effort required to successfully prepare and engage in the longitudinal assessments may reduce perceived cost.

Lastly, discrepancy theory (Fox & Miner, 1999) offers a method for measuring and instilling motivation that, although originally proposed for other forms of continuing medical education, can readily be adapted for longitudinal assessment programs. First, physicians subjectively rate aspects of their clinical competency (e.g., knowledge of diabetes) and the desirability of those aspects (e.g., how important to you is it to have expert knowledge about diabetes?). Then, they take an assessment of the target competency. Finally, a discrepancy score between perceived and actual competency is computed and presented. 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—especially in the context of a longitudinal assessment program, where they could focus on those topics when studying for the next assessment.

Summary and conclusion

In this paper, we took a situated expectancy-value approach to thinking about the role of motivation in continuing certification programs. We reviewed basic motivational theory and empirical work from laboratory and classroom settings, and we 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.

Many of the motivational benefits for testing can be understood from the equation of having the perceived benefits of a test outweigh the perceived costs of preparing for and taking it. We found that a sufficiently challenging test can increase 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. Two ways to make clear the level of difficulty are to describe the specific task items used and to give representative problems to practice and receive feedback on.

We also reviewed three components of value (intrinsic, utility, and attainment) that should be attended to when designing an 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). Similarly, framing the longitudinal assessment and feedback as an opportunity to learn and develop can further facilitate the adoption of mastery-approach goals 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. High-stakes assessment can also carry an emotional cost in the form of test anxiety, but the move to a longitudinal assessment scheme of more frequent testing may reduce test anxiety relative to less frequent, higher-stakes tests. We also discussed the related phenomenon of stereotype threat, which can be mitigated by emphasizing the assessment as an opportunity to improve (as opposed to a high-stakes, evaluative test), highlighting the components of value previously described, asking demographics at the end of the assessment or at some other time not right before the assessment, and including diverse demographic features in the testing clinical scenarios.

By considering how both motivational and cognitive factors relate to the benefits and costs of longitudinal assessment, future work can build theories that integrate across these frameworks and a practical opportunity to design multi-purpose assessments that are both engaging and useful.

Acknowledgements
We thank Andrew Bazemore, Rebecca S. Lipner, David B. Swanson, and Thomas O'Neill for feedback on earlier drafts of this work.

Author contributions
TN-M wrote the first draft of the manuscript. SF, ZC, and BR provided feedback. All authors contributed to revising the manuscript.

Funding
This work was funded by a grant from the American Board of Internal Medicine (ABIM), American Board of Medical Specialties (ABMS), and American Board of Family Medicine (ABFM). Individuals from ABIM, ABMS, and ABFM provided feedback on the overall goals of the review and on earlier drafts of the manuscript, but approval of the final manuscript rested with the authors alone.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interest
The authors were not involved with the peer review process of this work.

Received: 1 March 2022 Accepted: 27 September 2023
Published online: 10 October 2023

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.

Agarwal, P. K., D'Antonio, L., Roediger, H. L. III, McDermott, K. B., & McDaniel, M. A. (2014). Classroom-based programs of retrieval practice reduce middle school and high school students' test anxiety. *Journal of Applied Research in Memory and Cognition*, *3*(3), 131–139.

Alpert, R., & Haber, R. N. (1960). Anxiety in academic achievement situations. *The Journal of Abnormal and Social Psychology, 61*(2), 207–215.

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.

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.

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.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajaras & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). Information Age.

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: Motivational psychology* (2nd ed., pp. 261–271). 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: Theory and application* (2nd ed.). 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.
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., & 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.
Bullock, J. L., Lockspeiser, T., Pino-Jones, A. D., Richards, R., Teherani, A., & Hauer, K. E. (2020). They don't see a lot of people my color: A mixed methods study of racial/ethnic stereotype threat among medical students on core clerkships. *Academic Medicine, 95*(11), S58–S66.
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. (2019a). 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. (2019b). 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.
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. (1999). *Self-theories: Their role in motivation, personality, and development*. Psychology Press.
Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.
Dweck, C. S., & Yeager, D. S. (2019). Mindsets: A view from two eras. *Perspectives on Psychological Science, 14*(3), 481–496.
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.
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.
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.
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.
Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology, 54*(1), 5–12.
Fassiotto, M., Li, J., Maldonado, Y., & Kothary, N. (2018). Female surgeons as counter stereotype: The impact of gender perceptions on trainee evaluations of physician faculty. *Journal of Surgical Education, 75*(5), 1140–1148.
Fencl, H. S., & Scheel, K. R. (2005). Research and teaching: Engaging students—An examination of the effects of teaching strategies on self-efficacy and course climate in a nonmajors physics course. *Journal of College Science Teaching, 35*(1), 20–24.
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.

Fraundorf, S. H., Caddick, Z. A., Nokes-Malach, T. J., & Rottman, B. M. (2023). *Cognitive perspectives on maintaining physicians' medical expertise: IV. Best practices and open questions in using testing to enhance learning and retention*. Principles & Implications.

Galla, B. M., Plummer, B. D., White, R. E., Meketon, D., D'Mello, S. K., & Duckworth, A. L. (2015). The academic diligence task (ADT): Assessing individual differences in effort on tedious but important schoolwork. *Contemporary Educational Psychology, 39*(4), 314–325.

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.

Greeno, J. G., The Middle-School Mathematics Through Applications Project Group. (1997). Theories and practices in thinking and learning to think. *American Journal of Education, 106*, 85–126.

