A Vision for Certification in Internal Medicine in 2020

Assessment 2020 Task Force
September 2015

Recommendations to the American Board of Internal Medicine (ABIM): A Vision for Certification in Internal Medicine in 2020

ASSESSMENT 2020 TASK FORCE

Harlan Krumholz (Chair)	Yale University
Richard Baron	American Board of Internal Medicine
Lee Berkowitz	University of North Carolina (Chapel Hill)
Jack Boulet	Educational Commission for Foreign Medical Graduates
David Coleman	Boston University
Ezekiel Emanuel	University of Pennsylvania
Kevin Eva	University of British Columbia
Ted Eytan	Kaiser Permanente
David Johnson	University of Texas Southwestern (Dallas)
Rebecca Lipner	American Board of Internal Medicine
Marilyn Mann	Patient Advocate
William McGaghie	Loyola University Chicago
André A. Rupp	Educational Testing Service
Martín J Sepulveda	IBM Corporation
Candace Thille	Stanford University Graduate School of Education
Abraham Verghese	Stanford University School of Medicine
Robert Wachter	University of California San Francisco
Patrick Alguire *(Ex-officio)*	American College of Physicians
Steven Durning *(Ex-officio)*	Uniformed Services University

Acknowledgments

On behalf of the Assessment 2020 Task Force, we submit these recommendations to the American Board of Internal Medicine (ABIM) Board of Directors and Council.

This Task Force, initiated almost two years ago, brought together a remarkable interdisciplinary group of individuals who devoted themselves to learning about each other's fields, studying leading advances in certification across professions and envisioning a future that would be better able to ascertain knowledge and skills, more efficient in its administration and more enjoyable in its experience. I owe a great debt of gratitude to every member of this Task Force. Each of them participated in the hope that the joint product would improve medical care and relieve the burden of assessment. I also wish to thank members of ABIM staff, in particular, Rebecca Lipner for her wisdom and insight as my principal partner in leading this effort; and Bryn Herrschaft and Helene Brooks for all their efforts administering the work of the Task Force; and Lorie Slass, Lisa Miller, Jenny Brosseau and Leslie Tucker for their work on the Assessment 2020 website and outreach efforts. I also would like to particularly thank Dr. Robert Wachter, whose idea it was to establish this Task Force while he was Chair of the ABIM Board of Directors. Finally, I thank the ABIM Board for the privilege of leading this group.

It is important to note that the Task Force recognized that much work would be required to determine the feasibility of these recommendations and to develop implementation plans. The group sought to envision what the future could be and considered that what is possible tomorrow may be very different from what can be done today.

The spirit of this document is to stimulate discussion about the future state of certification. Society and the profession need a credible, trustworthy and valid approach to assessing the ability of physicians to provide excellent clinical care over the course of their careers. Certification is part of that accountability and can play an important role in being a fair arbiter, a source of information to the public and a catalyst for continual improvement. Advances in the science of assessment and changes in medical practice require us to continually question and renew our approaches. These efforts are ongoing, but there are junctures in which a deep look can be beneficial, or even transformational. We hope that our efforts here constructively contribute to our aspirations to be ever better and may play a role in positive change that is welcomed by the profession and the public, and advances considerably the value of certification in medicine.

Sincerely,

Harlan M. Krumholz, MD
Chair, Assessment 2020 Task Force

Table of Contents

EXECUTIVE SUMMARY: A Vision for Certification in Internal Medicine in 2020

In order for ABIM to fulfill its mission "to enhance the quality of health care by certifying internists and subspecialists who demonstrate the knowledge, skills and attitudes essential for excellent patient care," it must keep pace with evolving trends, adapt to new circumstances and embrace the latest science. As a result, in 2013, the ABIM Board of Directors commissioned a Task Force called *Assessment 2020* to develop a vision for the future of assessment for certification (initial and maintenance) in internal medicine and associated subspecialties.

The effort was designed to be forward-thinking, taking into account what might be possible in the near future and the direction of advances in medical practice, technology, cognitive psychology, performance and skill assessment, and pedagogy. The work was guided by the needs of patients and society while also being attentive to the burden on and benefit to physicians. The desired program was envisioned to be iterative, adaptive, feasible, valid, defensible, and one that would drive learning and incentivize excellence. As a result, physicians would find the program relevant, engaging and efficient, and the public would find it informative and useful in assessing physicians.

The recommendations were not constrained by what is currently possible. The Task Force sought a diversity of perspectives. Comments were actively solicited from stakeholders (including patient groups, physicians, health plans and insurers). Moreover, the ABIM Board Certified physicians were informed of the initiative and invited to provide comments through the Assessment 2020 website. Blog posts and polls were commissioned by Task Force members and experts in the field of assessment to start the conversation with the community.

The Task Force relied on established frameworks for what makes a good doctor (ACGME competencies[1]) and what makes a good assessment (van der Vleuten's utility framework[2] and Kane's validity argument[3]). The group explored the principles and values that are appropriate for certification, the competencies physicians will most likely need to practice in the near future, the approaches that might be sufficiently rigorous to assess performance in these competencies, and innovations currently in the research and development phases at ABIM and elsewhere.

The recommendations that derive from the work of the Task Force are as follows:

1) **Replace the 10-year Maintenance of Certification Exam with More Frequent, Less Burdensome Assessments.**
   The Task Force recommends replacing the 10-year Maintenance of Certification (MOC) exam with more frequent assessments that could be taken at home or at the workplace.

The new format would be designed to assess competency in essential contemporary knowledge. Some aspects of the assessment would be "open-book" and some would represent knowledge that should be known without outside references, but specified in advance by the profession. The results of the smaller, more frequent, lower-stakes assessments would provide insight into performance and accumulate over time and culminate in a high-stakes pass/fail decision. A failure at this point may necessitate taking a longer exam or another form of assessment in order to maintain certification. This approach would emphasize learning as an integral part of the program, but would also provide meaningful criteria to the public as to whether a physician is remaining current. The existing self-assessment component of MOC would likely no longer be a separate requirement as the new exam format would provide knowledge assessments on a frequent enough basis to obviate the need for it.

2) **Focus Assessments on Cognitive and Technical Skills.**
The Task Force recommends that ABIM focus on its MOC efforts on assessing cognitive and technical skills relevant to the practice of internal medicine. The rationale is that there are specific competencies in these domains that are unique to the internist and that may degrade over time. In addition, there are rigorous and scalable assessment methods that are currently available or will be available soon to measure these competencies. Assessment of cognitive skills will assure the public that physicians are keeping up with the clinical knowledge that is relevant to patient care. Assessment of technical skills will assure that physicians can apply that knowledge to adequately perform the technical procedures. ABIM should continue to focus on developing assessments of these competencies that closely align to actual practice through innovative approaches.

Other competencies, such as communication, teamwork, empathy and quality improvement, are also vital for effective patient care, but formal assessment of them for practicing physicians is challenging. These skills have some special attributes. They may be context dependent in that the systems and teams may influence the ability of an individual to demonstrate them. Merely participating in programs such as those focused on quality improvement, although important, may not indicate meaningful performance of such activities. Direct observation may be critical for assessing competence. The Task Force recommends that ABIM should continue to include the demonstration of these skills as part of initial certification requirements as these are assessed in a standardized and uniform way in training programs and under direct observation. However, the best approach to assess these skills at the individual level outside of a training program is not clear. ABIM should continue to emphasize the importance of these skills and encourage health care organizations to promote and assess these skills locally and in context. As methods emerge that are effective and efficient—that can account for context and convey meaningful information without undue burden—ABIM should re-evaluate its role in assessing these competencies.

3) **Recognize Specialization.**

   The Task Force recommends a movement toward certification in specialized areas, doing so without the need for underlying certificates (e.g., Cardiovascular Disease is the underlying certification for Interventional Cardiology). Thus, an underlying certification should no longer be required for *maintaining* certification in subspecialty areas that currently require them. The subspecialty area would stand on its own for MOC. A natural extension of this recommendation includes recognition of additional specialization in relevant practice areas. ABIM will need to consider feasible approaches to recognize these focused areas of practice. In doing so, a strategy for how to represent the scope of practice to the public will be essential. Ultimately, the goal is for the customization of MOC, so that it represents an individual's practice and is appropriately transparent and meaningful for the public.

I. INTRODUCTION

A. Charge to the Assessment 2020 Task Force

ABIM's mission is "to enhance the quality of health care by certifying internists and subspecialists who demonstrate the knowledge, skills and attitudes essential for excellent patient care." To fulfill this mission and remain current, ABIM must stay current with evolving trends. To that end, the following charge was given to the Assessment 2020 Task Force in 2013 by the ABIM Board of Directors to help guide ABIM into the future.

1. Charge

The Task Force is assembled to develop a vision for assessment for certification (initial and maintenance) in internal medicine and associated subspecialties for the near future. Taking advantage of advances in medical practice, technology, cognitive psychology, performance and skill assessment and pedagogy, this team will develop an approach to cognitive assessment that produces a relevant, valid and reliable assessment process for the future. The group will also leverage the experience of other professions in evaluating performance and competence along with its own experience as an assessment organization. Although the focus is on internal medicine, the approach ought to be applicable to other areas of the profession.

The vision for the future will take into account the following: scientific validity, face validity, technical feasibility, financial viability – and likelihood that the approach will improve medical practice, provide confidence in medical practitioners and produce a safer, more effective, more efficient, more equitable, more patient-centered and systems-based health care system. Additionally, the approach should be understood and acceptable to physicians.

The charge is both strategic and tactical in that practical considerations should guide the strategic vision. The group should consider various perspectives but be principally guided by the needs of patients and society, while also being attentive to the burden on and benefit to physicians. The time horizon is the near future, with a goal for implementation by 2020. The group will be informed by past approaches to assessment, but is not constrained by current methods being employed by ABIM.

2. Methods

The Task Force members were chosen to have expertise in diverse backgrounds including medicine, technology, cognitive psychology, assessment, education, health policy and patient advocacy.

The process used by the Task Force was designed to be open and inclusive. We engaged in a conversation with the community by actively seeking comments from key stakeholders including patient groups, physician groups, physicians, health plans and insurers. From the outset of the project, ABIM physicians were informed of the initiative and invited to provide constructive ideas and comments through the Assessment 2020 website (http://assessment2020.abim.org). To further encourage a thoughtful conversation with our stakeholders, blog posts and polls pertaining to assessment issues were written by Task Force members and experts in the field of physician assessment. Current ABIM research and development efforts related to assessment innovations (e.g., more detailed performance feedback reports for the high-stakes exam) were made transparent on the website to inform this open conversation.

Two background white papers were commissioned to inform the work of the Task Force. The purpose of the first paper, which is included in Appendix A, was to more formally evaluate the perspectives of thought leaders in health care including patients, physicians, health care administrators, educators and payers on the future of medicine. The purpose of the second paper, which was published in *Academic Medicine* in October 2015 and included herein as Appendix G, was to address research comparing the effects of open- versus closed-book exams.

Additionally, a review of a portfolio of potential assessment methodologies from both inside and outside of the medical profession was performed. The goal of the review was to understand the extent of their evidence base in terms of the validity of the assessment and to gauge how that might be applied to enhancing ABIM's assessment programs. Finally, a formal evaluation of the methodologies specifically used for assessing cognitive and technical skills was done by evaluating their complexity, cost and value.

B. Background on the Evolution of Assessment 2020

The Assessment 2020 initiative was commissioned in 2013 by the ABIM Board of Directors to begin to address how ABIM could adapt to changes in medicine and assessment while incorporating knowledge from other professions to facilitate change.

Historically, ABIM's certification efforts have been best known for high-stakes cognitive assessment (also referred to as the "secure exam"). This current analysis of ABIM's certification programs is specifically designed to evaluate a physician's cognitive skills in a broad domain of internal medicine or its subspecialties. The focus of the cognitive assessment has been on testing clinical judgment, not factual recall. The exam content is based on a detailed blueprint developed by a committee composed of practicing physicians whose experience covers the breadth of the discipline. The questions are presented as patient vignettes in settings that reflect current medical practice. By following the Standards for Educational and Psychological Testing[4] (a well-established professional consensus concerning appropriate and fair test use that is based on psychometrics, the science of assessment), ABIM has consistently produced exams that are rigorous and fair assessments of cognitive skills.

Prior to 1990, certification was completed once in a physician's lifetime. With research indicating that physicians' clinical skills tended to decline over time,[5] that an individual physician's ability to independently and accurately self-assess was poor,[6] and that few physicians were examining their own data from practice,[7] a new initiative eventually known as Maintenance of Certification (MOC) was introduced. This new initiative required physicians to recertify once every 10 years after initial certification as a way of demonstrating to the public that they had kept up with changes in medicine.

Since the introduction of MOC, several important developments have occurred:

1. American Board of Medical Specialties Introduces 2015 standards
In 2012, the American Board of Medical Specialties (ABMS) began establishing 2015 Maintenance of Certification (MOC) standards for all ABMS Boards (http://www.abms.org/media/1109/standards-for-the-abms-program-for-moc-final.pdf) that would continue to incorporate additional competencies other than cognitive skills into the requirements but also make MOC a more continuous program rather than a 10-year cycle. The MOC program was created to have an integrated four-part framework that addressed 1) Professional Standing and Professionalism; 2) Lifelong Learning and Self-Assessment; 3) Assessment of Knowledge, Skills and Judgment; and 4) Improvement in Medical Practice.

The standards for ABMS Programs for MOC are common across the ABMS Member Boards while permitting appropriate distinctions in programs across individual Member Boards. Three general goals that each ABMS Member Board's Program for MOC is asked to incorporate are:

- To include the 1999 six Accreditation Council for Graduate Medical Education (ACGME)/ABMS Core Competencies in the program: Practice-Based Learning & Improvement; Patient Care & Procedural Skills; Systems-Based Practice; Medical Knowledge; Interpersonal & Communication Skills; and Professionalism.
- To enhance the value of its Program for MOC and the experience of physicians engaged in its Program including taking actions to increase the Program's quality, relevance and meaningfulness with sensitivity to the time, administrative burden and costs (monetary and other) associated with participation.

- To engage in continual quality monitoring and improvement of its Program for MOC and participate in the ABMS MOC Review Process.

As ABIM was working to put the ABMS 2015 standards into place, changes occurred in the practice of medicine, the field of assessment and the environment.

2. Changes in Medicine

The most profound change in the practice of medicine is a direct result of the fact that decisions about high-quality patient care have become more complex. Many factors must now be considered by physicians in making recommendations for high-quality patient care. These decisions must consider risk, patient preference with shared decision-making, and response to treatment. New areas, such as precision medicine and the use of a patient's genetic profile to help customize decisions for prevention, diagnosis and treatment, are examples of this complexity. Additionally, decisions for care are more influenced by context, the increasingly important clinical role of allied health professionals, digital tools such as electronic health records and mobile devices, and the explosion of information available through the proliferation of journal articles, evidence-based clinical support systems and decision support systems.

Furthermore, an increased focus has been placed on patient-centered care, on directly measuring and improving the quality of patient care, and on physicians playing a more active role in reducing the cost of health care.[8] As a result, physicians have been challenged with playing critical—but not yet fully defined—roles related to the full experience of care for each patient, the costs of that care and the results.[9]

Concurrently, the rates of development of information technology are expected to continue to increase, with rapid development of tools for health care systems, providers and patients, such as clinical decision support (CDS) systems and mobile applications.[10] Clinical decision support systems may become sophisticated enough to manage large amounts of knowledge and yield smarter and more powerful diagnostic capabilities, yet evidence of their effectiveness is yet to be determined.[11] Patients may also use mobile applications to have greater access to medical information and to track their health and fitness data on a regular basis.[12]

As a direct result of these changes in the availability of technology, as well as the sheer abundance of journal articles and evidence-based clinical support systems, physicians are faced with an overwhelming amount of information. Consequently, physicians will likely need a set of unique competencies to understand this information and be able to apply it to clinical decision-making and the improvement of care quality and cost.[13]

3. Changes in Assessment

There also have been significant changes in assessment methodology for licensure and certification that have led to new developments in ways to measure the competence of physicians.[14] This evolution includes the shift from measuring the process to measuring the outcomes (in which, depending on the outcome, physicians may play only a supporting role), the need for continuous learning throughout a physician's career, and rapid changes in technology and psychometrics. The shift to competency-based medical education recognizes that assessment is part of an ongoing learning experience where the emphasis is not only on medical knowledge but also on what a physician should actually be able to do in practice.[15] As a result of this shift, assessment systems designed to measure competency must meet growing requirements including adaptability, continuity and comprehensiveness, and be able to address the progression of physician skills in the clinical workplace.[16]

While multiple-choice questions (MCQs) remain an excellent method of assessing cognitive expertise, advances in technology allow for processing different methods of testing including simulations which mimic real-world practice more closely and are used across disciplines to assess, for example, airline and submarine pilots. Natural language processing and automated scoring techniques offer scalable approaches to measuring the clinical reasoning process and not simply whether a physician is able to pick out the correct answer from a list of options. It is likely that the future of assessment will rely more on these multifaceted approaches, especially when evaluating complex skill sets such as communication or clinical reasoning.

Even simple enhancements to typical MCQs may better assess higher order skills like clinical reasoning. It is possible to focus questions on common misconceptions and undesirable actions. New approaches could allow for open-book exams (a form of which is used on the Certified Public Accountant exam) to acknowledge the abundance of information that cannot be memorized. Different item types including short-answer responses or compact mini-performance tasks could be used to assess higher order skills.

Advances in assessing patients' experience of care (i.e., the participation of patients in decisions of care and respect and understanding for their beliefs, values, concerns, preferences and their understanding of their condition) seem promising. These include validated patient surveys such as Consumer Assessment of Healthcare Providers and Systems (CAHPS), Objective Structured Clinical Examinations (OSCEs) that would examine the patient experience of care in a standardized but richer context, and audio or video recordings of live patient-physician encounters. Likewise, to assess an individual's role in a team, a multi-source feedback approach or direct observation of how the individual functions in a team (used in business to evaluate leadership performance) in a live or simulated context is promising. To make the scoring of these assessments more scalable, new approaches to rating patient-doctor encounters that do not involve expert raters are being studied for their effectiveness and rigor in educational settings.

4. Changes in the Environment

As the Task Force started its work, several other changes occurred that made the work of Assessment 2020 even more vital than it had previously been. ABIM changed its once-every-10-year Maintenance of Certification (MOC) program to a more continuous one that resulted in much criticism among internists and medical specialty societies. The criticism focused on the burden of the new requirements and concerns that the program was not as relevant and meaningful as it aspired to be.

It was fortuitous that the Task Force was already in place with the intent to be forward-looking and proactive in nature. While these critical sentiments developed, we were simultaneously listening to constructive feedback from physicians and patients to the extent that we were able to incorporate their concerns into our vision of the future. As colleagues in the medical community, it was important to be responsive to these voices, and this initiative cannot succeed without such constructive input.

In sum, the practice of medicine is evolving, as is the science of assessment, both being influenced heavily by advances in technology. These developments have questioned the way we think about the knowledge, skills and attitudes that physicians will need in order to practice effectively in the future and how our assessments will need to change accordingly. We interviewed thought leaders in health care to help us better understand how the health care system may change in the next decade, and we evaluated a portfolio of assessments to better understand the most promising approaches to assessment.

II. FRAMEWORK: COMPETENCIES & ASSESSMENTS

A. What Makes a Good Doctor

While scholars continue to investigate the attributes of a good physician, and all agree that cognitive skills are essential to being a successful physician, global consensus is growing that being a good physician entails a number of additional competencies that reflect important attributes of successful practice. In the United States, six competencies were first adopted in 1999 by the Accreditation Council for Graduate Medical Education (ACGME) and American Board of Medical Specialties (ABMS).[14]

Support for the use of these ACGME/ABMS competencies is based in both empirical evidence as well as theory.[17,18] The evidence includes, but is not limited to, patient outcomes, physician performance and physician quality improvement.[15] In addition, a recent literature review concludes that physicians who are board certified generally provide better patient care.[18] Given this convergence of evidence, the adoption of the competencies by various constituents in the U.S., as well as the absence of alternative frameworks with superior characteristics, the Task Force concluded that the use of ACGME competencies should guide its work. The six competencies are briefly described below:

1. **Medical Knowledge**: Demonstrating knowledge about established and evolving sciences relevant to the care of patients and their application, and ensuring that this knowledge does not degrade over time.

2. **Patient Care and Procedural Skills:** Providing care that is compassionate, appropriate and effective. This includes diagnosis and treatment of active conditions and promoting health.

3. **Interpersonal and Communication Skills**: Demonstrating skills that result in effective information exchange (communication) as well as teaming with patients, their families and professional associates (e.g., fostering a therapeutic relationship that is ethically sound, demonstrating skills with both non-verbal and verbal communication; serving as both a team member and at times as a leader).

4. **Professionalism:** Showing a commitment to carrying out professional responsibilities, adherence to ethical principles and sensitivity to diverse patient populations. This competency embodies a physician’s promise of duty and expertise.

5. **Practice-Based Learning and Improvement:** Showing an ability to investigate and evaluate patient care practices, appropriately appraise and use scientific evidence, and improve the practice of medicine. This competency emphasizes the commitment to ongoing quality assurance in care of patients.

6. **Systems-Based Practice:** Demonstrating awareness of and responsibility to the larger system of health care. An ability to understand and use system resources to provide optimal care (e.g., coordinating care across geographic sites or serving as the primary patient manager when care involves multiple specialties).

B. What Makes a Good Assessment Program

Developing an effective assessment program entails compromises that vary for each specific assessment context.[2] A framework for what makes a good assessment must consider several aspects of the assessment including its purpose, its design and its value.

1. Purpose of the Assessment

a) Assessment *for* learning versus assessment *of* learning. In designing an assessment, it is important to clearly define its goal.

The goal of an assessment *for* learning is to provide ongoing feedback to learners about their strengths and weaknesses so they can target areas that need improvement. The goal of an assessment *of* learning is to measure the attainment of a certain level of knowledge at a particular point in time and compare it to a benchmark. That said, the output from an assessment *of* learning can also be used for improvement. ABIM uses an assessment *of* learning to decide whether a physician has demonstrated whether they have acquired enough knowledge to be board certified. Both assessment designs have been shown to drive the learning process.

b) High-stakes versus low-stakes assessment. What distinguishes a high-stakes assessment from a low-stakes assessment is not how it was designed but how the results are used (i.e., the consequence). If the results of the assessment are used to determine an important outcome, such as whether one graduates from college or achieves certification, the assessment would be considered high-stakes since there is a significant consequence of achieving or not achieving the goal. If the results are made public, then the consequences are even more critical. By contrast, an assessment that carries little significant or public consequences (e.g., only the examinee is aware of his/her results) would be considered low-stakes.

c) Verification of the exam taker. Verification of the identity of the individual taking the assessment plays a crucial role in the legitimacy of the inferences we can make from the individual's assessment. Verification can be done using human proctors at the testing site, remote proctoring where proctors are viewing examinees live, or an audit that typically occurs after the administration by reviewing videos of the testing experience. If the goal is to make a consequential statement about an individual, it is critical to know that the statement is based on performance that can be reliably and confidently attributed to that individual and not someone else.

2. Design of the Individual Assessment

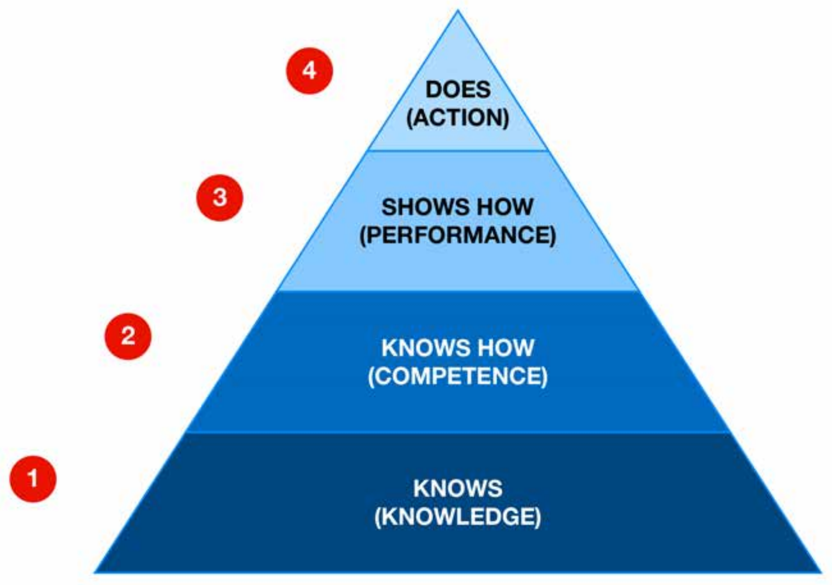
The design of the individual assessment follows from the purpose of the assessment but is often limited by cost and feasibility of administration or implementation. In designing an assessment, we need to consider the way in which the construct is measured – does the physician's behavior need to be observed (performance-based assessment) or should the physician's thought process be evaluated (cognitive assessment)? For the assessment to be credible, we also need to factor in the assessment frequency, e.g., how often the physician needs to be assessed in order to ensure continued competence. Other factors to consider include the length of the assessment, which is important in being able to make reliable decisions. Item formats are also critical to the design as we may choose multiple-choice items, essays, standardized patients, workplace-based assessments and whether the testing experience should allow access to external resources (i.e., open- versus closed-book assessments).

3. Value of the Assessment

In determining the value of assessments (i.e., they are supposed to be relevant, useful, meaningful and efficient) we apply van der Vleuten's model of utility[2] as well as Kane's well-respected argument-based approach to validation.[3] The van der Vleuten model, described below, considers two important psychometric characteristics—reliability and validity—as well as three other central factors: educational impact, acceptability and cost. All elements are considered when evaluating the value of the assessment, but some are more important (or weighted more) than others depending on the purpose of the assessment. If the assessment has high-stakes consequences for the examinee, reliability and validity should be weighted more heavily, in part because legal defensibility is critical for the assessment organization.

a) Reliability. When we measure a skill or attribute, if we sample across conditions appropriately, we can achieve reproducible scores and decisions that will allow evaluators to be confident in high-stakes decisions that are being made about a physician. Reliability is always about the consistency of measurement across occasions, such as different forms of the test and different administration dates.

b) Validity. An assessment should measure the construct or trait it purports to measure so that interpretations and decisions on the basis of assessment results are defensible. In this context, the authenticity of the measurement and its relevance to the physician's practice are important aspects of validity. Miller's competency pyramid[19] is helpful in thinking about assessments that become more authentic as one moves from testing what a physician "knows" (e.g., via multiple-choice questions) to testing what a physician "knows how" to do (e.g., via complex, multiple-choice case simulations), to testing what a physician "shows how" to do (e.g., via performance simulations) and, finally, testing what a physician "does" (e.g., via structured observations in practice). However, these assessments also move from being able to sample broadly from the content of the domain to only being able to sample narrowly due to high costs and feasibility of administration or implementation at the top of the pyramid. Building reliable and valid assessments that move through these stages presents many challenges.


Theory: Miller's Framework for Clinical Assessment (1990)

c) Educational impact and value of feedback (usefulness). There is increasing evidence that all assessments drive learning to some degree. Good performance feedback can be quite valuable in identifying areas in need of improvement and should be provided regardless of the purpose of the assessment.[20]

d) Acceptability**.** The acceptability of the assessment has to do with its perceived fairness, its fidelity to practice, the level of anxiety it provokes and whether the experience drives the learning process.

e) Feasibility/Cost. For both developers and consumers of the assessment, the feasibility and cost are important considerations. For example, if an assessment lasts for three days in order to produce a reliable measure of performance, that assessment is likely to be too costly and time-consuming for the individual. In addition, an assessment that includes multiple performance components is likely to be expensive to administer (e.g., result in high human scoring costs) or to develop (e.g., when immersive digital environments need to be designed to reduce human scoring needs).

Kane's argument-based approach to validation[3] underscores the fact that interpretations of the performance on an assessment and associated decisions about stakeholders are valid/defensible if they are stated clearly and there is trustworthy evidence to support them in light of possible alternative explanations. Like van der Vleuten's utility model, [2] the evaluation of whether interpretations and decisions are defensible is dependent on the stakes of the decisions, best practices in the discipline and stakeholder values. Evidence that is brought to bear for this purpose is typically collected over a series of pilot studies and involves how well the assessment meets its claims in terms of scoring inferences (are the scores meaningful), generalization of findings (would we get similar results if the assessment were repeated), extrapolation to real-world performance (do the scores relate to the traits we are trying to measure), and decision inferences (are the decisions we make from the scores meaningful).

4. Design of the Assessment Program

We also need to consider the design of the entire assessment program that would incorporate a variety of assessments whose purpose may be to assess different competencies. The program could include a profile of a physician's performance on these various competencies. This may entail a combination of some lower stakes assessments for competencies whose purpose is for learning and higher stakes assessments for competencies where thresholds of performance should be met. Multiple methods of assessment on multiple occasions typically provide a richer environment in which to base important decisions about an individual physician.

III. WHAT WE HEARD: SKILLS & ASSESSMENTS FOR THE FUTURE

As part of the Assessment 2020 initiative, we listened to our stakeholders (physicians, consumer groups and health systems) to gain an in-depth understanding of how they view the value and design of our assessments and the future competencies that will be needed to provide quality patient care. In addition, we incorporated input about Maintenance of Certification (MOC) through data that are routinely collected and analyzed by ABIM, including the post-exam surveys and surveys filled out by the physicians upon completion of self-directed assessments. Likewise, social media outreach, including the Assessment 2020 website, blogs and Twitter, engaged our community of stakeholders on the future of physician assessment and the MOC program. Interviews with 28 thought leaders and eight public interest/consumer advocacy groups proved to be quite informative as was the analysis of diplomate surveys completed following engagement in ABIM assessment products. Response on social media was steady but marginal since efforts to promote Assessment 2020 were scaled back due to the launch of the continuous MOC program in early 2014 and then the program changes in early 2015.

ABIM researchers also conducted in-depth interviews with “thought leaders” in health care – individuals who would likely have an original and important perspective on how the U.S. health care system may change in the coming decade and how these changes would impact physicians and the role of ABIM. These interviews are summarized in a white paper (Appendix A). Targeted outreach calls with important consumer groups were also conducted to incorporate the patient voice into the discussion of the future of certification. Through these efforts, the Task Force was able to develop a robust picture of how the community viewed the skills that physicians would need in the future to deliver quality patient care and how to assess them. In this section, we present an aggregated summary of our different sources of feedback. More detailed information concerning the health care thought leaders interviews (Appendix A), social media outreach (Appendix B), and physicians’ exam and product feedback (Appendix C) is attached for your reference.

A. Values

One of the most common themes we heard about assessment of the future involved the need for increased alignment of the assessments to actual practice. Physicians have frequently expressed a desire for the high-stakes exam to reflect more closely what they do in practice. Physicians find that sometimes the clinical issues presented on the exam are not reflective of the clinical issues they encounter in practice. This feedback is not exclusive to the exam, as physicians have expressed a desire that other parts of MOC, such as the practice assessment component, be more relevant to the work they do. Furthermore, some consumer groups also expressed a need for more workplace-based assessments to reflect real practice.

Physicians claim that they want less "busy work" and more opportunities to participate in activities that are useful, efficient and engaging. Some physicians feel that the work done for MOC places an unfair burden on their time, and ABIM should focus on reducing this burden. Suggestions include reducing redundancy for program requirements since other entities such as the health care system may be measuring the same competency. Physicians also reported that they often lacked staff resources to collect data for parts of the existing MOC program or were forced to duplicate data that already existed in their practice. Thought leaders suggested that the MOC program could find ways to leverage existing data from electronic health records (EHRs) and other collective data sources in order to effectively relieve physicians of the noticeable burden of data collection and data entry. Seamless transfer of data may make the practice assessment component a more acceptable part of the MOC program in the future.

Overall, physicians stated that they valued the parts of the program that helped them to recognize a weakness in their clinical knowledge and practice and empower them to improve in those areas. Physicians, however, expressed concern whether there is substantial evidence demonstrating that a physician's completion of the MOC program would yield improved quality of patient care. Some consumer groups expressed the need for a program that engages physicians in continuous improvement of their practice and the delivery of quality care to their patients. Consumer groups also expressed the opinion that giving physicians the ability to make modifications to their practice with a measurable impact on patient outcomes is one of the most important values for physician assessment.

B. Competencies

Major shifts in the U.S. health care system identified by thought leaders may necessitate physicians developing particular sets of skills that go beyond traditional cognitive and technical skills. Some of these new areas agreed upon by thought leaders included doctor-patient communication, teamwork and the use of emerging technologies.

1. Cognitive/Technical Skills

The broader community (i.e., respondents who were neither the thought leaders nor ABIM diplomates) felt strongly that the more traditional physician competencies would continue to be important in the future, including cognitive and technical/procedural skills. However, they thought that although physicians might know the right thing to do if asked, they might not always be able to perform it effectively and accurately. For example, for cardiac catheterization, knowing the right stent size and position does not mean that the stent was inserted properly. They indicated a need for more workplace-based assessments or simulated practice environments for technical skills. Others also believed that a physician's skills in taking a patient's history and completing a physical exam would continue to remain important competencies. A consensus, however, around the continued importance of a traditional competency in diagnostic skills was less clear. While some members of the community and thought leaders felt that diagnosis will become less important for physicians with the advancement of medical decision support systems, others concluded that diagnostic skills will continue to be important since it is unlikely technology will be able to entirely substitute for the physician.

2. Doctor-Patient Communication

Another area of unanimous agreement by thought leaders was that patients would become more health literate and have higher expectations for shared decision-making. As a result, physicians would need to focus increasingly on nontraditional competencies including a deeper understanding of how a patient's environment impacts health and wellness. This would include more effective communication and engagement, understanding of personal needs and preferences, and the ability to adapt a treatment plan that incorporates the patient's values and preferences while still addressing clinical needs.

Feedback from the broader community and targeted consumer groups also placed significant importance on patient engagement and shared decision-making skills. The community recognized a need for physicians to have skills in careful listening and understanding of the patient's goals and values.

3. Teamwork

Thought leaders were also unanimous in predicting that physicians and other health care providers would need to work together as part of organized inter-professional teams to increase both quality and efficiency of care. These teams may exist in the physical environment of a common practice or in virtual settings through online collaboration that will become more commonplace as technology advances even further. As a result, physicians will need specific skills in working cooperatively with others, delegating tasks to other team members and developing values and attitudes reflective of a collaborative environment. A few thought leaders also foresaw a potential need for physicians to relinquish the value of forming a strong personal relationship with their patients and to rely on other health care providers to maintain direct contact with patients. This scenario would result in less focus on delivering direct patient care and more focus on managing teams and designing a system for care delivery involving other health care providers in direct contact with patients.

The broader community thought it was possible to assess an individual physician on teamwork skills and ABIM could possibly accomplish this through assessments like simulated clinical situations and observations of physicians working in teams.

4. Use of Technology

One of the potential new areas cited by thought leaders was the ability to apply technology to patient care. Thought leaders agreed that embracing information technology, including the effective use of electronic health records (EHRs), keeping up with rapid advances in technology and adapting their practices to new forms of data and tools will become even more important skill sets for physicians than they are today.

Social media outreach engaged the broader community to think about whether or not competency in technology would be important for physicians in the future. There was a general consensus, although not surprising for those who use social media regularly, that physicians will need to demonstrate their technological competency in the future, but that technology would be unlikely to replace the role of the physician. Being able to retrieve information efficiently and to effectively utilize EHRs in practice were two critical skills noted. Other skills mentioned include the use of mobile technology to manage and improve patient health in practice. However, others indicated that many patients would continue to want to meet with physicians face-to-face.

C. Assessment Methods

Some specific ideas related to assessment methods including adaptive testing, open-/closed-book design, simulations and other methods of assessing some potential new competencies were posed to the broader community on social media to solicit feedback about their utility and value.

Feedback on the debate between designing an open- versus closed-book assessment was robust. Though the ability to identify information has rapidly developed through technological advances, the broader community continues to believe that physicians should still maintain a set of core information without needing to consult external resources. However, physicians who engaged with us on social media felt that assessments should more closely mirror what physicians actually do in practice, which most commonly involves consulting external resources in an “open-book” fashion, at least on occasion, to provide care. Similar sentiments have been expressed by physicians through comments on MOC program surveys. Some physicians felt that for questions on the closed-book exam that focus on infrequently seen but important illnesses or uncommon patient populations, a physician in practice would most likely consult external resources (assuming it was not an urgent decision).

With this feedback in mind, it is apparent that physician and public stakeholders lean toward assessments that incorporate both open- and closed-book designs. Consequently, physicians would need to continue to demonstrate competency in core medical knowledge without using external resources, but also be able to demonstrate competency in accessing external resources and applying that information on an assessment.

Feedback on the value of adaptive testing, simulations and other potential assessments to measure new physician competencies was limited. Those who responded to targeted poll questions believed that the potential benefits of adaptive testing (shorter exams and less measurement error) make it worth implementing despite the added complexities. Simulations were recognized by stakeholders as a potentially valuable tool making use of more advanced technology to measure competency in areas like procedural skills. Unique methods of assessing some newer competencies, such as patient engagement and shared decision-making, were also posed as possibilities. The broader community believed that video recordings of a patient-physician interaction using either real patients or patient actors would be valuable in assessing a physician's skills in delivering patient-centered care using trained rater groups, comprised not only of physicians, but also non-physician health care professionals and patients.

IV. WHAT WE CONCLUDED

A. Values

We propose that the principles supporting certification and Maintenance of Certification (MOC) be mainly guided by the needs of patients and society while also being attentive to the burden on, experience of and benefit to physicians. A successful assessment program of the future should be valued by both physicians and the public. The program itself should be iterative, adaptive, feasible, valid, defensible and efficient, and one that drives learning. The process by which the program is developed and implemented should be transparent.

The program should acknowledge that a one-size-fits-all approach may not be appropriate in all situations. Yet, at the same time, high-quality assessments that are valid and trustworthy are still essential for maintaining the credibility of the program. The physician user should find the program relevant, engaging and efficient so that it reduces the burden placed on the physician while still fulfilling its purpose for both patients and society. The public should find the results of the program credible.

B. Competencies

The thought leaders identified a number of salient competencies that the Task Force also believes are important for providing high-quality patient care. These include cognitive and technical/procedural skills, doctor-patient communication skills, teamwork and the ability to effectively use technology (e.g., electronic health records and decision support systems). We believe that the use of technology will play a key role in the future (e.g., continued explosion of knowledge, decision support systems, electronic health care records), and quality improvement will become even more prominent in practice as physicians seek to provide timely, efficient, cost-effective and high-value care that is continuously improving. Indeed, the commitment to being a lifelong learner, which is a trait embodied by all the six core ACGME competencies, may be most essential for providing high-quality care in the future.

As the Task Force reflected on the competencies that were delineated by the thought leaders, three new issues emerged that needed further consideration. These issues include context-specific competencies, further specialization in medicine and the importance of assessing technical skills.

1. Differences Among Competencies

The Task Force recognized that all six ACGME competencies are important for providing high-quality patient care. However, the competencies have some special and different attributes. Some competencies, such as cognitive and technical skills, are unique to the internist and have rigorous and scalable assessment methods available. Physicians tend to accept that assessment of cognitive and technical skills are critical since these skills are likely to degrade over time as changes in clinical care occur. Attestation to the public about whether a physician's skills are current is an important service to patients. This particular effort is currently not being regulated for certified physicians by other entities in the health care system.

Other competencies, such as teamwork, communication and quality improvement, are more difficult to measure. They may be context dependent in that the health care systems and teams may influence the ability of an individual to demonstrate them. These competencies are also multidisciplinary in nature and not unique to internal medicine. Merely participating in programs such as those focused on quality improvement, although important, may not indicate meaningful performance of such activities. Direct observation may be critical for assessing competence. It is often difficult to determine the role of the individual physician in producing these outcomes. If there were standardized ways to effectively measure these competencies, we might be able to measure the characteristics of the individual physician that contribute to the performance of the system. Currently, a framework exists for assessing these multidisciplinary competencies during residency and fellowship training as program directors can attest to their mastery through observation of performance in that setting. As physicians exit the training environment and enter practice settings that are very distinct, the landscape becomes more complex and the approach for formally assessing these competencies is less clear.

However, the Task Force recognizes that these other roles that physicians play in making the health care system better are equally important as the cognitive and technical skills. In fact, the Task Force believes that these issues are beyond just internal medicine and are so fundamental to the culture of medicine itself that they need to be engaged in as broadly as possible. Therefore, the Task Force believes that ABIM should serve in a leadership role to work with and encourage health care organizations to create a consistent framework for assessing multidisciplinary competencies across the profession and do so locally and in context. This approach would avoid duplication of efforts and redundancy as we strive toward an integrated and cohesive system across medicine more generally. This partnership could have the power to transform the profession of medicine itself, where all physicians share common values, behaviors and standards rather than different specialty groups setting these standards in isolation. As methods emerge that are effective and efficient—that can account for context and convey meaningful information without undue burden—ABIM should re-evaluate its role in assessing these competencies.

2. Specialization

Although the group coalesced around the idea of measuring cognitive and technical skills for a discipline, the question arose about how broad each discipline would be in the future and what specific cognitive and technical skills should be measured. The inevitable future appears to be one in which there will be significantly more specialization in practice (e.g., breast cancer oncologists, thyroid endocrinologists). Presently, the Task Force recommends that underlying certifications (e.g., Cardiovascular Disease is the underlying certification for Interventional Cardiology) should no longer be required for *maintaining* certification in subspecialty areas that currently require them. The subspecialty area would stand on its own for MOC as the specialization in those areas has already occurred.

A natural extension of this recommendation includes recognition of additional specialization in practice areas that are seeing more subspecialization (such as an endocrinologist who only treats thyroid disease). ABIM will need to consider feasible approaches to recognize these focused areas of practice. In doing so, a strategy for how to represent the scope of practice to the public would be essential. The group recommends that if ABIM decides to address this issue, it should be transparent in the labeling of these focused areas for the public and do it in a way that is equitable for all physicians.

The Task Force realized that there was not enough information about how extensive an issue specialization is beyond what has already been recognized in certification and therefore recommends further exploration before offering any specific options in the MOC program. ABIM's specialty boards and exam committees could be tasked with pursuing these questions and using external data sources such as national surveys to better understand the extent of the sub-subspecialization rates, the feasibility of implementation, and, if reasonable, the best approach for presenting it to the public.