Greeno, J. G., The Middle-School Mathematics Through Applications Project Group. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53*, 5–26.

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.

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.

Hall, J. W., Grossman, L. R., & Elwood, K. D. (1976). Differences in encoding for free recall vs. recognition. *Memory & Cognition, 4*(5), 507–513.

Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002a). 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. (2002b). 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. (2016b). 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 & T. Urden (Eds.), *Motivational Interventions, advances in motivation and achievement* (Vol. 18, pp. 71–105). Emerald Group Publishing.

Hattie, J., Hodis, F. A., & Kang, S. H. K. (2020). Theories of motivation: Integration and ways forward. *Contemporary Educational Psychology, 61*, 101865.

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* (pp. 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* (pp. 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.

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: Theory and application* (2nd ed.). Guilford Press.

Jiang, Y., Rosenzweig, E. Q., & Gaspard, H. (2018). An expectancy-value-cost approach in predicting adolescent students' academic motivation and achievement. *Contemporary Educational Psychology, 54*, 139–152.

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.

Klein, S. B. (2019). *Learning*. Sage publications Inc.

Krapp, A. (1999). Interest, motivation and learning: An educational-psychological perspective. *European Journal of Psychology of Education, 14*(1), 23–40.

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.

Lave, J. (1988). *Cognition in practice*. Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.

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.

Leyens, J. P., Desert, M., Croizet, J. C., & Darcis, C. (2000). Stereotype threat: Are lower status and history of stigmatization preconditions of stereotype threat? *Personality and Social Psychology Bulletin, 26*(10), 1189–1199.

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., Durik, A. M., Conley, A. M. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., & Harackiewicz, J. M. (2010). Measuring situational interest in academic domains. *Educational and Psychological Measurement, 70*, 647–671.

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.
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.
Myers, S. P., Dasari, M., Brown, J. B., Lumpkin, S. T., Neal, M. D., Abebe, K. Z., Chaumont, N., Downs-Canner, S. M., Flanagan, M. R., Lee, K. K., & Rosengart, M. R. (2020). Effects of gender bias and stereotypes in surgical training: A randomized clinical trial. *JAMA Surgery, 155*(7), 552–560.
Mylopoulos, M., Kulasegaram, K., & Woods, N. N. (2018). Developing the experts we need: Fostering adaptive expertise through education. *Journal of Evlauation in Clinical Practice, 24*(3), 674–677.
Nagengast, B., Marsh, H. M., Scalas, L. F., Xu, M. K., Hau, K., & Trautwein, U. (2011). Who took the "x" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science, 22*(8), 1058–1066.
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* (pp. 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.
Perez, T., Cromley, J. G., & Kaplan, A. (2014). The role of identity development, values, and costs in college STEM retention. *Journal of Educational Psychology, 106*(1), 315–329.
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.
Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement, 53*, 801–813.
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.
Putwain, D. W., von der Embse, N. P., Rainbird, E. C., & West, G. (2021). The development and validation of a new multidimensional test anxiety scale (MTAS). *European Journal of Psychological Assessment, 37*(3), 236–246.
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* (pp. 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 volume in the educational psychology series. Development of achievement motivation* (pp. 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* (pp. 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.
Rosenzweig, E. Q., Song, Y., & Clark, S. (2022). Mixed effects of a randomized trial replication study testing a cost-focused motivational intervention. *Learning and Instruction, 82*, 101660.
Rosenzweig, E. Q., Wigfield, A., & Hulleman, C. S. (2020). More useful or not so bad? Examining the effects of utility value and cost reduction interventions in college physics. *Journal of Educational Psychology, 112*(1), 166–182.
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, 25*, 101860.
Sarason, I. G. (Ed.). (1980). *Test anxiety: Theory, research, and applications*. 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., Wild, K. P., & Krapp, A. (1995). Course-specific interest and extrinsic motivation as predictors of specific learning strategies and course grades. In *6th EARLI conference in Nijmegen*.
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* (pp. 183–212). Lawrence Erlbaum Associates Inc.
Schmidt, S. R. (1988). Test expectancy and individual-item versus relational processing. *The American Journal of Psychology, 101*(1), 59–71.
Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. S. Eccles (Eds.), *A volume in the educational psychology series. Development of achievement motivation* (pp. 15–31). Academic Press.
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.
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.
Spencer, S. J., Logel, C., & Davies, P. G. (2016). Stereotype threat. *Annual Review of Psychology, 67*, 415–437.
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* (pp. 3–14). Taylor & Francis.
Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613–629.
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.
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.
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. (2018). 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). 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). 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., Paunesku, D., Walton, G. M., & Dweck, C. S. (2013). How can we instill productive mindsets at scale? A review of the evidence and an initial R&D agenda. In *White paper prepared for white house meeting excellence in education*. The Importance of Academic Mindsets.
Yeager, D. S., Hanselman, P., & Dweck, C. S. (2019). A national study reveals where a growth mindset improves achievement. *Nature, 573*, 364–369.
Yeager, D. S., Romero, C., Paunesku, D., Hulleman, C. S., Schneider, B., Hinojosa, C., Lee, H. Y., O'Brien, J., Flint, K., Roberts, A., Trott, J., Greene, D., Walton, G. M., & Dweck, C. S. (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), 374–391.
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*. Plenum Press.
Zeidner, M. (2009). Test anxiety. In I. B. Weiner & W. E. Craighead (Eds.), *The corsini encyclopedia of psychology* (pp. 1–3). Wiley.

Publisher's Note