3. Technical Skills

Although ABIM has not traditionally assessed technical skills of practicing physicians in its high-stakes assessment, the group believed that key technical skills were important. Consequently, setting standards would ensure that physicians are educated in procedures that they routinely do. Many agreed that there was a distinct difference between stating that a physician is fit to perform a procedure, and a rigorous assessment of those technical skills. Technical skills are likely to degrade over time since they are not entirely related to the number of procedures performed, but are more about the dexterity and actual physical skills involved in performing them.

Currently, ABIM’s Cardiovascular Board Interventional Cardiology Exam Committee has a workgroup that is developing a simulation for evaluating procedural skills as a supplement to their high-stakes multiple-choice assessment of cognitive skills. This project is currently in pilot form but could potentially be used to assess technical skills that are important in this discipline. Likewise, the American Board of Anesthesiology is working to incorporate simulation as an assessment in their MOC program. The Task Force sees new opportunities to work with medical societies and other Boards to determine the “blueprint” or list of technical skills that are appropriate for specific specialties. ABIM could then work to identify and pilot new assessment methods for evaluating these skills. The group agreed that technical skills were within the responsibility and role of ABIM and should be included in MOC programs.

C. Assessment Methods

1. Examples Across Diverse Competencies

To better understand the state of assessment across a variety of competencies, the Task Force reviewed the assessment and education literature for three seemingly diverse competencies to determine which assessments showed promise and why. These three were clinical diagnostic reasoning, patient experience of care (as one measure of doctor-patient communication skills) and teamwork. Detailed concept papers are included in Appendices D—F. Each concept paper contains a current toolbox of the approaches currently being used to assess these specific competencies.

a) Clinical Diagnostic Reasoning (Appendix D)

Assessing the clinical reasoning process should provide more information on whether a physician gets to the correct diagnosis for the right reasons. In reviewing the literature there was a variety of ways to assess this process, ranging from approaches that are more or less authentic to medical practice.

The first approach includes enhancements to the current multiple-choice exam, which has the benefit of a large robust psychometric evidence base but is less authentic to practice. These enhancements may entail questions that focus on common misconceptions and undesirable actions. They could also include compact mini- performance tasks – sequences of selected response formats that build on one another and focus on key features of the case that identify the critical path needed to arrive at an appropriate reasoning process. A second approach used by the Medical Council of Canada in their qualifying exam for licensing physicians includes the use of short answers (i.e., short constructed response tasks) scored through natural language processing techniques so that answers are not being prompted by multiple-choice options, making the assessment approach slightly more true to practice, and more evidence about aspects of reasoning could be collected. A third approach, similar to the one used on the United States Medical Licensing exam, includes computer-based case simulations. This approach is more sophisticated but more costly. However, it is more authentic to practice in that it includes virtual patients. Furthermore, this approach would involve the use of a partial credit scoring model and could provide even more robust feedback to physicians on where their reasoning process went wrong. It is not clear as to whether this approach is as psychometrically robust as the multiple-choice format and definitely requires more testing time to cover the breadth of a field, but it certainly is a promising approach in need of further exploration.

b) Patient Experience of Care (Appendix E)

The Task Force felt that patient experience of care and shared decision-making in particular were legitimate values and goals in their own right. That is, outcomes from the patient's experience of care do not need to be related to standard clinical outcomes in order for them to be deemed important. For instance, when a physician demonstrates empathy for a patient, it does not necessarily have to imply better outcomes of clinical care. Instead, it is a primary goal of medical practice.

A number of approaches for assessing patient experience of care were reviewed. The first approach is the use of well-established and validated patient surveys such as Consumer Assessment of Healthcare Providers and Systems (CAHPS) and Physician Achievement Review (PAR). Although they can broadly sample the patient population, they lack the rich context of specific patient-physician encounters. A second approach is Objective Structured Clinical Examinations (OSCEs), which have a richer context and yield reliable assessment measures when constructed and administered properly. However, these are more costly and less feasible to implement as the encounters need to be done with live patients or patient actors. The third approach is audio or video recordings of live patient-physician encounters that could be rated based on predetermined criteria. The rating could be done by physician experts or possibly through modern learning environment techniques such as crowdsourcing – a method to obtain ratings by soliciting contributions from an online community who have been given appropriate criteria and instructions.

c) Teamwork (Appendix F)

The thought leaders we interviewed clearly felt that teamwork and care coordination were skills that would be essential for quality patient care in the future even more so than today. We were able to identify three assessment approaches that could be useful in understanding how well an individual functions in a team using either rating scales or other measurement rubrics. The first includes a multi-source feedback approach in which members of a team assess the effectiveness of the individual on the team using rating scales. The second involves observation of the individual functioning in a team in several live contexts. And the third approach uses simulated case scenarios in which a team performs but the specific individual is rated for his/her team behavior.

We learned from reviewing assessment approaches for these three competencies that there are innovative approaches to assessment that should be explored to either expand testing beyond multiple-choice questions (e.g., clinical diagnostic reasoning) or to make the assessments more authentic to practice (e.g., using case-based simulations). Crowdsourcing, a new method for inexpensively evaluating behaviors, could also be explored as to its reproducibility and validity. In addition, various advances in automated scoring systems will make some performance-based assessments less costly to score in the long run. Such systems will be able to connect evidence from different sources, including written and spoken responses, action sequences and log file entries, and features such as gaze, stance and facial expressions. Each new type of assessment brings with it challenges that must be addressed through well-thought-out research studies to ensure that assessments remain valid and defensible.

2. New Approaches to the Format of the High-Stakes Examination

As cognitive skills were seen to be critical to certification, the group deliberated as to what format changes might occur in this arena to ensure that the assessment remains relevant in the future. Three specific areas were considered including the use of external resources during a high-stakes assessment, the frequency, length and consequences of the assessment, and whether assessing performance in practice through patient outcomes could serve as a direct assessment of patient care in specific areas and might supplement the high-stakes cognitive exam. A discussion of each of these follows:

a) External Resources (Appendix G)

There was a clear sense from the physician community that with the rapid proliferation of information, the use of external resources (e.g., tools like UpToDate or DynaMed) is becoming more mainstream practice and that ABIM should consider changing its assessment practices to enable access to these types of resources during the exam. As there was little readily available information about the effectiveness and relevance of open- versus closed-book exams, Dr. Steven Durning led several Task Force members along with other research colleagues in the field of medical education to conduct a systematic review of the literature on the topic. This peer-reviewed paper was published in *Academic Medicine* in October 2015.

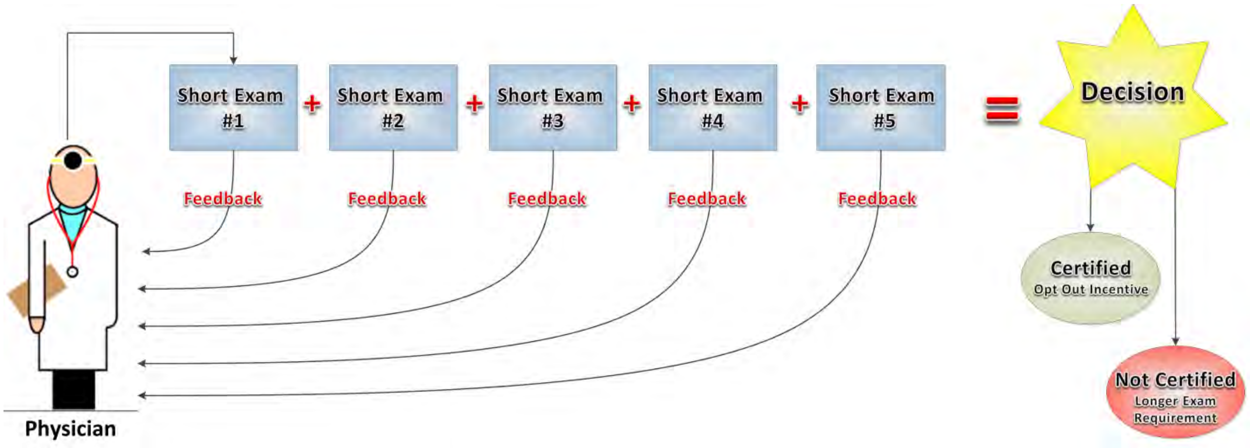
b) Frequency, length and consequences (Appendix H)

The Task Force addressed some of the major criticisms that we have heard from physicians with respect to the high-stakes examination. These include that taking a high-stakes exam once every 10 years is anxiety-provoking and there are significant consequences to physicians' careers if they fail the exam. A number of designs were presented that addressed the frequency, length and consequences of failing the exam. These designs are presented in Appendix H.

Based on these designs, the Task Force recommends that MOC have more frequent assessments, taken in a secure setting (possibly at home with some element of remote authentication), with a potential for some portion of the assessment to be open-book but still timed. The closed-book portion could assess core knowledge in a discipline that would be important to know from memory and could be pre-specified in advance. For example, a discipline could identify the facts and concepts that an individual should know from memory and the assessment could simply involve assessing these facts and concepts by random selection. When new knowledge supersedes older knowledge, people participating in the program would be notified that there were changes in the facts and concepts. The open-book portion could contain questions that are more focused on the clinical reasoning process where access to materials is perfectly suitable within a contained time frame. Other ideas for assessing cognitive skills would be to assess knowledge of key peer-reviewed articles that are identified in advance. These articles would be selected based on their impact. These questions would address advances in the field related to the article. This assessment could be done in a non-secure, open-book setting. The specific content and the decision about open- or closed-book would need to be researched further to better determine the value and feasibility of the assessment.

Initially, the more frequent exams would begin as low-stakes assessments. Physicians would be given detailed feedback to understand their areas of strength and weakness, and would be able to build to a steady state of competency. At all times, the physician would know how close he/she was to meeting the threshold for competency. These low-stakes decisions would, at some point (for example, every five years), be rolled up into a high-stakes decision as to whether the physician remained certified. If unsuccessful, physicians would likely have to take a longer assessment similar to the one currently given every 10 years. This approach should conceivably be less stressful and would encourage physicians to maintain a steady state of knowledge rather than cramming for a high-stakes exam. Variations on this theme include rewarding high performance by allowing physicians to opt out of some of the assessment after high performance was shown for a stable period of time. The details of frequency of testing, when and how to roll up to a high-stakes decision, the scoring approach and what happens if a physician is unsuccessful all have to be studied carefully before implementation.

As this is a more continuous approach for assessing cognitive skills, it would eliminate the need for separate self-assessments (Part II of the MOC program) as continuous learning would be contained in this approach. A diagram of how this might work is pictured below.



Feedback will include information about areas for improvement & probability of remaining certified following all short exams.

c) Patient Outcomes (Appendix H, continued)

The Task Force considered the issue of measuring performance in practice through patient outcomes or a composite of patient outcomes. This measure could serve as a direct measure of patient care in narrow domains of practice and supplement the cognitive assessment that would continue to cover the breadth of the field.

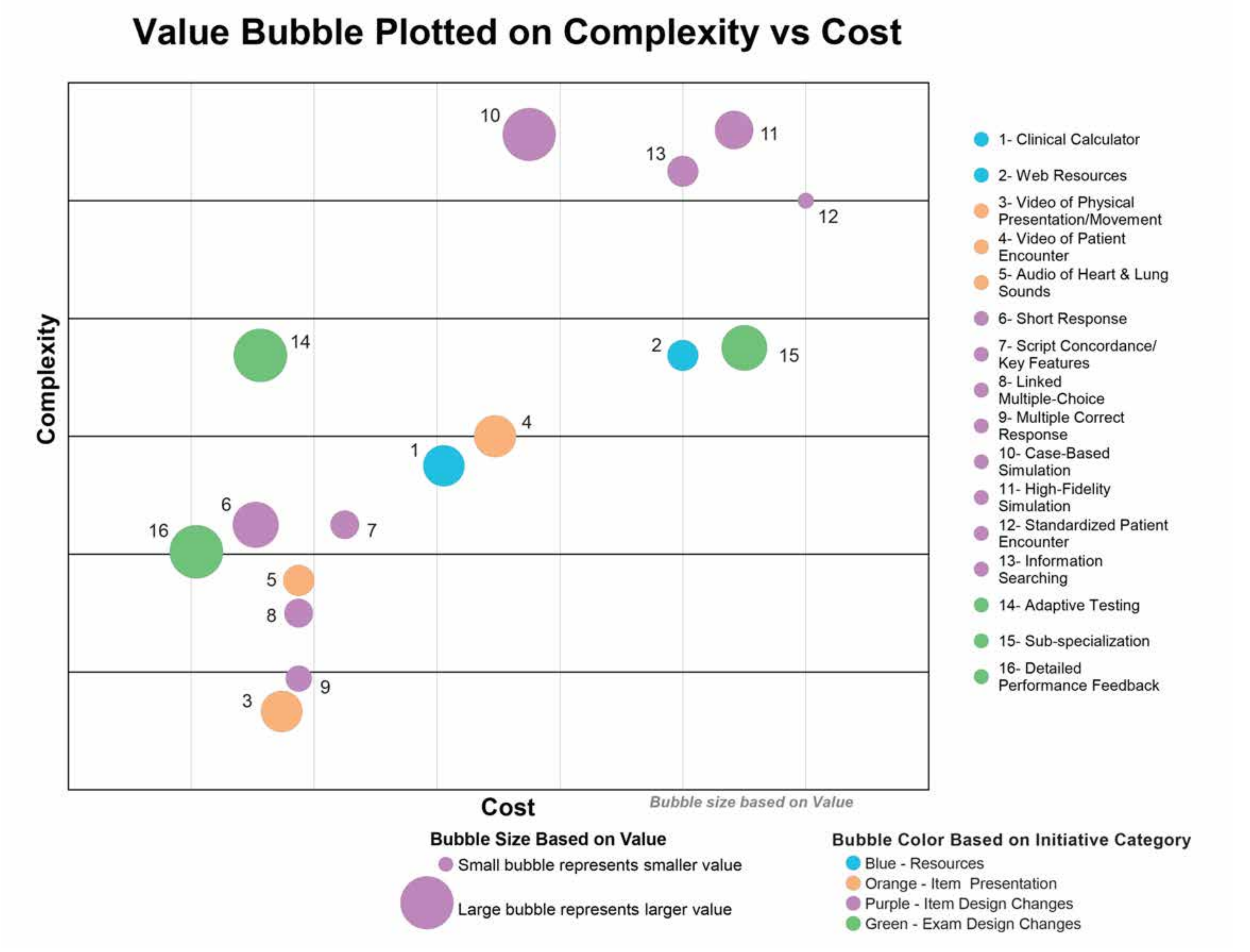
Good patient outcomes have the potential of being indicative of good patient care assuming they are properly risk-adjusted and do not incur unintended consequences (i.e., physicians limiting their practice to healthier patients). Physicians in larger, more organized health systems are already being held accountable by their employers for their patient outcomes wherever evidence-based measures are available. However, physicians in smaller practices may not have the resources needed to collect and report objective quality measures.

The Task Force saw measurement of performance in practice through patient outcomes as an important trend in health care that could not be ignored. However, at this time, it was not clear how much of that care could be attributed to an individual physician and what the unintended consequences might be for patient care if ABIM held the individual physician responsible for that outcome. The Task Force recommends that ABIM continue to monitor the environment to determine if and when the data become meaningful, reliable and with few unintended consequences, as well as flow freely in all electronic health records. At that time, ABIM should consider working with health systems to set benchmarks/standards for performance in the different specialty areas and incorporate the assessment into MOC requirements.

It is clear that the skills and knowledge of individual physicians are an important component of producing desirable patient outcomes. Consequently, ABIM should presently focus on better elucidating the critical individual physician skills that are important to producing better outcomes (e.g., knowledge and/or practice of quality improvement) and develop approaches to assessing those rather than focusing on outcomes themselves.

D. Evaluation of Assessment Approaches: Complexity, Cost and Value

As mastery of cognitive and technical skills were seen as critical to the certification of individual physicians, we evaluated different kinds of changes to the current multiple-choice assessment that might enhance its relevance and make it more engaging to the physician as well as more meaningful to the public. These changes include small as well as large enhancements. A group of ABIM assessment experts evaluated each change based on three dimensions: its complexity, cost/risk, and its value (validity/opportunity) relative to the current multiple-choice examinations. The chart below evolved from prior work to help prioritize enhancements. A team of 21 measurement experts within and outside ABIM rated specific aspects of the three dimensions. These ratings were then aggregated, discussed as group and adjusted through a consensus process. The chart depicts the result of this evaluation. The three dimensions are depicted as: Cost on the horizontal axis, Complexity (including risk) on the vertical axis, and the size of the bubble as the Value (including opportunity). The bigger the bubble, the more value it adds.



There are four different types of changes listed in the chart:

1. **Resources**: This is depicted in blue and shows two types of external resources that could be added to the cognitive assessment. For example, the addition of clinical calculators (i.e., specific formulas that are used in practice such as CHADS2 for Atrial Fibrillation Stroke Risk would be provided in the form of a calculator) is a smaller enhancement than adding Web resources such as UpToDate or DynaMed.

2. **Item Presentation:** This is depicted in orange and involves adding additional stimuli to the particular item. For example, there are three instances of item presentation including videos of a patient showing a physical presentation or movement (e.g., a patient's gait), videos of a patient encounter with a doctor where there is communication between the two, and audio sounds of the heart or lungs.

3. **Item Design:** This is depicted in purple and shows eight different item designs. These include short answers, script concordance questions that examine the physician's ability to interpret medical information under conditions of uncertainty, multiple-choice questions that are linked together to present mini-cases, multiple correct response that requires more than one correct answer, case-based simulations that present open-ended patient vignettes on the computer, high-fidelity simulation for assessing technical skills such as stent placement in a catheterization laboratory, a standardized patient encounter which uses patient actors to simulate a doctor-patient encounter, and finally information-searching skills.

4. **Exam Design:** This is depicted in green and shows three approaches. The first is adaptive testing which is a computerized test that adapts to the examinee's ability level. The assessment typically is more reliable than a linear test. The second type is subspecialization or focused practice/modular exams with specific choice. The third is detailed performance feedback which provides information on specific items that have been missed on the exam (this is a new enhancement to ABIM's high-stakes exam that will be rolled out beginning with physicians who took exams in spring 2015).

As reflected in the bubble chart, the most valuable changes (based on the assessment experts' experience) include more detailed performance feedback (low cost, medium complexity), adaptive testing (low cost, medium complexity), and case-based simulations (medium cost, very high complexity). Other lower cost and complexity options include videos, audio, short answers and the clinical calculator.

This approach to evaluating changes has guided research and development efforts at ABIM to advance our knowledge of new assessment approaches so that we are better able to provide more meaningful, relevant and engaging assessments to our stakeholders. In alignment with Miller's competency pyramid, an assessment portfolio for initial certification and MOC could contain various assessments that achieve different levels of authenticity to practice and can cover both the breadth and depth of the content of the domain.

V. RECOMMENDATIONS

The recommendations that derive from the work of the Task Force are as follows:

1) **Replace the 10-year Maintenance of Certification Exam with More Frequent, Less Burdensome Assessments.**
   The Task Force recommends replacing the 10-year Maintenance of Certification (MOC) exam with more frequent assessments that could be taken at home or at the workplace. The new format would be designed to assess competency in essential contemporary knowledge. Some aspects of the assessment would be "open-book" and some would represent knowledge that should be known without outside references, but specified in advance by the profession. The results of the smaller, more frequent, lower stakes assessments would provide insight into performance, and would accumulate over time and culminate in a high-stakes pass/fail decision. A failure at this point may necessitate taking a longer exam or another form of assessment in order to maintain certification. This approach would emphasize learning as an integral part of the program, but also provide meaningful criteria to the public as to whether a physician is remaining current. The current self-assessment component of MOC would likely no longer be a separate requirement as the new exam format would provide knowledge assessments on a frequent enough basis to obviate the need for it.

2) **Focus Assessments on Cognitive and Technical Skills.**
   The Task Force recommends that ABIM focus its MOC efforts on assessing cognitive and technical skills relevant to the practice of internal medicine. The rationale is that there are specific competencies in these domains that are unique to the internist and that may degrade over time. In addition, there are rigorous and scalable assessment methods that are currently available or will be available soon to measure these competencies. Assessment of cognitive skills will assure the public that physicians are keeping up with the clinical knowledge that is relevant to patient care. Assessment of technical skills will assure that physicians can apply that knowledge to adequately perform the technical procedures. ABIM should continue to focus on developing assessments of these competencies that closely align to actual practice through innovative approaches.

   Other competencies, such as communication, teamwork, empathy and quality improvement, are also vital for effective patient care, but formal assessment of them for practicing physicians is challenging. These skills have some special attributes. They may be context dependent in that the systems and teams may influence the ability of an individual to demonstrate them. Merely participating in programs such as those focused on quality improvement, although important, may not indicate meaningful performance of such activities. Direct observation may be critical for assessing competence. The Task Force recommends that ABIM continue to include the demonstration of these skills as part of initial certification requirements as these are assessed in a standardized and uniform way in training programs and under direct observation. However, the best approach to assess these skills at the individual level outside of a training program is not clear. ABIM should continue to emphasize the importance of these skills and encourage health care organizations to promote and assess these skills locally and in context. As methods emerge that are effective and efficient—that can account for context and convey meaningful information without undue burden—ABIM should re-evaluate its role in assessing these competencies.

3) **Recognize Specialization.**
The Task Force recommends a movement toward certification in specialized areas, and to do so without the need for underlying certification (e.g., Cardiovascular Disease is the underlying certification for Interventional Cardiology). Thus, an underlying certification should no longer be required for *maintaining* certification in subspecialty areas that currently require them. The subspecialty area would stand on its own for MOC. A natural extension of this recommendation includes recognition of additional specialization in relevant practice areas. ABIM will need to consider feasible approaches to recognize these focused areas of practice. In doing so, a strategy for how to represent the scope of practice to the public will be essential. Ultimately, the goal is for the customization of MOC, so that it represents an individual's practice and is appropriately transparent and meaningful for the public.

VI. IMPLICATIONS

1) **Interactions with Physicians and the Broader Community.** The valuable input that was obtained from physicians and the broader community implies that it would be beneficial for ABIM to expand its efforts to engage physicians and the broader health care community and the public into continuously evolving its programs to stay meaningful, relevant and attentive to the impact on physicians. These efforts could involve more localized visits to physician offices to understand changing practice, and more focus groups or surveys involving patients, physicians and health systems.

2) **Research and Development Efforts in Assessment.** Although traditional multiple-choice questions currently have the benefit of a large robust psychometric evidence base, the analysis provided in this document shows promising new approaches to assessment that should continue to be studied. Advances in techniques such as automated scoring enable more sophisticated item types. The use of computer case-based simulations and high-fidelity simulations will likely become more scalable in the future because of this technique. This implies that ABIM would benefit from continued research and development efforts in concert with other assessment organizations and research entities. Efforts are also needed to make the assessments less expensive, less burdensome, and more engaging and relevant to practicing physicians.

In sum, throughout this report we have attempted to identify principles—guided by the needs of patients and society—for certification and MOC programs that are relevant, engaging and efficient for physicians while remaining a trustworthy and valid assessment system. It is our hope that this report will stimulate important dialogue that will result in improvements in the certification process and better care for patients.

REFERENCES

1. ACGME. ACGME outcome project: table of toolbox methods. [Internet]. 2009 Jan 2. [Cited 2015 Sept 2]. Available from http://www.acgme.org/Outcome/assess/Toolbox.pdf
2. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Med Educ. 2005;39(3):309-17.
3. Kane M. Validating the interpretations and uses of test scores. J Educ Meas. 2013;50:1-73.
4. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; American Psychological Association; National Council of Measurement in Education; 2014.
5. Choudhry NK, Fletcher RH, Soumerai SB. Systematic review: the relationship between clinical experience and quality of health care. Ann Intern Med. 2005;142(4):260-W-230.
6. Davis DA, Mazmanian PE, Fordis M, Harrison RV, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence. JAMA. 2006;296(9):1094-102.
7. Audet AM, Doty MM, Shamasdin J, Schoenbaum SC. Measure, learn, and improve: physicians' involvement in quality improvement. Health Aff. 2005;24(3):843-53.
8. Controlling health care costs while promoting the best possible outcomes. Philadelphia, PA: American College of Physicians;2009.
9. Gawande A. Big Med. The New Yorker. [Internet]. 2012 Aug 13. [Cited 2015 Sept 2]. Available from http://www.newyorker.com/magazine/2012/08/13/big-med
10. Berner ES. Clinical decision support systems: state of the art. Rockville, MD: Agency for Healthcare Research and Quality;2009. 09-0069-EF.
11. Friedman LF. IBM's Watson supercomputer may soon be the best doctor in the world. Business Insider. [Internet]. 2014 April 22. [Cited 2015 Sept 2]. Available at http://www.businessinsider.com/ibms-watson-may-soon-be-the-best-doctor-in-the-world-2014-4
12. News KH. How smart phone apps help doctors track your health. Health Talk. Vol 2014: AARP; 2014.
13. Krisberg K. Big data key to improving health care. AAMC Reporter. [Internet].2014 Jan. [Cited 2015 Mar 1]. Available at https://www.aamc.org/newsroom/reporter/january2014/366338/big-data.html
14. Norcini JJ, Lipner RS, Grosso LJ. Assessment in the context of licensure and certification. Teach Learn Med. 2013;25:S62-S67.
15. Carracchio C, Englander R. From Flexner to competence: reflections on a decade and the journey ahead. Acad Med. 2013;88:1067-73.
16. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32(8).
17. Hawkins RE, Lipner RS, Ham HP, Wagner R, Holmboe ES. American Board of Medical Specialties maintenance of certification: theory and evidence regarding the current framework. J Contin Educ Health Prof. 2013;33:S7-S19.

18. Lipner RS, Hess BJ, Phillips RL. Specialty board certification in the United States: issues and evidence. Contin Educ Health Prof*.* 2013(33):S20–S35.

19. Miller GE. The assessment of clinical skills/competence/performance. Acad Med*.* 1990;65(9 Suppl):S63-67.

20. Eva KW, Regehr G, Gruppen LD. Blinded by “insight”: self-assessment and its role in performance improvement. In: Hodges BD, Lingard L, editors. The question of competence: reconsidering medical education in the twenty-first century. London: Cornell University Press; 2012.

Appendix A:
Thought Leader Interviews

Assessment 2020 –
Interviews with thought leaders about coming changes in health care, and how ABIM should respond

Benjamin Chesluk, Bryn Herrschaft, Elizabeth Bernabeo, Siddharta Reddy, Helene Brooks

I. Purpose

In order to support the Assessment 2020 initiative, a team of ABIM staff researchers conducted interviews (N=28) with individuals identified by senior ABIM staff and members of the Assessment 2020 Task Force as "thought leaders" in health care – individuals who would likely have an original and important perspective on how the U.S. health care system may change in the coming decade, and how ABIM should respond. The interviews focused on how health care in the U.S. is likely to change, how physicians will have to respond, and what this will imply for physician assessment and ABIM more generally.

II. Methods

Each interview lasted approximately 30-45 minutes, and all were conducted by phone with a team including one or more interviewers (BC, EB, SR) and a dedicated note-taker (HB, BH). Each member of the research team (BC, EB, SR, HB, BH) read all the notes, both as the interviews were conducted and at the end of the data collection periods, and analyzed these for emergent themes and patterns. The team met frequently and worked together throughout the research to compare individual analyses, in order to refine the analytic approach and better define the findings.

We conducted a first round of interviews with a purposeful sample of 13 participants recommended by the members of the Assessment 2020 Task Force as experts in health care research, assessment and education. These were primarily physicians who were familiar with ABIM's role in setting standards and creating assessments for the practice of internal medicine. This initial round of participants focused on assessment – what new skills and competencies physicians may need to practice in a changing health care system; what information about physicians will be important to patients and to employers and payers; and how ABIM should develop its programs to address these changes.

Our analysis indicated that several additional themes were important to participants, including: imminent changes that would impact the U.S. health care system; what these changes would mean for physicians and the practice of medicine; and what ABIM should do, both internally to adapt our program and develop new assessments, as well as externally, engaging with other organizations to help these changes develop in the most beneficial way possible. These findings were presented to the Assessment 2020 Task Force. The Task Force recommended that we continue to collect data to pursue these important insights. Using similar "thought leader" criteria to the first round, we purposefully sampled another 12 participants to reflect a diverse range of backgrounds, both professionally (social scientists, patient representatives, and non-physician providers as well as physicians) and sociodemographically (representing a broader range of ages and racial/ethnic backgrounds). Sampling was terminated when we felt our participants' responses were repetitive and no new themes were emerging (i.e., when we had reached theoretical saturation).

The second set of interviews were analyzed individually (BC, EB, SR, BH) for emerging themes and collectively (N=25) to see if and where new, emergent themes could help us interpret findings from the first round. At this step of the analysis, we noted that multiple participants spontaneously brought up the issue of current and future developments in genetic testing and personalized medicine. This subject had been raised by many participants, but none of our participants had specific expertise in this area. To fill this gap, we conducted a final round of interviews with three experts in medical genetics (N=3). (See Figure 1 for a visual representation of the three groups of participants.) This report presents our analysis of all 28 interviews together.

III. Results – grouped by overarching theme

Our analysis revealed several organizing themes classifying participants' predictions about coming changes in health care and how physicians will have to adapt:

1. Science/technology – how medical science and health care technology will advance, particularly focusing on the electronic health record (EHR) and decision support systems.
2. Patients – how the population of patients will change, both in terms of the medical conditions that require care as well as what they expect from health care providers.
3. Health care system – how the system of organizations that employ providers or pay for their services will change.
4. Teamwork/roles – how the roles that physicians and other providers play in working together to provide care will change.
5. Identity – how all these changes may impact physicians' professional identity and the status of the profession of medicine.

Our analysis focused on identifying our participants' claims and assumptions within these headings. We identified areas where participants shared common assumptions, areas where they shared some assumptions but also reached differing conclusions, and areas where comparing participants' statements indicated significant debate or lack of commonly shared assumptions. It is important to note that our analysis did not focus on tabulating the precise number of times specific ideas were mentioned, but rather sorting participants' statements into these categories: relatively unanimous shared assumptions; some alignment between assumptions; and significant disagreement. Our presentation of the data below highlights these areas of shared or contrasting assumptions within each thematic heading.

It is also important to note that our description of "what ABIM can do" within each section is meant to document all the various recommendations for ABIM from our participants. The ideas described in these sections do not represent the recommendations of the research team for ABIM to follow – our goal is to be comprehensive in representing what our participants thought ABIM may want to consider, without specifically recommending or critiquing any of these in particular. In addition, there may be things ABIM should do that our participants did not mention in these interviews. We hope that this material can provide the springboard for further discussion about which, if any, of these recommendations ABIM should actually pursue.

1. Science/technology

What may change:

Our participants were unanimous that the rate of development of medical science and technology may continue to increase, particularly in the area of information technology, where tools for health care systems, providers, and patients could advance exponentially.

Some felt that the rapid development of medical knowledge and available procedures, medications, and tools may present clinicians with an overwhelming amount of information – too much for an individual to master. This may complicate the definition of "general internal medicine" and make it harder to define what should be the common knowledge all physicians should have at their recall. Generalists may find it more challenging to stay current in such a broad field, and subspecialty areas of medicine may become ever more developed and specialized.

At the same time, physicians may have access to much more information about each specific patient, as new tools and apps for monitoring and recording patient information may become ubiquitous. Participants had varying perspectives on what this would mean for physicians. Some predicted that this could require physicians to become extremely sophisticated at distilling large amounts of patient data and available medical knowledge in the course of providing care. The hope is that EHR systems would become at least somewhat more useful and user-friendly to physicians in these tasks – in essence, a better version of current tools such as UpToDate. By contrast, many participants predicted that EHR systems and decision support tools may become sophisticated and “intelligent” enough to take the demand off physicians to manage amounts of knowledge and data beyond what any individual could do. In this scenario, providers may give over much of their current work of analysis and diagnosis to smarter and more powerful artificial intelligence, like IBM’s Watson.

Other ways that technology could impact the practice of medicine were mentioned, including: patients having greater access to medical information and their own data, via online resources and personal health apps; physicians and patients having access to an increasing set of tools to enable contact and data sharing outside of the traditional office visit (tools such as videoconferencing and remote monitoring of patient data via smartphones or wearable monitors); and the growing ability to leverage “big data” (massive new sets of patient data) to make care decisions for individual patients and for populations. What all these have in common is the likely reduction in importance of the face-to-face physician-patient visit. More tools may be available to gather information about patients, communicate with them, and make decisions regarding their care without direct contact.

An additional area of likely rapid change is the field of genetics and personalized medicine. Participants talked about the continuing trend of identifying genes responsible for specific diseases and conditions and about this information’s potential to radically change current categories of disease, as well as the increasing availability and routine use of whole-genome screening, perhaps even at birth. Faster and more accurate diagnosis of many conditions, and more efficacious and personally-tailored treatment options, could become available, though treatment is likely to lag behind diagnosis, presenting patients and providers alike with challenges in deciding how to act on information arising from genetic testing.

How physicians may need to adapt:

The rapid increase of medical knowledge and available procedures, devices, and medications may challenge all physicians to keep up with these advances and adapt their practices to them. Subspecialists may need to become increasingly super-specialized in order to be able to maintain mastery over a defined body of medical practice. Practitioners of general internal medicine could find their role challenged, and may be pushed to a more specialized role of expert diagnostician/care planner for patients with unusual or complex conditions. (Several participants referred to this new role as requiring general internal medicine physicians to become “Dr. House.”)

All physicians will need to embrace information technology and become sophisticated in its use, though participants disagreed on how this might play out in practice – if health information technology stays unwieldy at the point of care, as many systems currently are, then physicians will need a great deal of individual sophistication in using such systems to filter through increasing amounts of patient data and making clinical decisions. If health IT becomes both ubiquitous and highly usable, then physicians will need less individual skill in this area. And if health IT becomes both more usable and much more like true “artificial intelligence,” then physicians may find their professional role changing, with less emphasis on managing information to make diagnoses and more on guiding patients through making decisions about information and treatment options provided by the expert system.

Beyond the use of expert systems to manage patient data and make clinical decisions, physicians may need to be open to staying current on other advances in technology, such as telepresence tools, and to integrating these into their practice – as well as helping patients make decisions on what consumer tools to use to monitor and manage their own health information.

The massive increase in available data and advances in genetic medicine could also open a new competency area for physicians: analyzing population-level "big data" and using that information to help make decisions for individual patients. Physicians and other providers could need to keep up with a rapidly-developing field, and understand how to apply it appropriately to the care of individual patients and their families. Physicians may need to know when to order genetic tests, how to interpret their results, how to integrate this information with other streams of patient data, and how to work with patients to make decisions based on this information.

What ABIM can do:

ABIM can assess physicians in the many new or rapidly changing competency areas mentioned above, including:

- Public health and population care (including risk analysis, preventive care, and the design of care systems).
- Use of technology, particularly EHRs (focusing on how effectively physicians can manage the patient information that is available to them).
- Genetics and personalized medicine (including how to take a genetic history, when to order genetic tests and how to interpret them, when and how to refer to specialists in this area, and especially how to frame diagnoses and treatment options for patients and their family members).

In addition, ABIM can expand its program to include new assessment methods to provide different perspectives on physician knowledge, decision making, and practice performance. Some of the specific new assessment methods mentioned by our participants included:

- Increased use of patient survey data (for example, including data from patient surveys in reporting about certification on the ABIM website).
- Assessing how well physicians are using the data available to them by looking directly at data extracted from physicians' EHR systems.
- Standardized simulation of individual- or team-level situations to assess decision making and clinical judgment.

2. Patients

What may change:

Another unanimous observation by our participants was that the population of patients that physicians care for could grow in size, in complexity, and in the level of demand they place on physicians to accommodate their wishes.

The Affordable Care Act (ACA) is anticipated to greatly increase the number of patients seeking providers for primary care. The proportion of elderly people in the population will increase, and if current rates of increasing obesity and diabetes continue, then the population of the U.S. will have even more complex health needs.

In addition, as mentioned above, many participants noted that patients may likely continue to become more informed about their health needs thanks to online resources, personal health apps, and monitoring devices. These patients may be entering their relationship with health care providers with a large amount of data they may need help interpreting, and with ideas about their conditions and the care they want that providers will need to engage and perhaps accommodate.

Related to changes in the health care system that will be described below, a number of our participants noted that patients may expect providers to cater to their preferences, applying models from the service industries of measuring consumer engagement and satisfaction. To the extent that patients feel empowered to "shop around" among providers for those that cater to their preferences, physicians and other providers could be expected to include measures of consumer experience alongside more traditional clinical measures. Many participants assumed that this trend would increase, and hoped that these measures of experience would focus on meaningful areas such as patient engagement and shared decision making.

How physicians may need to adapt:

A larger population of patients with more complex combinations of conditions exacerbated by age and obesity could present numerous challenges to physicians. The health care system will need to provide primary and specialty care to an older, sicker, more complicated population. Caring for these patients may require physicians and other providers to engage in more integrated care planning, making connections between medicine and areas that have traditionally been more the purview of public health, such as nutrition, fitness, and environmental health. The increasing demand for primary care could create a need for more general internal medicine physicians at precisely the same time that the concept of "general" medicine is being challenged by exponentially increasing medical knowledge. Many participants stated that this demand for primary care would likely be met by physicians and other providers working in teams and/or by other providers (such as nurses) filling roles currently occupied by physicians. Whatever the case, the demands may result in physicians having less direct contact with all but the most complex patients.

Physicians may need to focus increasingly on competency areas traditionally associated with public health: the social determinants of health, understanding how a patient's environment impacts health and wellness, and accessing available resources to address these issues to improve patient outcomes. Physicians may also need to adopt a more service-oriented focus on measuring and improving their patients' experience of care. Many participants predicted that measures of patient experience will need to be taken as seriously as clinical process and outcome measures.

Training and certification entities may need to focus on competencies related to communicating with and engaging patients, understanding their needs and preferences, and adapting treatment plans accordingly. One aspect of this will be that patients could require providers to help them manage the growing amount of personal data and medical information available to them.

What ABIM can do:

A substantial number of our participants suggested that ABIM should focus on assessing physician competencies in delivering patient-centered care, to the extent that these can be shown to lead to better outcomes. Competencies mentioned by participants included physicians' effectiveness in listening to and communicating with patients; engaging patients and families at the appropriate level for their needs and capabilities; and integrated care planning (including nutrition, fitness, and education). A few participants mentioned the idea that ABIM could evaluate practices as to whether they have features associated with patient-centeredness; specific examples of possible features included use of care checklists (to ensure patients understand communication from providers) and providing patients with access to the same data available to physicians.

Many participants suggested that ABIM could provide patients with more detailed and relevant information to help them select among physicians, including information about clinical outcomes, efficiency/cost (this could be especially relevant if payers shift more costs to patients), and standardized measures of patient experience and satisfaction. Connected to this was the idea that ABIM could offer patients tools and information to improve their interaction with health care providers, such as checklists and templates to follow in conversations with physicians.

3. Health care system

What may change:

Our participants were nearly unanimous in assuming that the health care system in the U.S. may move to being more integrated and organized, more effective, and more efficient. They articulated a common, aspirational vision of health care in the U.S. becoming a more truly integrated system, with organizations following an accountable care model and designed to deliver care more proactively and efficiently to a population, rather than largely reacting to sick individuals' needs. Many cited the ACA and the current accountable care organization (ACO) movement as indicating the direction of change, as integrated care organizations could assume responsibility for a population of patients.[1]

This new system was commonly described as entailing a shift in providers' roles, with a new emphasis on preventive and primary care taking precedence over hospital and subspecialty care. Participants envisioned a number of likely structural/systemic changes, including:

- A reduction in the number of hospitals, and thus a reduced need for hospital-based providers.
- An increase in the scale and integration of primary-care delivery systems – including a likely shift from physician-led practices to primary care being delivered by teams of providers working under the auspices of larger health care organizations.
- Connected to the above, a likely continuation of the current trend of steep reduction in the number of small and physician-owned independent practices.

If physicians increasingly work within care organizations such as these, rather than in independent practices, they will need to work within the practice parameters set by the organizations that employ them, operating within constrained budgets and following guidelines and protocols set by others. This may entail practicing in a more monitored and constrained environment. Participants envisioned the organizations that employ physicians more carefully scrutinizing their practice processes and outcomes, and pushing physicians to practice more efficiently.

For patients, participants described a health care system that will hopefully work more effectively and provide better experiences of care – more organized, more convenient, more proactive, and more understanding and accommodating of patients' individual preferences and needs, again following the models set by service industries. Of course, this might be an overly optimistic view – some participants described the possibility that patients' experience of care could become worse, as more patients with more complex conditions seek to engage with a shrinking number of providers

[1] It is interesting to note that many interviewees took it for granted that fee-for-service would decline as part of the U.S. health care system, to be replaced with other models: bundled payment, capitation, etc. However, some experts in this field (not interviewed in this project) see fee-for-service as so deeply entrenched in U.S. health care that dislodging it might take more time and effort than our interviewees foresaw. It remains to be seen how these new payment models will actually be implemented.

working through a more complex bureaucracy. If the health care system does become harder for patients to navigate, this potentially creates even more demand for someone to help them get the care they need; some participants optimistically saw this as a role physicians might be asked to play.

How physicians may need to adapt:

For physicians and other providers, a reorganized health care system would certainly have a tremendous impact on their working lives, one that many current providers would find quite disruptive. Even the most optimistic participants described these changes as extraordinarily disruptive, potentially requiring physicians and other providers to work in new ways, within tighter constraints and with less of the satisfaction of providing direct patient care. It is interesting to note that participants characterized these new demands as challenging for physicians, but hopefully worthwhile, as they could be working as part of a care system that delivers better care, more proactively and more efficiently, to more people.

An increased demand for primary and preventive care and a decreased demand for specialty care would require major workforce shifts. The impact of this type of reorganization would vary by specialty and practice setting. For primary care physicians, they could expect to increasingly be called on to take a public health focus, helping to design the care that others—members of other professions, particularly nurses—could deliver to populations of patients. Participants also predicted an increase in other professions being called on to help meet the increased need for primary care, with a rise in nurse-led practices, etc. There was no clear consensus among our participants about what physicians' specific role in primary care could be. Some envision physicians moving to a largely managerial role, leading care teams rather than getting directly involved in patient care. Others look at the likely growing complexity of health care and patients' issues and foresee a continued or even intensified role for physicians as the patient's primary advocate and "information broker."

For hospital providers and specialists in technologically complex fields, this may be a period of painful contraction, uncertainty, and anxiety, as hospitals close and reorganize under the auspices of ACOs seeking to conserve resources and reduce the need for specialized physician care. There may continue to be a demand for their services, but these could likely be much reduced overall, as the organizations that employ physicians seek to reduce patients' needs for these by providing more preventive care.

To meet these demands, physicians may need to develop the competencies required to practice effectively in an integrated care organization. A major component of this will be teamwork, a topic discussed in more detail in the next section. Other ACO competencies mentioned by our participants include:

- Ability to use population-level patient data ("big data") for risk analysis and the design of care protocols.
- Public health – including understanding the social determinants of health and how to design care for patients from different sociodemographic contexts.
- System design – the ability to design care systems and protocols to be implemented by health care teams.
- Ability to work as part of an ACO – to practice effectively within resource constraints, protocols, and performance targets set by a larger organization.

What ABIM can do:

Working as part of an ACO-type organization will entail new physician competencies, including the new skill/knowledge areas mentioned above, as well as attitudes that participants predicted physicians would find new and, for some, challenging to adopt: a willingness to work as part of a larger organization, follow organizational guidelines, etc. Several participants mentioned that ABIM could assess physicians in these areas, with particular foci on clinical decision making in ambiguous situations and the cost/benefit implications of treatment decisions. The question for many was whether ABIM could provide information to an ACO on how well and efficiently a physician practices that is different in type or quality to what the ACO can learn from its own systems.

A particular area that was singled out by one interviewee was to assess physician abilities in “hot-spotting” type analysis – assessing whether physicians could use population data to identify specific areas where their health care system was not functioning well, analyze this problem, and identify and implement possible solutions.

4. Teamwork/roles

What may change:

Our participants were unanimous in predicting that physicians and other providers may increasingly need to be able to work together as part of organized interprofessional teams embedded within larger care systems. They referred to teamwork as a path to increasing quality and efficiency, two goals that could be of paramount importance to ACO-type integrated health care organizations.

Participants frequently predicted increasing pressure for providers to work at the “top of their license,” meaning providing only the services that they are uniquely capable of providing based on their professional training. Many participants observed that this may require training and residency programs to focus more explicitly on working as part of formal teams and the competencies involved in teamwork and team design/management.

Participants envisioned that this increasing focus on teamwork and specialization of roles could have various outcomes for patients. Patients could have much improved experiences of working with well-organized teams backed up by capable organizations. Or the patient experience could be one of losing any sense of personal relationship with physicians or other providers.

How physicians may need to adapt:

According to our participants, working with other providers as part of a team may require physicians to step out of the role of directly providing routine care. One scenario referred to is that nurses and other health professionals will care for patients with routine needs, with physicians directly caring only for patients with complex conditions and needs. A more extreme scenario described would entail physicians managing teams and designing care systems for others to deliver to patients – something many current physicians would find quite unfamiliar, and not currently focused on in training.

On a practical level, physicians may need to be adept at working in close coordination with others, and be comfortable both with delegating tasks to other team members and with accepting direction set by supervisors and managers. The knowledge, skills, and attitudes required to work as part of an interprofessional team could become of paramount importance to the organizations that train and employ physicians.

On a more general level, many participants predicted that the profession of medicine may need to learn to let go of the value currently placed on professional autonomy and forming personal relationships with patients. Instead, physicians may need to shift to a new set of values focused on teamwork and fitting into a role defined by a larger organization, with less direct patient care for all but the most specialized physicians.

What ABIM can do:

ABIM can address these changes by assessing physicians in the competencies related to teamwork and their changing roles within the health care system. Some of these competencies have been specified above, including competency in public/population health and system design. Specific teamwork competencies mentioned by our participants as becoming increasingly important for training programs to focus on include:

- Leading or managing care teams
- Sharing decision making with other providers
- Care coordination/integrated care planning

In addition, many participants mentioned the potential for ABIM to move from certifying individual providers to certifying care teams, though they acknowledged this would be a departure for the organization.

A particular assessment method for looking at these teamwork competencies mentioned by several participants, either on the individual or the group level, would be the use of standardized simulation, in which individuals or teams could be observed facing a defined challenge or problem.

5. Identity

What may change:

A theme that surprised the project team was how physicians' professional identity could be impacted by the changes outlined above. Physicians and the teams and organizations they work within may be faced with more constrained financial resources and more need to compete on the dimensions of cost, effectiveness, and patient experience. These increased pressures may originate from payers, employers or ACO managers who, working with real-time data, seek to identify high-cost, low-quality providers. Improving all three of these at the same time – understanding and improving patient experience, while simultaneously improving population outcomes and reducing the resources used (including physician time) – could be a complex undertaking. This could pose significant challenges to the profession of medicine.

Many participants spoke at length about how physicians may need to grapple with a loss of professional autonomy (due to working within larger care organizations that will seek to manage the resources they use in caring for patients, and due to working as part of a team, perhaps in a more managerial capacity) combined with a loss of prestige and wealth (as health care in the U.S. moves away from fee-for-service and towards efficiency and accountable care). Several participants described this as presenting a major challenge to practicing physicians (who may need to learn to be effective and find professional satisfaction in different roles than they were trained for), to learners and new physicians (who may be facing a very different organizational climate than they may have expected), and to the organizations that train, employ, and certify physicians and other providers.

Our participants expressed the hopeful side of this, too: the hope that these changes may give physicians and other providers the satisfaction of being able to provide better care to more patients more efficiently, more reliably, and more proactively. In sum, being a physician could mean something different; the question is whether individual physicians and the profession will embrace these changes or attempt to resist them.

How physicians may need to adapt:

Many participants predicted that the profession of medicine may need to find ways to "scale" itself. In other words, physicians may need to leverage their time- and resource-intensive training and expertise in order to provide better services more efficiently to much larger populations of sicker patients, all within tighter cost constraints and more narrowly-defined roles. This need to scale relates directly to the almost-certain development of new technological tools that could change what physicians do. Physicians may need to learn to incorporate these new tools, which will allow them to work in different ways, but also allow other providers or patients themselves to take over tasks they have historically monopolized.

The most typical way participants referred to this general challenge of scale was to compare it with the historical transition from artisan labor to mass-production manufacturing. In other words, physicians may need to learn to design and manage large-scale systems, rather than serving patients on an individual, one-at-a-time basis. Certain procedural specialties may continue to work in this way, but overall, physicians could need to learn to design systems of care. Several participants mentioned other professions that have faced this challenge of scale, including:

- Pharmacists, who in the past few decades have faced a professional bifurcation: a few work in hospitals, as highly specialized members of patient care teams, while many more work in chain pharmacies, where their scope of work is highly determined by their employers, and they have little to no ability to develop relationships with patients/customers.
- Travel agents and bank tellers, who have found their professions significantly disrupted as consumers use new IT capabilities to take over the roles they previously monopolized.
- Airline pilots, who over the course of the past 100 years have moved from flying "by instinct" to gradually improving the work they do with the continuous advent of better instrumentation and navigation systems, safer and smarter planes, etc. This is a useful comparison in many ways because pilots still maintain an important and prestigious professional role, but now work in much more team-oriented ways, with major demands to maintain and upgrade their skills via simulation training and assessment, and with airplanes that are now intelligent enough to become active partners in flying themselves.
- Financial advisors, who mediate between consumers and their particular contexts/needs and complex financial management tools and algorithms.

What ABIM can do:

Participants described different roles ABIM could play in helping the profession of medicine recognize and adapt to these challenges. Some mentioned the potential value of providing practicing physicians with a venue for sharing experiences and resources with each other. Possibilities for this that were mentioned include ABIM creating an online community for physicians to share experiences of change with each other, or providing real-world or virtual opportunities for physicians to shadow more experienced providers or mentor residents or novices.

In addition, ABIM is seen to wield influence over training program curricula and assessments. Several participants mentioned that ABIM could engage directly with training programs to promote awareness of changes in the health care system and emphasize new skills and competencies related to teamwork, efficiency, focus on patient experience, and working as part of an ACO-type organization.

A final area of potential advocacy that some participants mentioned was related to the challenge of "scaling" physician work. The time- and resource-intensive nature of physician training is seen to pose a challenge for the profession. With this in mind, a few participants mentioned that ABIM could advocate for making physician training faster and more efficient. It was unspecified how to reconcile this image of faster, more focused physician training with the increasing complexities and competencies outlined in the previous sections. There are, however, pilot programs to test the feasibility of competency-based medical education, as opposed to primarily time-based systems; theoretically, such programs, if successful, could make physician training faster.

IV. General implications for ABIM

Much of what our participants predicted has implications for ABIM, both in terms of what the organization should do to prepare for the changes they describe (assess different competency areas, use different assessment methods, etc.), as well as in terms of what ABIM should do to lead or support policy initiatives to help the future take shape in the best way possible. In the section above, we outlined specific actions ABIM could take that were mentioned by our participants related to the themes of the interviews; in this section we provide a general overview of what our participants said might be in store for ABIM.

In general, our participants felt there can be a continued role for ABIM in updating and assessing knowledge and practice standards for internal medicine physicians. However, most predicted that certifying individual physicians, particularly via multiple-choice questions, could become both less viable and less relevant, for several reasons: First, the knowledge that individual providers can keep "in their heads" will become less important, as physicians will likely increasingly rely on expert systems to manage clinical knowledge and help make decisions about patient care. Second, fewer physicians will practice independently and will instead work as part of organized teams within care systems; the care patients receive will be less directly connected to their physicians' individual medical knowledge. Third, physicians in general may be less directly involved in patient care; their work may focus more on managing teams and designing care-delivery protocols and systems.

With this in mind, participants described several ways in which ABIM can stay relevant and helpful to the health care system as a whole. (Note that these are not mutually exclusive, but overlap and could be pursued in combination.)

One possible future direction for ABIM that many participants mentioned would be to move beyond certification of individual providers to certifying teams and care organizations. This came up particularly when participants focused on the question of what information patients will want about their providers in the future – individual providers' competence and performance may be less relevant than the competence and performance of the teams and systems within which they work.

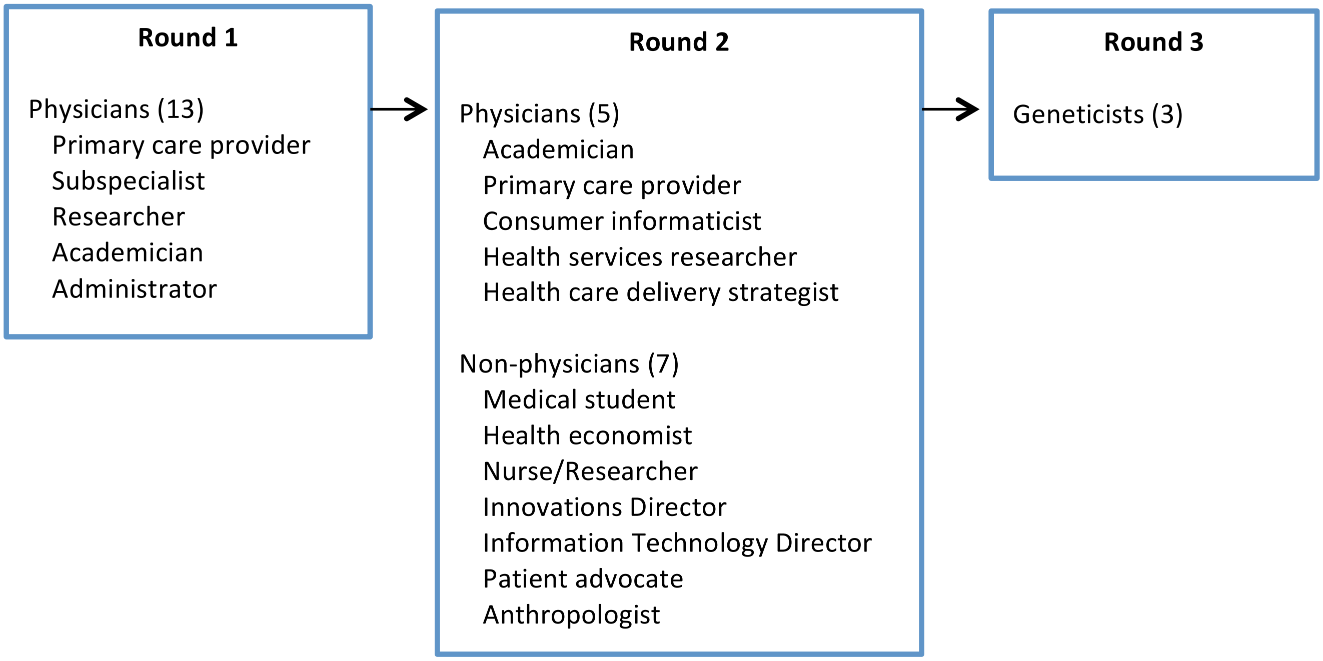
A second possible direction commonly mentioned would be to innovate assessments that look at physician performance in practice, particularly focusing on how physicians make decisions in ambiguous situations, and on the cost implications of their clinical decisions. Participants speculated on several approaches ABIM and other certification organizations could take, including extracting cost data directly from physicians' EHR systems, evaluating how well physicians integrate data about patient experience into their practice, and looking more directly at performance in practice, through individual or group simulation exercises or even through direct observation of physicians at work.

A third possible direction would be to assess physicians on the new competencies the changing health care system will require. Some of the specific competency areas that were most commonly mentioned by our participants were:

- Teamwork
- System design
- Public health/population care
- Genetics/personalized medicine
- Efficiency of care
- Integrated care planning
- Patient-centeredness (including communication, patient "activation," and attention to patient experience)

Finally, a substantial number of our participants expressed the hope that ABIM could help lead the profession of medicine towards engaging productively with the many disruptive challenges outlined in the sections above. They described several ways ABIM could do this: by engaging directly with training programs to promote educating future physicians to meet the new roles health care will require of them; by providing information to practicing physicians on how to adapt to the changing health care system and perhaps offering venues for practicing physicians to share information with each other or even to observe each other at work; and by defining and disseminating a clear vision for how physicians can adapt to the new roles that this changing, hopefully improving health care system will require.

Figure 1: Participants



Round 1

Physicians (13)
- Primary care provider
- Subspecialist
- Researcher
- Academician
- Administrator

Round 2

Physicians (5)
- Academician
- Primary care provider
- Consumer informaticist
- Health services researcher
- Health care delivery strategist

Non-physicians (7)
- Medical student
- Health economist
- Nurse/Researcher
- Innovations Director
- Information Technology Director
- Patient advocate
- Anthropologist

Round 3

Geneticists (3)

Interview guides

First round of interviews

1. First, we'd like to ask you to think about how the world has changed in the past 10 years or so for individual physicians. What have been the most significant changes, and how have they impacted individual physicians?
2. Thinking about how the world may change in the coming 10 years or so, what changes do you foresee that may have an impact on individual physicians?
3. How will the coming changes you've described impact the skills and competencies physicians will need to practice medicine?
4. Do you think there will be any new skills or competencies physicians will need in the next 10 years?
5. What will be the most effective tools and methods to know if physicians have the knowledge and skills they need to provide high quality patient care?
6. How important will the ability to access and integrate up-to-date information during the process of providing care be as a competency for individual physicians? How could a physician's ability in this area be assessed?
7. What information will patients want/need about their physicians in the future?
8. What else will be important for ABIM to do in 10 years?

Second round of interviews

1. Let's think about how you think the world may change for physicians in the coming 10 years or so. What changes do you foresee – or what present-day trends may continue – that will have an impact on individual physicians?
2. Now let's think about how the coming changes you've described for the next 10 years will impact how physicians will practice medicine. What are the skills and competencies physicians will need to practice medicine if these changes take effect?
3. What will be the most effective ways to know if individual physicians have the knowledge and skills they need to provide high quality patient care?
4. What will the patients in the world 10 years from now want to know about physicians? And how will they get that information?
5. What about the organizations that employ physicians or reimburse for their services – what information will these organizations want to know about physicians? And how will they get that information?
6. What do you think will be important for ABIM to do to be relevant and helpful in the future you've described?
7. To the extent that ABIM can influence practice, training and policy through its requirements – what should the organization do over the next 10 years to help bring about what you would see as the best possible future for the U.S. healthcare system?

Third round of interviews – geneticists

1. How are advances in genetics over the next 10 years likely to affect the practice of medicine?
2. What will these changes mean for physicians – the knowledge and competencies they will need, their relationship with patients, their use of technology, etc?
3. What should ABIM do over the next 10 years to account for these changes – how should our assessments and requirements change?

Appendix B:

Community Outreach Summary

Appendix B - Table of Contents

Assessment 2020 Blog Posts

- **Assessment 2020: What Skills Will Physicians Need in the Future?**
  (Harlan Krumholz & Richard Baron, Task Force)
  Posted on 12/17/2013

- **If Quality Care Begins with a Correct Diagnosis, Why Is the Diagnostic Error Rate So High?**
  (Robert Wachter, Task Force)
  Posted on 02/12/2014

- **If Everyone Agrees Teamwork Is Crucial to Providing Quality Patient Care, Why Is It So Hard to Achieve?**
  (Benjamin Chesluk, ABIM)
  Posted on 02/27/2014

- **Performing Common Ambulatory procedures: Is it Time to Reverse the Downward Trend?**
  (Patrick Alguire, Task Force)
  Posted on 03/20/2014

- **How Can Physicians Help Make American Health Care More Patient-Centered?**
  (Marilyn Mann, Task Force)
  Posted on 04/28/2014

- **The Necessity of Stewardship**
  (John Benson, President Emeritus, ABIM)
  Posted on 05/22/2014

- **Diagnostic Accuracy – Computers vs. Physicians (POLL)**
  Posted on 6/23/2014

- **How Does Technology Affect Patient Care? (POLL)**
  Posted on 8/18/2014

- **Closed-Book Exams in the Age of "Just Look It Up" Medicine**
  (Steven Durning, Task Force)
  Posted on 08/25/2014

- Open- or Closed-Book Exams? (POLL)
  Posted on 9/3/2014

- Dear New Doctor…
  (Donna Cryer, CEO of CryerHealth)
  Posted on 09/11/2014

- The Doctor-Patient Relationship (POLL)
  Posted on 9/18/2014

- Physician Assessment and the Hidden Curriculum
  (Harlan Krumholz, Task Force)
  Posted on 09/25/2014

- What Effect Does the Hidden Curriculum Have on Assessment? (POLL)
  Posted on 10/2/2014

- Can Medical Simulation Be Used to Assess Physicians' Procedural Skills?
  (William McGaghie, Task Force)
  Posted on 10/09/2014

- Should Simulation Be More Widely Used? (POLL)
  Posted on 10/17/2014

- The Challenges of Assessing Skills: Taking Patient Histories and Conducting Physical Exams
  (Jack Boulet, Task Force)
  Posted on 10/23/2014

- The Challenges of History-Taking/Physical Exam Skill Assessment (POLL)
  Posted on 10/30/2014

- Weighing the Benefits of Adaptive Testing
  (Bradley Brossman, ABIM)
  Posted on 11/06/2014

- Adaptive Testing: Yay or Nay? (POLL)
  Posted on 11/13/2014

- Building the Patient-Physician Relationship (VIDEO)
  Posted on 11/21/2014

- **How Do Patient Online Doctor Ratings Rate?**
  (Bradley Gray, ABIM)
  Posted on 12/04/2014

- **Online Physician Ratings (POLL)**
  Posted on 12/11/2014

- **Changing Technology and Physicians of the Future (VIDEO)**
  Posted on 01/09/2015

- **The Physician's Role in Reducing Costs (VIDEO)**
  Posted on 01/16/2015

- **Cost Effectiveness in Health Care (POLL)**
  Posted on 01/23/2015

- **How Can Technological Advancements Be Used to Improve Physician Assessment?**
  (Robin Guille, ABIM)
  Posted on 01/30/2015

- **Open- or Closed-Book Exams? (VIDEO)**
  Posted on 02/13/2015

- **Automated Scoring of Complex Performance Tasks: Caring for a Sick Patient**
  (Andre Rupp, Task Force)
  Posted on 02/19/2015

- **Assessing Teamwork Skills (POLL)**
  Posted on 02/27/2015

- **The Future of Physician Assessment (VIDEO)**
  Posted on 03/12/2015

- **"Patient Experience" Measures: How Would You Rate?**
  (Gerald Arnold & Rebecca Baranowski, ABIM)
  Posted on 04/03/2015

- **Striving Towards Assessment That Is Valuable to Those Being Assessed**
  (Kevin Eva, Task Force)
  Posted on 04/16/2015

- **Does Modern Technology Redefine What Makes a Good Doctor?**
  (Lauren Duhigg & Jonathan Vandergrift, ABIM)
  Posted on 04/30/2015

- **Assessment Engineering: The Science of Planning and Producing a Fair Exam**
  (Robert Cook, ABIM)
  Posted on 05/15/2015

- **Diplomates to Receive Additional Feedback on their Exams – Incorporates Diplomate Recommendations**
  (Robin Guille, ABIM)
  Posted on 6/4/2015

- **Can Professionalism Be Taught and Assessed?**
  (Elizabeth Bernabeo, ABIM)
  Posted on 6/22/2015

Assessment 2020: What Skills Will Physicians Need in the Future?

Harlan Krumholz, MD – Assessment 2020 Task Force Chair

Richard Baron, MD – Assessment 2020 Task Force member

Certification from the American Board of Internal Medicine (ABIM) is a trusted marker to patients when selecting doctors. As physicians, maintaining our certification reflects a commitment to our patients and to the profession. Certification and Maintenance of Certification (MOC) demonstrate to our patients, our peers and ourselves that we know what we need to know and do what we need to do to provide high-quality care. ABIM's obligation is to ensure that our products and programs are relevant and meaningful to both patients and physicians. To inform future enhancements to Certification and MOC, the Assessment 2020 initiative seeks to determine what competencies physicians will need as the field of medicine continues to evolve and to find the best ways to evaluate these skills.

The members of the Assessment 2020 Task Force include not only physician leaders, but also experts across a spectrum of professions related to performance evaluation and assessment.

Through this Assessment 2020 Blog, we seek to stimulate conversations among physicians and the public alike on a variety of topics such as:

- physician assessment;
- patient quality of care;
- the skills and competencies all physicians should have; and,
- how to factor advances in medicine into assessment.

Every few weeks, we'll add new blog posts which will be authored by experts both inside and outside the profession of medicine. We encourage your comments. Be sure to subscribe to the blog to receive the latest updates.

The success of this initiative depends on feedback from *you* – as physicians, patients and other stakeholders – to help us find out what it means to be a good doctor in the 21st century and how we can best evaluate whether doctors are meeting those standards. Thank you for your participation in this important work. We look forward to hearing from you.

If Quality Care Begins with a Correct Diagnosis, Why Is the Diagnostic Error Rate So High?

Robert Wachter, MD – Assessment 2020 Task Force member

There is a saying in the quality world that one should try to avoid creating situations in which clinicians are hitting the target but missing the point. The quality and safety movements of the last decade illustrate this danger. Namely, there are now scores of quality and safety measures, most of them reasonably evidence-based but none of which capture whether the initial diagnosis was correct. In a 2010 *Health Affairs* article (pdf), I wrote:

As one vivid example of how far we need to go, a hospital today could meet the standards of a high-quality organization and be rewarded through public reporting and pay-for-performance initiatives for giving all of its patients diagnosed with heart failure, pneumonia, and heart attack the correct, evidence-based, and prompt care – even if every one of the diagnoses was wrong.

This wouldn't be a big deal if diagnosis was easy, but it's not. In fact, making a correct diagnosis may well be the hardest thing we do in medicine. Diagnostic error remains the most common form of medical mistake, according to studies of closed malpractice claims. Tens of millions of dollars have been invested in trying to build computers that are better diagnosticians than doctors. So far, none has come into widespread use.

Improving diagnostic accuracy will require substantial work, in many forms:

1) Clinicians need to have the right information available at the right time. You can't diagnose the patient's hyper-coagulable syndrome if the medical record doesn't remind you that the patient had a thrombosis a few years back. The rapid wiring of American health care has skyrocketed in the past four years because of the federal Meaningful Use incentives and provides a hopeful start. However, we need better computer systems and seamless interoperability.
2) Clinicians need reliable sources of feedback on their diagnoses. The clinician who misses the diagnosis of lung cancer will never get better if he or she never learns that the patient later proved to have this tumor. Here too, the uptick in IT implementation is a helpful start, but more needs to be done to provide meaningful feedback and to encourage physicians to use it effectively.
3) Clinicians need better skills in heuristics and meta-cognition – thinking about their thinking. Here, we're making real progress. I recently presented a tricky case at our department's morbidity and mortality conference. The resident said that he thought the diagnosis was X, "but I'd be worried about anchoring bias." I nearly cheered.
4) Ultimately, computerized decision support, in the form of diagnostic checklists ("did you remember to consider…?") or even more advanced artificial intelligence will be helpful. The Isabel system is now reasonably effective and IBM's Watson team is focused on health care computing. These systems will need to be plugged into the EHR (without requiring separate data entry) and they have to be continuously learning from experience. One can envision a future decision-support system that constantly mines a hospital's database for outcomes, linked to admission variables: On admission, such a system may say to the clinician, "patients like yours ultimately proved to have lupus," in the same way that Amazon.com says "customers like you also liked *Harry Potter*."

In a generation or two, physicians' diagnostic skills may become less important when this task, like so many others, is taken over by computers. Until then, ensuring that physicians have the knowledge and skills they need to be expert diagnosticians will remain a crucial role for training programs and certifying boards.

If Everyone Agrees Teamwork Is Crucial to Providing Quality Patient Care, Why Is It So Hard to Achieve?

Benjamin Chesluk, PhD, American Board of Internal Medicine staff

Teamwork matters to good health care. When physicians and other care providers communicate well and collaborate, it makes the care patients receive better and safer, and it can make clinicians happier and more fulfilled by their work [1].

Everyone recognizes this, and has for decades and more, but actually improving teamwork in the diverse settings where providers practice has proved a complex and enduring challenge. The fee-for-service model—and short visit times required by any payment system—pressure physicians to work alone. Workplace hierarchies often interfere with open dialogue and collaboration between members of different professions and specialties. The members of a patient's care team may share the same space, but inhabit different social worlds, each with its own focus, jargon and professional culture [2]. In the face of all this, is it realistic to expect effective teamwork among providers in the U.S. health care system?

Many think it is, and are putting their time and money into making it so. For example:

- Numerous hospitals are instituting ways to bring patients' care teams together, such as geographic grouping of patients or interprofessional rounds [3].
- Primary care practices are implementing team-based models that aspire to take some of the pressure off the individual physician by allowing everyone in the practice to work at the top of their license [4].

All these innovators benefit from the new HRSA-supported National Center for Interprofessional Practice and Education, which brings together innovators from around the country and all over the world to learn from one another and share new ideas and approaches.

At ABIM, we are exploring looking at teamwork as a physician competency. In October 2012, we introduced TEAM, the Teamwork Effectiveness Assessment Module. TEAM is an innovative new self-assessment tool physicians can use to identify their team and get feedback from their teammates on how they, as physicians on that team, can improve their teamwork. We are excited about this new module and the enthusiastic response it's received from early users. We hope it helps contribute to the spread of effective teamwork throughout the health care system.

REFERENCES

1. Reeves S, Lewin S, Espin S, Zwarenstein M. Interprofessional Teamwork for Health and Social Care (Promoting Partnership for Health). 1st ed. 2010: Wiley-Blackwell.
2. Garman AN, Leach, DC, Spector N. Worldviews in collision: conflict and collaboration across professional lines. J Organ Behav. 2006. 27(7): 829-849.
3. O'Leary KJ, Sehgal NL, Terrell G, Williams MV. Interdisciplinary teamwork in hospitals: A review and practical recommendations for improvement. J Hosp Med. 2012. 7(1): 48-54.
4. Grumbach K, Bodenheimer T. Can health care teams improve primary care practice? JAMA. 2004;291(1): 1246-51.

Performing Common Ambulatory Procedures: Is It Time to Reverse the Downward Trend?

Patrick Alguire, MD – Assessment 2020 Task Force ex-officio observer

Many patients, especially those in primary care ambulatory settings, expect their personal physician to perform certain “minor” but needed procedures. Patients benefit from the continuity, convenience and (in certain situations) decreased cost when their personal physician performs a procedure. Yet, many general internists indicate that they do not feel comfortable performing common ambulatory procedures, citing inadequate training.[1] This is confirmed by other data showing the number of procedures performed by internal medicine physicians is declining.[2]

For reasons of safety and economy, it is understandable that hospitals restrict the performance of invasive procedures—particularly ones that may incur harm if done incorrectly—to physicians with special expertise. The declining experience and competence in common and relatively simple procedures in the ambulatory setting are less understandable.

Patients are now being treated by a large number of primary care physicians who do not have the skills to perform common ambulatory procedures. This is despite the finding that practicing physicians rate their interest in learning procedures, particularly ambulatory procedures, as highly as learning about scientific and other clinical topics. [3] What can be done?

Systematic instruction with simulators in workshop settings is reliable and cost-effective, and offers the opportunity to practice skills in a “safe setting”, thereby enhancing skill retention and minimizing errors. [4] Workshops with simulations have been shown to improve perceived competence, operator safety and patient outcomes.[5]

- Should professional societies strive to make procedural simulation training for common ambulatory procedures more accessible?
- Should professional societies establish ambulatory procedure registries that can facilitate earning Maintenance of Certification Self-Evaluation of Practice Assessment points and potentially earn higher reimbursement from insurers by documenting higher quality outcomes?
- Should ACGME incorporate a greater emphasis on procedural competence in residency training?
- Should ABIM assess procedural competence using simulation in the Certification and Maintenance of Certification programs?
- What are the next steps toward making this a reality?

We look forward to your feedback.

REFERENCES

1. Wickstrom GC, Kolar MM, Keyserling TC, et al. Confidence of graduating internal medicine residents to perform ambulatory procedures. J Gen Intern Med. 2000; 15: 361-65.
2. Wigton RS, Alguire P; American College of Physicians. The declining number and variety of procedures done by general internists: a resurvey of members of the American College of Physicians. Ann Intern Med. 2007 Mar 6; 146(5): 355-60.
3. Alguire PC. Teaching physicians procedural skills at a national professional meeting. Med Educ Online [serial online] 2004; 9:1. Available from http://www.meded.online.org. Accessed February 1, 2014.
4. Heppell J, Beauchamp G, Chollet A. Ten-year experience with a basic technical skills and perioperative management workshop for first-year residents. Can J Surg. 1995 Feb; 38(1): 27-32.
5. Fincher RM, Pogue LN, Cowan CF. Teaching correct and safe bedside procedures. Acad Med. 1991 Jul; 66(7): 396-7.

How Can Physicians Help Make American Health Care More Patient-Centered?

Marilyn Mann, Assessment 2020 Task Force member

In its 2001 report, *Crossing the Quality Chasm*, the Institute of Medicine (IOM) defined patient-centeredness as "providing care that is respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions"[1]. The IOM proposed that patient-centeredness be adopted as one of the key aims for quality improvement in health care.

One of the most important aspects of patient-centeredness is shared decision-making. Shared decision-making involves clinicians and patients making decisions together based on the best available evidence and patients' values, beliefs and preferences. Shared decision-making is not only essential for respecting patient autonomy (a patient's right to refuse or choose their treatment), but is also needed for beneficence (balancing the benefits of treatment against the harms) and non-maleficence (avoiding harm) [2-3]. Why then is shared decision-making not the norm in clinical practice?

A 2013 *Health Affairs* article argues that insufficient attention has been paid to the specific competencies needed by patients, providers and health care systems to optimize patient engagement [4]. Competencies needed by physicians include:

- Agreeing that patients should be part of the decision-making process.
- Establishing the patient's preferred role in decision-making.
- Identifying choices and evaluating the evidence in relation to the individual patient.
- Presenting the evidence, taking into account the patient's competencies.
- Helping the patient reflect on and assess the impact of alternative decisions with regard to his or her values and lifestyle - negotiating decisions with the patient; agreeing on a care plan; arranging for follow-up.

On a system-level, the authors argue that structural changes are necessary to facilitate shared decision-making. These could include information systems to link patients with decision aids and other resources, and team-based care that engages professional staff in helping patients with self-management and health literacy. For our work in determining the future of physician assessment, my fellow task force members and I are exploring the following:

- How can medical education and continuous professional development be modified to train physicians in the competencies needed to engage patients in shared decision-making?
- How can clinical practice guidelines promote shared decision-making?
- Do we need payment models that support and reward efforts to practice shared decision-making?
- How can we measure and assess whether shared decision-making is occurring?

Can shared decision-making mitigate overuse and underuse?

REFERENCES

1. Committee on Quality of Health Care in America and Institute of Medicine. Crossing the quality chasm: A new health system for the 21st century. Washington, DC: National Academies Press; 2001.

2. Elwyn G, Tilburt J, Montori, V. The ethical imperative for shared decision-making. Eur J Pers Cen Healthc. 2013; 1:129-31.

3. Stiggelbout AM, Van der Weijden T, De Wit MPT, Frosch D, Legare F, Montori VM, Trevena L, Elwyn G. Shared decision-making: Really putting patients at the centre of healthcare. BMJ. 2012;355:e256.

4. Bernabeo E, Holmboe ES. Patients, providers, and systems need to acquire a specific set of competencies to achieve truly patient-centered care. Health Aff. 2013 Feb; 32(2): 250-8.

The Necessity of Stewardship

John Benson Jr., MD - President Emeritus, American Board of Internal Medicine and ABIM Foundation

The prospect of health care consuming 20% of the GDP by 2020 is unconscionable so corrective actions have enormous urgency. There are some initiatives underway that address this issue and still others that need to happen in order to bring stewardship to the forefront of individual physicians and organizations at-large.

Through its admirable Choosing Wisely® campaign, the ABIM Foundation has promulgated the concept of stewardship of limited resources—especially unnecessary, even harmful, costs—as a clinical competence to be stressed to trainees. None too soon, especially since only 36% of physicians polled in 2013 feel they are responsible for rising costs or their reduction. Obvious proof that there is so much more ground to cover in this area.

As a start:

Some teaching hospital administrators, who see Graduate Medical Education's acolytes as a risk to their current modus operandi, must stop acting as competitors in a local technology arms race: pricing services without relationship to costs, skimping on nurse/inpatient ratios, counting outpatient clinics as losers and regarding premature readmissions as revenue.

ABIM could require candidates to achieve a perfect score on questions related to costs and redundant care as a requirement for admission to secure exams for initial certification or MOC.

ACP could grade use of resources through MKSAP questions.

CMS, which has the ultimate negotiating position in the form of reimbursement for Medicare services, could only accept negotiated bundled charges. It could also refuse payment for non-compliance with the Choosing Wisely recommendations.

Educators, if forced to adhere to stricter ACGME's accreditation standards, can reward suitable ordering behavior by trainees or require meaningful interventions.

The time is well past exhortation. The issue has been recognized for decades. Hard choices and penalties must go beyond training the next generation. 2020 is closing in.

Closed-Book Exams in the Age of “Just Look It Up” medicine

Steven Durning, MD – Assessment 2020 Task Force ex-officio observer

The field of medicine continues to grow at an extraordinary pace. Every year brings new advances in clinical care and best practices, and research shows physicians cannot possibly keep up [1].

Of course, today’s physicians have more resources than ever in the exam room to help them fill that knowledge gap and make appropriate clinical decisions. With a number of online, on-demand reference systems replete with the latest guidelines and information, there is no limit to what is available with a few quick clicks of a keyboard or swipes on a smartphone.

As we explore how to enhance physician assessment, one issue that the Assessment 2020 Task Force is considering is how much physicians should be allowed to rely on “looking it up” versus their formal training and education. Where this becomes an issue is during the ABIM Certification and Maintenance of Certification exams. At present, both are “closed-book” exams, meaning the look-up of information during the test is expressly forbidden and expectations are that physicians should know the material without having to look it up.

- In the modern era of medicine with endless resources available to physicians during the delivery of care, does a closed-book examination align with the actual practice of medicine?
- Are we assessing physicians’ skills in the best way possible?

Studies from the field of education have shown that just the act of taking an examination improves performance above and beyond simply studying for it, a fact that has been shown with both open- and closed-book exams [2,3]. Preliminary work in other fields suggests that this testing effect is either equivalent in open- and closed-book exams or may favor closed-book exams [4], but more research is needed to determine if this applies to the field of medicine.

How much does the “closed-book” examination emulate real-life experience and care delivery?
Some have argued that it would be better to assess the ability of physicians to deliver the best care with all of the resources available to them in their real-world practice settings.

Indeed, “looking up” specific aspects of a patient’s presentation can be helpful. For example, if there is too much unknown and/or unexpected information, it can aid the physician with making good decisions and delivering high-quality care.

Given the practical benefit of these resources in patient care, should assessment begin to allow “look it up” medicine in the testing room, too, in some instances?

REFERENCES

1. Adair JG and Vohra N. The explosion of knowledge, references, and citations: psychology’s unique response to a crisis. Am Psychol. 2003;58(1):15-23.
2. Roediger HL and Butler AC. The critical role of retrieval practice in longterm retention. Trends Cogn Sci. 2011;15(1):207.
3. Larsen D, Butler A and Roediger H. Testenhanced learning in medical education. Med Educ. 2008;42:959-66.
4. Moore R and Jensen PA. Do openbook exams impede longterm learning in introductory biology course? Journal of College Science Teaching. 2007;36(7):46.

Dear New Doctor…

Donna Cryer, JD – CEO of CryerHealth

Dear New Doctor:

Hello and welcome! I am your patient. They may not have mentioned this to you in medical school, but I am your partner. I am keenly invested in your success so that you can help me lead a successful, healthy life. The better you become at helping me reach my full physical, emotional and social potential, the more satisfied you will be with your career and the legacy you leave behind as a member of the medical profession.

While you may have been told that medicine today is about winning grants, authoring publications, following guidelines, checking boxes in EMRs or jumping through hoops for administrators or regulators, I am here to tell you a secret – it is still about you and me. **And as our society and your peers continue to include elements of our relationship in their assessments of how well you do your job, I think it is safe to say that we need each other now more than ever.**

I am active, engaged, empowered and—like many of the patients you will encounter—medically complex with multiple chronic conditions. I find it almost impossible to get the care I need to optimize my health potential and meet my life goals and I need your help. If I may be so bold, I'd like to make the following suggestions (since I've been a patient for a long time and you are just starting out):

1. **Seek first to understand, then to be understood.** Both Steven Covey and Dr. William Osler, stress the importance of listening. Ask me about my goals, preferences, values, literacy, health literacy, social supports and life circumstances. Your advice, prescriptions and referrals won't stand a chance if they don't fit into the context of my life. Even if we only have seven or 15 minutes, investing a portion of that time in understanding me will improve everything that comes later.
2. **Join a system or create a practice that prioritizes coordinated care.** Whether you choose internal medicine or a specialty, you and I will both benefit from systems, processes and staff that support you in focusing on diagnosis and treatment and supporting me as I carry out the treatment plan. Don't turn a blind eye to the importance of technological and/or personal infrastructure for things like appointment scheduling, refills of medication, exchanges of my labs and imaging data, and error recognition. If they aren't working for us, I need you to stand up and say so. Protest, serve on committees, change the vendor contract specs, etc. Lack of care coordination isn't simply inconvenient: it can kill me.
3. **Recognize that being a patient and being part of the health care environment is not my job.** *You* chose this field, went to school, took tests, interviewed--*you* wanted to be here. *I* did not. I have another job, I have a life. Work *with* me, train me, support me in becoming a better patient and thus, a better partner to you. How can I answer questions about my symptoms in a meaningful way if you never told me what I was supposed to be tracking or provided a framework in which to track them? Prescribe an app, a journal, a spreadsheet, a sticky note; send me reading materials or point me to a website or forum you judge credible before or after the visit. Provide a way for me to send my questions ahead of time so you have a chance to research and answer them. Send me your goals for the visit ahead of time while we're at it. Help me make the most of this opportunity to spend time with you and benefit from your expertise.

4. **Learn from me.** I have the advantage of being with me 24 hours a day, 7 days a week, present for every symptom and drug reaction. I do everything possible to be healthy – listening, reading, studying about patients just like me with my characteristics, disease progression and medication regimens. Don't be afraid or too prideful to listen to me, request my thoughts or opinions, or even suggest the strategies I have developed to your other patients as appropriate. If I had people offering to work for me for free, I would take them up on it. Besides, this is what partners do for each other.

So, New Doctor, I hope this was helpful. I look forward to working with you for many years to come.

Here's to both our health and happiness!

Sincerely,

Your New Patient

Physician Assessment and the Hidden Curriculum

Harlan Krumholz, MD – Assessment 2020 Task Force Chair

Lately, I have been contemplating the future of physician assessment and reflecting on the "hidden curriculum", or the attitudes, values and beliefs to which we are indirectly exposed throughout our education [1].

Traditionally, assessment questions seen throughout all physicians' medical training consist of a series of facts followed by a prompt to select the best test or treatment for that patient from a pre-defined multiple-choice list. To answer these questions, a physician would have to accept that it was possible to know the "right" answer without even having spoken with the patient – something most of us would object to.

On several occasions, I have found myself wanting to reject all the answers because the question did not provide sufficient information. How can physicians—and their subsequent assessors—know a right answer without understanding the values and preferences of the patient?

Consider the following question as an example:

A 70-year-old man is referred to you for advice on treatment following a pulmonary embolism. The pulmonary embolism was diagnosed when he presented with acute shortness of breath and chest discomfort three months ago. He was treated initially with low molecular weight heparin, and then with warfarin, aiming for an INR of between 2.0 and 3.0. No underlying risk factors for the embolic event were identified. The patient has had no bleeding episodes during therapy. His only other active medical issue is hypertension, which is controlled with a diuretic. His referring physician asks for your recommendation concerning anticoagulation at this time.

The best recommendation now is:

A) Stop the warfarin, and start low dose aspirin.
B) Continue his current dose of warfarin for three more months, then switch to low dose aspirin.
C) Continue warfarin indefinitely, but reduce the dose, aiming for an INR of 1.5 to 2.0.
D) Continue his current dose of warfarin indefinitely.

Let me be clear that this is not a question from an ABIM exam, but is intended to point out that there are situations where multiple answers could be correct, depending on the patient's preferences and underlying conditions. Here, any of these could be acceptable strategies.

Suppose it has been difficult to maintain an INR of 2.0 to 3.0, and the patient has been greatly troubled by the need for monitoring. Let's say he also enjoys biking, but has curtailed this activity because of concern about the risk of bleeding. If I explain to him in a way that is easy to understand that trials have shown a risk of a recurrent DVT of 7-9% per year with no treatment, and that this could be reduced by starting low dose aspirin to about 5% per year. If he prefers this risk to continued warfarin therapy (which could further reduce his risk), then this would be an appropriate strategy. But suppose he—or a family member—is very concerned about a recurrent pulmonary embolus and has not been bothered by the warfarin therapy or monitoring; in this case, continuing his current dose indefinitely would give him the lowest risk of a recurrent thrombotic event. Option B could reduce his chances of post-thrombotic symptoms, and if he were willing to continue warfarin for a few more months, might be best for him. If he had concerns about bleeding as well as a recurrent blood clot, then option C might be best.

While it might be tempting to boil down treatment decisions to multiple-choice answers, those who practice medicine know that the knowledge we gain through interactions with patients is critical to making the *right* decisions and the *best* recommendations.

As we look at the future of physician assessment through Assessment 2020, one of our tasks is to consider how we can account for the "hidden curriculum" in treatment.

- If a patient's values and preferences are elemental to decisions about care, how can they be incorporated into the testing environment?

- How should physicians best tested on core knowledge in an environment that does not allow for patient interaction (i.e., multiple-choice tests)?

REFERENCES

1. Jackson, PW. Life in the classrooms. 1st ed. New York(NY): Holt, Rinehart, and Winston, 1968.

Can Medical Simulation Be Used to Assess Physicians’ Procedural Skills?

William McGaghie, MD – Assessment 2020 Task Force member

Research shows that medical simulation in many forms can be a powerful mechanism to help physicians acquire and maintain clinical skills. There is no doubt that medical simulation works as a teaching tool, but what about physician assessment?

Medical simulations ranging from very low to very high “fidelity” to practice are currently being explored for physician assessment. An example of a low-fidelity simulation is a case-based multiple-choice question (MCQ), perhaps with an image or graphic data [1]. High-fidelity simulations are ones such as a computer-based interactive endovascular simulator in an angiographic suite, which presents interventional cardiology problems with or without complications [2]. A growing body of evidence shows that both levels of fidelity are promising in assessing physicians for things such as accuracy, quality, safety or even ethical character [3].

Medical simulation is a means, not an end. Simulations can be used in assessments either *for learning* (formative) or assessments *of learning* (summative), depending on the goals. The key is to match these goals with physician assessment tools, whether grounded in high- or low-fidelity medical simulation or other formats.

Technological advancements—mannequins, computer-based clinical problems, virtual reality avatars and many others—will continue to produce increasingly lifelike human simulations for use in medical teaching and testing. Some certification boards, including those for anesthesiology, surgery and internal medicine, are using simulation now in their Maintenance of Certification programs, while others are considering its utility [4].

This is the wave of the present, not the future. The challenge for medical educators and physician evaluators is to use these technologies intelligently to ensure that they produce reliable assessment data that can be used to make valid judgments about physician competence and to ensure that the benefits outweigh the development and implementation costs [5].

Do you think the added cost and complexity are worth the realism gained by high-fidelity simulations?

REFERENCES

1. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. Educ Res. 1995; 24(5): 511, 35.
2. Lipner RS, Messenger JC, Kangilaski R, et al. A technical and cognitive skills evaluation of performance in interventional cardiology procedures using medical simulation. Simul Healthc. 2010; 5(2): 65-72.
3. Scalese RJ, Issenberg SB. Simulation-based assessment. In: Holmboe ES, Hawkins RE, eds. Practical Guide to the Evaluation of Clinical Competence. Philadelphia: Mosby Elsevier, 2008; 179-200.
4. Feldman M, Lazzara EH, Vanderbilt AA, Diazgranados D. Rater training to support high stakes simulation-based assessments. J Contin Educ Heal Prof. 2012; 32(4): 279-288.
5. Fletcher JD, Wind AP. Cost considerations in using simulations for medical training. Mil Med. 2013; 178: 1037

The Challenges of Assessing Skills: Taking Patient Histories and Conducting Physical Exams

Jack Boulet, PhD – Assessment 2020 Task Force member

The abilities to take a patient's history and to perform a relevant physical examination are fundamental components of being a doctor. It could be argued that skills in these domains are directly, or indirectly, related to all of the Accreditation Council for Graduate Medical Education six core competencies, but the competency most directly related is, of course, "Patient Care." Assessing history-taking (HX) and physical examination (PE) skills can be difficult and subject to many potential sources of error which is something we strive to account for in the ABIM Certification and Maintenance of Certification programs.

In medical school, students are typically evaluated by faculty, both with real and simulated patient encounters. As part of the initial licensure process for physicians in the United States, HX and PE are measured in objective structured clinical examinations that employ standardized patients. In residency and/or as part of the Board certification process, both less-structured (e.g., observation of actual patient encounters) and more-structured (e.g., simulation) assessments can be undertaken. Regardless of how HX and PE are evaluated, there is a need to provide evidence that any derived estimates of ability are reliable and valid. Given the nature and variability of patient complaints and conditions, and the narrowing practice domains of many physicians, this can be difficult.

While HX and PE are both thought of as "skills," they vary substantially as a function of the patient complaint(s). A workup of a patient presenting with elbow pain will be quite different, and likely less challenging, from one presenting with dizziness. In the measurement world, this is known as task specificity. To address this issue, and obtain a reasonably precise estimate of ability, we often need to sample broadly, observing and evaluating students, residents and practitioners across many different types of patient encounters. This can, however, be a costly undertaking. Also, when individuals are evaluated by faculty or other assessors, training the evaluators—which is often not done or done poorly—and gathering adequate assessment data can be quite demanding.

For PE skills in particular, the number of individual techniques is quite vast. Furthermore, there can be significant differences in opinion regarding the value of specific maneuvers in terms of helping to make a correct diagnosis. Finally, even if consensus can be reached regarding the value of a PE maneuver, judging the quality of the technique, especially via observation, can be difficult and error-prone. The combination of these factors makes it difficult to develop defensible PE assessment tools. Without these tools, it is not possible to gather meaningful assessment data.

The evaluation of HX and PE skills can be quite challenging, regardless of whether this is done in a standardized setting or as part of ongoing workplace assessment activities. Unlike assessment of other "competencies" (e.g., knowledge, clinical reasoning), secure examinations cannot be effectively employed. The practitioner needs to be observed (preferably multiple times, across different patient presentations), and this introduces many measurement challenges.

Ultimately, to properly assess these core competencies, the health care community must acknowledge that until new assessment systems for HX and PE are developed, it is necessary to balance the need for valid and reliable assessment scores with the feasibility of gathering relevant performance data.

Weighing the Benefits of Adaptive Testing

Bradley Brossman, PhD – American Board of Internal Medicine staff

Adaptive testing has become a common form of exam administration in recent years [1]. Although several types of adaptive testing exist (e.g., CAT, Ca-MST), the basic premise is that items are selected for each candidate to match their proficiency level. For example, candidates who perform very well on an initial set of items are subsequently administered more difficult items, whereas candidates who perform poorly on an initial set of items are administered easier items. As a result, not only is each candidate administered a different *set* of items, but each candidate may be administered items that are much different in *difficulty.*

Can scores be reasonably compared against each other if each candidate took a different set of items with different levels of difficulty? Studies have demonstrated that they can. One showed that not only are exam scores *comparable* when different candidates are administered different sets of items (assuming that proper statistical methodology is used), but that the scores are actually more *accurate* and *precise* when the difficulty of the items are selected to match the proficiency levels of the candidates [2]. How can this be? In short, it is because the statistical procedures that are used to determine them under the adaptive testing framework take into account the fact that some candidates faced harder items and others easier ones.

Adaptive testing applies the same principles as traditional testing but on a more individual level. Whereas traditional exams are constructed with content and difficulty levels appropriate for the average test-takers' level, adaptive tests are constructed with appropriate content for the entire population of test-takers but at difficulty levels best suited for measuring *each candidate's* proficiency. In so doing, we obtain a more *accurate* measure of each candidate's proficiency, typically using a *fewer number of questions* than what is required in traditional testing. As such, an additional benefit to test takers is that they may actually take a shorter exam.

Adaptive testing should be used when the purpose of the exam benefits from this type of testing and if the number of test questions available support the use of it (there are situations where this is not the case). In fact, I would argue that the basic premise behind any adaptive test—namely, matching the difficulty of the items to the proficiency of the candidates—is already used for nearly any well-constructed test, adaptive or not. For example, calculus questions would not appear on a basic mathematics exam given to third graders. Why? Because the calculus questions would not only be too difficult for them but they would also not test the appropriate knowledge domain in the population of interest: basic mathematics in elementary school students.

Along similar lines, the content on the ABIM Certification and MOC exams should—and do—align with the knowledge of the physicians that these exams measure. A recently published article demonstrated that exam scores and pass/fail decisions obtained under the adaptive framework were more accurate than those obtained under the traditional framework specifically for ABIM exams [3].

What do you see as the benefits and challenges of adaptive testing?

REFERENCES

1. Drasgow F, Luecht RM, Bennett RE. Technology in testing. In: Brennan, P, ed. Educational Measurement. 4 th ed. Washington, DC: American Council on Education. 2006; 471-516.
2. Luecht R, Sireci SG. A review of models for computer-based testing. New York: The College Board, 2011.
3. Brossman BG, Guille RA. A comparison of multistage and linear test designs for medium-sized licensure and certification examinations. J Comput Adapt Test. 2014;2:18-36.

How Do Patient Online Doctor Ratings Rate?

Bradley Gray, PhD – American Board of Internal Medicine staff

With the proliferation of online information sources, the first stop for most consumers before making a purchase, hiring someone or even just dining out is the Internet. The same is true for choosing a health care provider: among consumers in the U.S. who consulted physician website ratings, a third reported either selecting and/or avoiding physicians [1] or hospitals [2] because of ratings they find online. But how much do these online ratings correlate to trusted measures of high-quality care and outcomes? According to new research, not much.

A recent study published in *JAMA IM* found that websites featuring physician ratings are not an accurate assessment of the quality of care that patients receive. Researchers from the American Board of Internal Medicine (ABIM) compared patient-submitted online ratings against measures of the quality of care delivered in the practices and found no correlation. The researchers compared 1,299 physicians' results from an ABIM PIM Practice Improvement Module® against the ratings physicians received on eight leading, publicly available and free health-based websites. Sites were selected from Internet searches in which each physician's name, specialty and city were entered into the Google search engine.

The comparison found that there was no statistically significant association between the online ratings given by patients to their physicians and the quality of care delivered in the practices. There was, however, a small association between the website ratings and patient experience scores collected through widely-used and standardized patient surveys. The researchers did note, however, that the weak correlation between the website ratings and quality measures could have resulted from the low number of website ratings per physician or an unrepresentative sample of patients leaving the ratings. They also note that the associations might have been stronger had narrative patient evaluations (i.e., comments) from the websites been used in the evaluation.

Could the problem just be the way we view and approach health care as consumers in America? Interestingly, a study in the United Kingdom found that patient rating of physician practices on a National Health Service website were *strongly* related to offline measures of patient experience and had some associations with clinical quality measures[3]. Either way, it is critical that these unreliable sources of information be improved or replaced with more valid ones on which consumers can rely.

As we think through the future of physician assessment with Assessment 2020, I think the reliability of online physician ratings is an important topic for us to consider. How, if at all, can the assessment ABIM and other boards provide inform efforts to improve the physician rating information available to patients online? After all, the Internet and consumer hunger for guidance in selecting physicians isn't going away.

REFERENCES

1. Hanauer DA, Zheng K, Singer DC, Gebremariam A, Davis MM. Public awareness, perception, and use of online physician rating sites. JAMA. 2014; 311(7):734-35.
2. Bardach NS, AsteriaPenaloza R, Boscardin WJ, Dudley RA. The relationship between commercial website ratings and traditional hospital performance measures in the USA. BMJ Qual Saf. 2013; 22(3):194-202.
3. Greaves F, Pape UJ, Lee H, et al. Patients' ratings of family physician practices on the internet: usage and associations with conventional measures of quality in the English National Health Service. J Med Internet Res. 2012;14(5):e146.

How Can Technological Advancements Be Used to Improve Physician Assessment?

Robin Guille, PhD – American Board of Internal Medicine staff & Assessment 2020 Task Force member

At ABIM, exam writing committees work iteratively and collaboratively. They start by identifying important concepts then refine their ideas, ultimately producing fully-formed questions after several qualitative reviews. This workflow encourages the development of well-formed questions that test important points [1].

Structured content design approaches that encourage the development of more focused test questions [2] are being explored by ABIM. At the same time, modern computer technology affords the opportunity to include audio or video clips on the exam. Multimedia clips are used whenever their inclusion serves to better align the test with what a doctor commonly does in practice [3]. For example, audio clips may make a lot of sense on a cardiology exam but may not have relevance to a hematology exam.

The formats of test questions are also becoming more innovative with improved technology [4]. For instance, questions on the sequencing of tasks in practice can be conveniently tested using the *drag-and-drop* item format, where test-takers use their computer mouse to slide onscreen objects into pre-defined zones. This is a more natural way to capture the sequence than by presenting—say—five predetermined sequences as multiple-choice format options.

Overall test design is also improving with technology. One idea ABIM is currently experimenting with is adaptive designs for the certification exams, which choose in real-time the next exam question based on responses collected thus far. Research has shown that adaptive testing could cut exam time in half for a large portion of examinees [5].

In addition, ABIM is exploring packaging different kinds of assessments into small bundles of questions called "testlets," which can be scored independently but administered contiguously. For example, two testlets of our current multiple-choice questions, which test knowledge, could be combined with one testlet of medical simulation that tests performance. The final score might be a combination of the weighted testlet scores. This could allow for more focused examinations that provide richer feedback.

Regardless of how technology changes the particular features of assessment, standards of quality must remain. Summative assessments like ABIM certification exams are judged by how *valid* the scores are, i.e., by how meaningfully they can be used [6]. Formative assessments like the ABIM medical knowledge modules and PIMS - Practice Improvement Modules® are judged by how much *value* they add to the learning experience [7]. As long as these core testing values are honored, the addition of new technologies can only continue to enhance assessment.

REFERENCES

1. How exams are developed. American Board of Internal Medicine Web site. http://www.abim.org/about/examInfo/developed.aspx. Accessed January 28, 2015.
2. Mislevy, RJ, Almond, RG, Lukas, JF. A Brief Introduction to Evidence-Centered Design. Los Angeles, CA: University of California, Center for the Study of Evaluation; 2004.
3. Raymond, MR. A practical guide to practice analysis for credentialing examinations. Educ Meas. 2002; 21(3): 2537.
4. Parshall, CG, Harmes, JC, Davey, T, Pashley, P. Innovative items for computerized testing. In: van der Linden, WJ, Glas, CAW, eds. Computerized Adaptive Testing: Theory and Practice. Norwell, MA: Kluwer; 2000.
5. Brossman, BG, Guille, RA. A comparison of multistage and linear test designs for medium-sized licensure and certification examinations. JCAT. 2014; 2(2).
6. Messick, S. Validity. In: Linn RL, ed. Educational Measurement. 3 rd ed. New York, NY: MacMillan; 1989.
7. The value of formative assessment. FairTest: The National Center for Fair and Open Testing Web site. http://www.fairtest.org/value-formative-assessment-pdf. Accessed January 28, 2015.

Automated Scoring of Complex Performance Tasks: Caring for a Sick Patient

André Rupp, PhD – Assessment 2020 Task Force member

Developing assessments that can be reliably administered for large-scale activities like certification and licensure is a very difficult task. Demands for immediate or short-delay score reporting are often in conflict with the needs for high levels of authenticity and face validity in assessment. Agencies may often make various compromises when developing an assessment in order to provide reliable scores and valid interpretations of those scores.

There are many different methods to test someone's knowledge of a certain subject; the most popular method involves multiple choice questions (MCQ). MCQs are popular since they generally take a short time to answer and one assessment can cover a broad range of knowledge, skills and abilities. They can also be objectively scored in an automated way which yields reliable scores. For these reasons, MCQS are the general go-to format for high-stakes assessments that are administered at large scales.

At the same time, MCQs have natural limitations in how well they can measure more complex skills and abilities. Enhancing MCQs with embedded videos or small simulations can make questions feel more authentic, but the answer format is still limiting. Assessment agencies needing to measure more complex reasoning skills, such as arriving at a clinical decision using information from multiple sources, may include extended MCQS with a written component, interactive case studies with a sequence of decisions, and other more complex assessment methods [1].

More complex assessment methods are a valuable idea conceptually, but it can be very difficult to score them in automated ways. For example, automated scoring for an interactive case study designed to assess a complex performance task, like caring for a sick patient, might require the creation of scoring rules for every possible combination of choices. In an interactive environment, some of these choices may include:

- Did the physician obtain a thorough and appropriate patient history?
- Did the physician order the correct tests given the initial information?
- Did the physician evaluate the test results appropriately and come to a clinical decision?

Although this scoring method eliminates the possibility of scoring errors, the development of these extensive rules can be very time-intensive and requires that all logical choices be known and captured before the assessment is administered.

Parts of assessment with longer answers in either spoken or written form require natural language processing (NLP) to be automatically scored. The challenge in automatically scoring with NLP is that the assessment must identify features in an examinee's response that represent reasonable proxies for the kinds of performance characteristics that human rates would identify as the most important. Opponents of these methods challenge tools like this since it is very difficult to approximate scoring for creativity, nuanced reasoning and other more complex constructs (see [2] for writing assessment).

Finding the appropriate approach for automated scoring for an assessment is based on the integration of different scoring techniques used for different components of the assessment. This is as much a science as it is an art, especially for highly authentic interactive assessments. However, there are frameworks, principles and best practices to provide some guidance [3].

So what do you think – should assessment agencies try to assess complex performance tasks, like caring for a sick patient, with assessments that include automated scoring?

REFERENCES

1. Williamson DM, Mislevy RJ, & Bejar I I, eds. Automated scoring of complex tasks in computer-based testing. Mahwah, NJ: Erlbaum; 2006.
2. Perelman L. 2014). When "the state of the art" is counting words. Assessing Writing 2014;21:104-11.
3. Williamson DM, Xi X, & Breyer FJ. A framework for evaluation and use of automated scoring. Educ Meas; 2012: 31(1):213.

“Patient Experience” Measures: How Would You Rate?

Gerald Arnold, PhD, & Rebecca Baranowski – American Board of Internal Medicine staff

A natural consequence of the interaction between patients and physicians is how patients perceive their *experience* of health care. When asked about the quality of their health care, most Americans focus on how their doctor interacts with them and the issues related to their appointment [1]. What some researchers have deemed “humanity of care” includes aspects such as dignity and respect, privacy and wait time [2].
Like morbidity or re-admission rate, patient-reported experience measures (PREMs) have a distinct place in the way we assess a physician’s skills. PREMs cover a broad spectrum of quality factors:

- communication about care and treatment options;
- the perceived respect for patients;
- timely and coordinated care;
- patient participation in care decisions; and
- overall patient satisfaction.

A systematic review of clinical studies indicates that PREMs can correlate with the safety and quality of care provided by physicians [3]. However, PREMs are not usually associated with quality and safety processes outside the patient’s experience, such as the technical elements of a surgical procedure. Well-designed questionnaires can capture important patient viewpoints that are separate from whether a patient likes a physician and whether the parking lot is too far from the office [4].

Should PREMs be an important aspect of physician assessment? The Centers for Medicare and Medicaid Services (CMS) say yes. The Consumer Assessment of Healthcare Providers and Systems (CAHPS) surveys provide patients with an opportunity to rate their doctor on communication, shared decision-making and other important aspects of patient experience [5]. These ratings are publicly reported and used to determine incentive payments and payment adjustments by CMS. Moreover, PREMs can be used to drive quality-improvement efforts by individual physicians or institutions and are considered important assessments by health care advocacy organizations such as the Institute for Healthcare Improvement [6] and the Robert Wood Johnson Foundation [7].

Patient experience, when measured properly, shows sensitivity to a range of factors related to patient care and plays a critical role in demonstrating that certified internists provide excellent care for their patients.

Physicians, have you used PREMs to assess and improve your practice? How would your patients rate you on the humanity of care scale?

Patients, how would you rate your most recent experience receiving care?

REFERENCES

1. The Associated Press and NORC Center for Public Affairs Research. Finding quality doctors: How Americans evaluate provider quality in the United States. 2014.
2. Black N and Jenkinson C. How can patients’ views of their care enhance quality improvement? BMJ 2009;339:b2495.
3. Doyle C, Lennox L, and Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open. 2013 Jan 3; 3(1). pii: e001570. doi: 10.1136/bmjopen2012001570. Print 2013.
4. Manary MP, Boulding W, Staelin R, and Glickman, SW. The patient experience and health outcomes. New Engl J Med. 2013; 368:201-203.
5. 2014 CMScertified survey vendor reporting made simple v1.2 4/23/14. (pdf)
6. Institute for Healthcare Improvement: The Triple Aim.
7. Robert Wood Johnson Foundation: Aligning Forces for Quality. Good for health, good for business: the case for measuring patient experience of care. 2011.

Striving Towards Assessment That Is Valuable to Those Being Assessed

Kevin Eva, PhD – Assessment 2020 Task Force Member

As the ABIM nears completion of its Assessment 2020 project, it is a good time to pause and reflect on the foundational reasons for having Certification and Maintenance of Certification (MOC) practices. Regulatory organizations provide an essential service to the public by ensuring (as well as possible) that individuals are granted the privilege of practicing medicine only if they meet certain quality assurance standards.[1] To be effective, those processes should be designed and implemented with both fairness to the practitioners and responsibility to the public as core values. Fairness requires consideration of benefits to the practitioner, but it is not a license for individual control over MOC processes. While it is much easier to believe that individuals can determine for themselves whether or not they are practicing safely and effectively, that romantic notion is simply untenable.[2]

Assessment--even of the multiple-choice exam variety--can play an important role in identifying those who are capable of performing to the level of professional standards.[3,4] To achieve that goal, it is critical that rigorous standards are maintained. Achieving those standards in a way that considers fairness to the practitioner requires additional thought regarding how to build assessment processes that ***support*** rather than simply ***measure*** the continuing professional development of physicians. All licensing, certification and regulatory agencies for health professionals in North America invest considerable energy and expense to generate the best possible data regarding how well practitioners meet the expectations of their profession. It would be wasteful to then ignore how those data might be used to maximize impact on performance improvement.

Unfortunately, effectively using data for quality improvement initiatives is more complex than our intuition would generally lead us to believe. It is far too easy for those responsible for quality assurance processes to idealize the extent to which recipients of data will neutrally and rationally accept it as information that can be applied and utilized to improve their own performance.[5] Any data that conflict with one's professional identity create a threat to the recipient[6] that can be easier to overcome by discounting the data than by working to discern both why one's performance wasn't as strong as expected and what can be done to improve performance in the future.[7] This is especially true in complex environments where experience does not guarantee ability, where trustworthy data about performance is not routinely received, and where the influences on any outcome are multifaceted with only some residing in the practitioner's control.[8]

Most models of feedback in the health professions focus on how to deliver information effectively without recognizing that receptivity to the feedback is likely a greater determinant of its influence.[9] Receptivity to feedback relies on it being deemed credible both in terms of its validity and its intent (i.e., being perceived both as containing content that is meaningful and as coming from a trusted source motivated to enable the recipient to improve without threat). Anything less leads people to treat assessment exercises as hurdles that must be passed rather than as valuable educational opportunities.[10]

This is the ultimate challenge for regulatory authorities as credibility of intent is not easy to prove in high stakes contexts. Emerging literatures, however, suggest strategies that may yield benefit:

(1) Normalizing the improvement process to focus attention and activity beyond those individuals who reside at the bottom of the performance continuum. Setting a cut-score and ignoring everyone above it is misleading and can imply that those who "pass" are as good as they can be. This, unfortunately, ignores the fact that absolute mastery is rare, if not impossible, in the complex world of medicine;

(2) Providing guidance regarding what can be done to improve rather than simply telling recipients where they stand. It is assumed that most practitioners would not deliberately choose to practice below an understood standard and, as such, explicit information regarding how to achieve that standard will be considerably more valuable than data that only defines one's strengths or weaknesses; and,
(3) Striving for a continuous system that is integrated across the various stages inherent in the career of a health professional. Shared accountability is likely to be crucial given that practice does change over time and point-in-time hurdles are often treated simply as things that need to be overcome before returning to reality.

I do not mean to imply that these are easy issues to solve. We are, however, at a fortunate point in history in which there are many organizations and talented individuals working on these challenges. For example, many of these ideas in this blog post were developed and refined through the practice-derived insights and the research generated by members of the Assessment 2020 Task Force, the Medical Council of Canada's Medical Education Assessment Advisory Committee,[11] the Centre for Innovation at the National Board of Medical Examiners, and through consensus building processes recently engaged in jointly by the American Medical Association's Council on Medical Education and the American Board of Medical Specialties.[12] These types of collaborations are key because it is unlikely that any one organization could address all of these issues independently.

Promising strategies currently being discussed include creating assessment systems that directly follow from learning activities and result in the results cycling back to the practitioner to create further tailored learning plans. One goal in such a system is to facilitate the use of practice areas/professional development activities to generate authentic and meaningful assessment systems that mark progress and provide support. In such a model, considerable collaboration would be required to enable learners to discover the limits of their knowledge/ability rather than trying to convince feedback recipients that any set of data generated on their behalf defines their limits.

Prioritizing MOC systems that enable improved patient care must be the fundamental goal. Doing so requires that we don't assume synonymy between (a) quality assurance and quality improvement, (b) reliable and useful, (c) precise and actionable, or (d) the desire to practice well with the desire to be told how well one practices.[13]

REFERENCES

1. Randall GE. Understanding professional self-regulation. Ontario Association of Veterinary Technicians. oavt.org/self_regulation/docs/about_selfreg_randall.pdf Accessed October, 6, 2014.
2. Eva KW, Regehr G, Gruppen LD. Blinded by "insight": Self-assessment and its role in performance improvement. In: Hodges BD, Lingard L, editors. The question of competence: Reconsidering medical education in the twenty-first century. New York: Cornell University Press; 2012. pp. 131-54.
3. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. Annals of Internal Medicine 1989;110:719-26.
4. Tamblyn R, Abrahamowicz M, Dauphinee D, Wenghofer E, Jacques A, Klass D, Smee S, Blackmore D, Winslade N, Girard N, Du Berger R, Bartman I, Buckeridge DL, Hanley JA. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. 2007;298:993-1001.
5. Eva KW, Regehr G. Effective feedback for maintenance of competence: From data delivery to trusting dialogues. CMAJ 2013;185:463-4.
6. Kluger AN, van Dijk D. Feedback, the various tasks of the doctor, and the feedforward alternative. Medical Education. 2010;44: 1166–74.
7. Eva KW, Armson H, Holmboe E, Lockyer J, Loney E, Mann K, Sargeant J. Factors influencing responsiveness to feedback: On the interplay between fear, confidence, and reasoning processes. Advances in Health Sciences Education. 2012;17:15-26.
8. Eva KW. Diagnostic error in medical education: Where wrongs can make rights. Advances in Health Sciences Education. 2009;14:71–81.
9. Shute VJ. Focus on formative feedback. Review of Educational Research. 2008;78:153–89.
10. Sargeant J, Eva KW, Armson H, Chesluk B, Dornan T, Holmboe E, Lockyer J, Loney E, Mann K, van der Vleuten C. Features of assessment learners use to make informed self-assessments of clinical performance. Medical Education. 2011;45:636-47
11. Eva K, Bordage G, Campbell C, Galbraith R, Ginsburg S, Holmboe E, Regehr G. Medical Education Assessment Advisory Committee Report to the Medical Council of Canada on Current Issues in Health Professional and Health Professional Trainee Assessment. For the Medical Council of Canada; 2013.
12. Hawkins R, et al. The ABMS MOC Part III Examination: Value, Concerns and Alternative Formats. White paper produced by the American Board of Medical Specialties. 2015.
13. Mann KV, van der Vleuten C, Eva K, Armson H, Chesluk B, Dornan T, Holmboe E, Lockyer J, Loney E, Sargeant J. Tensions in informed self-assessment: How the desire for feedback and reticence to collect and use it conflict. Academic Medicine. 2011;86:1120-7.

Does Modern Technology Redefine What Makes a Good Doctor?

Lauren Duhigg and Jonathan Vandergrift – American Board of Internal Medicine staff

In the United States, the skills and characteristics that define a “good physician” have evolved over time. For example, a greater emphasis is now placed on communication and interpersonal skills, qualities that are increasingly being evaluated as part of the admissions process for medical school. In part, this is due to the realization that a patient’s experience with their health care provider is a key component of health care quality [1].

Additionally, new technology and tools to support physicians’ work are emerging at an increasingly rapid pace. These advances will likely alter the way we judge the cognitive and technical skills of a physician, similar the importance of patient-centered care in how the quality of our health care systems is judged. For example, the medical knowledge necessary to make clinical decisions needs to be readily and reliably accessible. In the past, this was largely via recall. Now, this information is increasingly accessible in real time, particularly via mobile devices. As stated by Dr. Stephen Klasko, president of Thomas Jefferson University, “It used to be if I knew 19 reasons someone had a disease and you only knew 15, I’d be considered the better doctor. But now we have all that information on our iPhones.”

Changes in required skills and technology, along with the modifications of assessments to account for these changes (e.g., via open- versus closed-book exams) have been discussed in prior blogs. However, there is little evidence to date documenting which of the characteristics and skills will become the most valuable in our rapidly changing health care environment. We strongly encourage all physicians reading this, young and old, to speculate on our behalf. In particular, we ask you to think about the colleagues you look up to as ideal physicians:

- What characteristics differentiate your exceptional colleagues?
- What skills do you see becoming more important over the next ten years?
- How will technological advancements affect the manner in which we assess physicians?

We welcome your feedback.

REFERENCES

1. Institute of Medicine. (2001). Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press.

Assessment Engineering: The Science of Planning and Producing a Fair Exam

Robert Cook – American Board of Internal Medicine staff

Designing an exam that fairly and accurately reflects proficiency in any area of knowledge or skill is not a simple task. The task becomes even more difficult when the exam is used to make high-stakes decisions concerning a subject matter that impacts the care and wellbeing of patients. Responsible exam development requires that the content represents the full depth and breadth of the subject being assessed while also providing enough psychometric information for valid inferences to be made about an examinee's ability in that subject area [1]. Although it is often difficult to create an exam that balances these concerns, the Assessment Engineering framework is a structured approach to exam development that replaces the traditionally delicate act of balancing these concerns with a systematic method that considers both of these important factors at every step of the exam development process [2].

ABIM is increasingly turning to Assessment Engineering methodology to inform how it develops exams and test questions that aim to accurately assess physician proficiency in Internal Medicine and its more than twenty subspecialties. Exam committee members use a systematic approach to derive testing points that reflect the skills and level of expertise ABIM Certification is intended to represent. Specifically, ABIM uses a process called prototyping to develop multiple choice test questions in a systematic way that keeps content experts focused on the core testing points while walking them through steps designed to ensure that each question includes essential content with an evidence-based, single-best correct answer along with plausible but distinctly incorrect alternatives. The final test content that is delivered to the examinee is then assembled automatically using criteria designed to ensure a fair balance of content is provided and that psychometric information goals are met [3].

As part of the Assessment 2020 initiative, now is the perfect opportunity to discuss ideas for how we can take Assessment Engineering approaches further at ABIM and in board certification writ large. For example:

- Our best prototypes can be turned into models for producing questions that address similar tasks in important but previously unaddressed, hard to write, or quickly changing content areas.
- Testing points can be generated independently of prototypes and test questions so they can provide a more thorough understanding of what a board certified physician is.
- Testing points can be established systematically through practice analyses that carefully walk through the tasks physicians perform and identify the things they do that are essential to quality patient care.

You do not build a bridge by starting at both ends and hoping they meet in the middle; you plan carefully before you start building. Similarly, you do not build a good exam by writing questions that you hope line up with what you want to assess. Instead, you plan carefully before you start building, which is what we do at ABIM.

As with medicine, assessment is a rapidly evolving field, with frameworks like Assessment Engineering representing some of the best thinking on current and future best practices. As ABIM strives to keep its assessment methods aligned with the current best practices, we look to you, the future of health assessment, to share your thoughts.

How can we use Assessment Engineering or other similar exam design methodologies to build a better bridge between medical credentialing and medical practice?

REFERENCES

1. Burke M, Devore R, Stopek J. Implementing assessment engineering in the Uniform Certified Public Accountant (CPA) examination. Journal of Applied Testing Technology. 2014. SI:134.

2. Luecht R. Assessment engineering in test design, development, assembly, and scoring. Keynote presentation at: Annual Meeting of the East Coast Language Testing Organizations (ECOLT); October, 2008; Washington, DC. http://www.govtilr.org/Publications/ECOLT08AEKeynoteRMLuecht 07Nov08[1].pdf. Accessed January 28, 2015.

3. How exams are developed. American Board of Internal Medicine website. http://www.abim.org/about/examInfo/developed.aspx. Accessed January 28, 2015.

Diplomates to Receive Additional Feedback on Their Exams – Incorporates Diplomate Recommendations

Robin Guille – American Board of Internal Medicine staff

We've heard from many physicians who, when they get their exam results, want to have more information so they can better understand their results. Working with input from ABIM Board Certified physicians, we are launching a new score report that we hope addresses many of the improvements physicians have asked for. The new score report is now distributed electronically and features clearer graphical explanations and more detailed feedback on performance.

Through our conversations with physicians, we learned together how the exam question feedback could be improved: Physicians want more specific information regarding the exam questions they missed. At the same time, we heard that the report needed to be simplified to be most useful. I am optimistic the improved report addresses both these important concerns.

Prototyping and Usability Study with Physicians

The process to revise the exam score report was rigorous. To identify best practices, ABIM staff researched similar testing organizations, bringing the insights gained to a focus group of randomly sampled ABIM Board Certified physicians. The focus group participants used those insights to inform an in-depth discussion on what physicians seek in a score report and what information might be missing, and to identify what was confusing about the old report. The focus group discussion led to the development of two initial prototypes, iteratively refined, based on comments from more physicians. A "hands-on" usability study—point-by-point review of each prototype—was then conducted by randomly selected ABIM Board Certified physicians. The ultimate draft design—a blend of the two prototypes— was presented by electronic survey to physicians representing all of the ABIM exam committees for final input and further refinement.

What We Learned

As an ABIM staffer, the honest face-to-face conversations with physicians were invaluable. While it was sometimes hard to hear that the old score report had too much technical jargon and missed the mark, it was wonderful to see that the new design is jargon-free and lengthy explanations have been replaced with hyperlinked, supplemental Web material.

There were so many voices heard in the development of the new report that this process can serve as a model for ways we can continue to work closely with the physician community to enhance our products and programs. We appreciate the feedback we receive from the internal medicine community, and I encourage you to continue sharing your thoughts and ideas with ABIM's CEO at rbaronmd@abim.org.

Link to Page 1 of Score Report
Link to Page 2 of Score Report
Link to Page 3 of Score Report

Can Professionalism Be Taught and Assessed?

Elizabeth Bernabeo – American Board of Internal Medicine staff

While most agree that the medical profession is responsible for promoting the professional behavior of students and practitioners, several ongoing challenges to teaching and assessing professionalism persist.

One challenge is that there is a lack of consensus around what professionalism actually is.
There are multiple ways to define 'professionalism,' some of which are not congruent with one another. Individuals using one definition may not acknowledge the potential validity of others [1,2]. Some believe that professionalism is not a stable construct that can be isolated, taught and assessed but rather a set of sophisticated and socially constructed skills that can be refined over a lifetime [3,4].

Another challenge is that context for professionalism varies.
Many believe that the assessment of professionalism requires consideration of the individual (their attributes, characteristics, attitudes, behaviors and identities) as well as interpersonal (relations, group dynamics) and societal (economic, political) dimensions [e.g., 2,4,5,6]. Context is essential in understanding lapses in professionalism [7]. In a more recent study, Ginsburg et al. identified a set of complex, interactive guiding principles and modifiers as critical to understanding physicians' responses to professional dilemmas [8]. One implication of this view is that professional practice may involve practitioners finding not so much the "right" answer (which may not always exist in an absolute sense), but rather deciding what is "best" in the situation in which they find themselves [9]. Recent work incorporating reflection into physician decision-making is consistent with this view [10,11].

Taken together, these challenges suggest that the teaching and assessment of professionalism is not a simple, one-size-fits-all task. In addition to situational context, assessors must consider the role of individual, social and institutional-level factors in their assessment, and be transparent about what the assessment effectively measures, as well as what it may not.

- Do you think that professionalism is a competence that can be assessed? If so, how?
- What are some of the most significant contextual issues that might impact the assessment of professionalism? Why are they important?

REFERENCES

1. Hafferty F, Castellani B. The Increasing complexities of professionalism. Acad Med. 2010;85(2):288-301.

2. Hodges BD, Ginsburg S, Cruess R, Cruess S, Delport R, Hafferty F, Ho MJ,Holmboe E, Holtman M, Ohbu S, Rees C, Ten Cate O, Tsugawa Y, Van Mook W, Wass V, Wilkinson T, Wade W. Assessment of professionalism: Recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(5):354-63.

3. Lesser CS, Lucey CR, Egener B, Braddock CH, Linas SL, Levinson W. A behavioral and systems view of professionalism. JAMA. 2010;304(24):2732-37.

4. Martimianakis MA, Maniate JM, Hodges BD. Sociological interpretations of professionalism. Med Educ. 2009;43(9):829-37.

5. Hafferty F, Levinson W. Moving beyond nostalgia and motives: towards a complexity science view of medical professionalism. Perspect Bio Med. 2008 Autumn; 51(4):599-615.

6. Wear D, Kuczewski MG. The professionalism movement: can we pause? Am J Bioeth. 2004;4(2):110.

7. Ginsburg S, Regehr G, Hatala R, et al. Context, conflict, and resolution: a new conceptual framework for evaluating professionalism. Acad Med. 2000;75(10 Suppl): S6S11.

8. Ginsburg S, Bernabeo E, Ross KM, Holmboe ES. "It depends": results of a qualitative study investigating how practicing internists approach professional dilemmas. Acad Med. 2012;87(12):1685-93.

9. Coles C. Developing professional judgment. J Contin Educ Health Prof. 2002 Winter;22(1):310.

10. Bernabeo EC, Holmboe ES, Ross K, CheslukB, Ginsburg S. Utility of vignettes to stimulate reflection in professionally challenging situations: theory and practice. Adv Health Sci Educ Theory Prac. 2013 August;18(3):463-84.

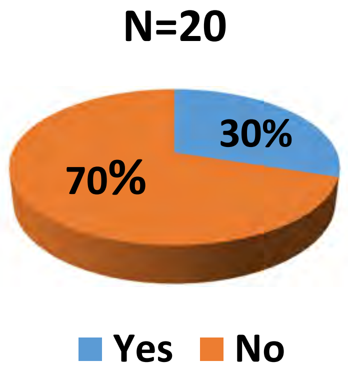
11. Bernabeo E, Reddy S, Ginsburg S, Holmboe E. "I am a wimp": internists reflect on factors that drive their approaches to professional dilemmas. J Contin Educ Health. In press.

Assessment 2020 Poll Questions & Results

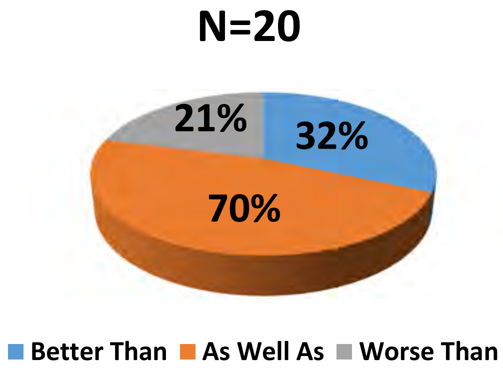
Responses as of 4/30/15

TECHNOLOGY

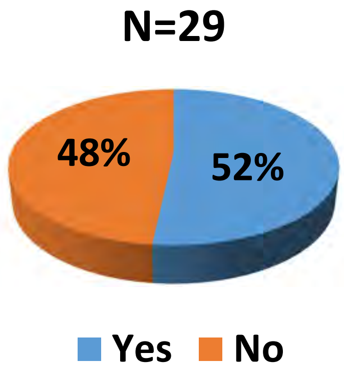
As technology advances, do you think computers will eventually replace physicians as diagnosticians?



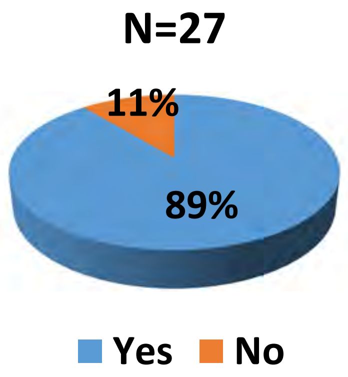
Do you think computers will diagnose health problems _______ physicians?



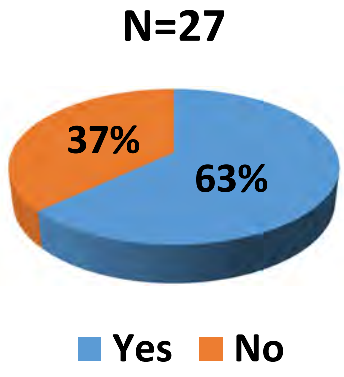
Does technology negatively impact physician-patient engagement?



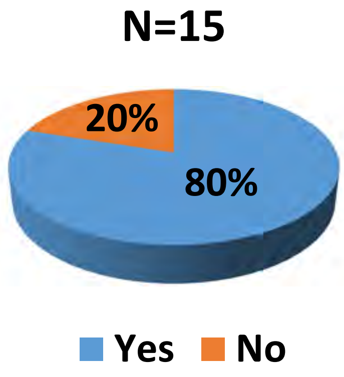
Will mobile technology be an important part of managing/improving patient health in the future?



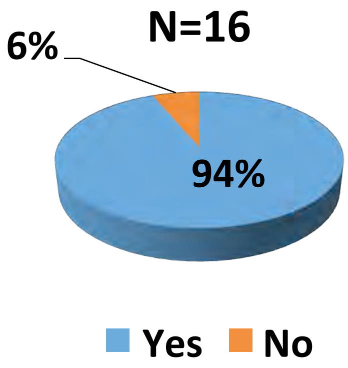
Are there downsides to patients using the Web/smartphone apps to understand their conditions?



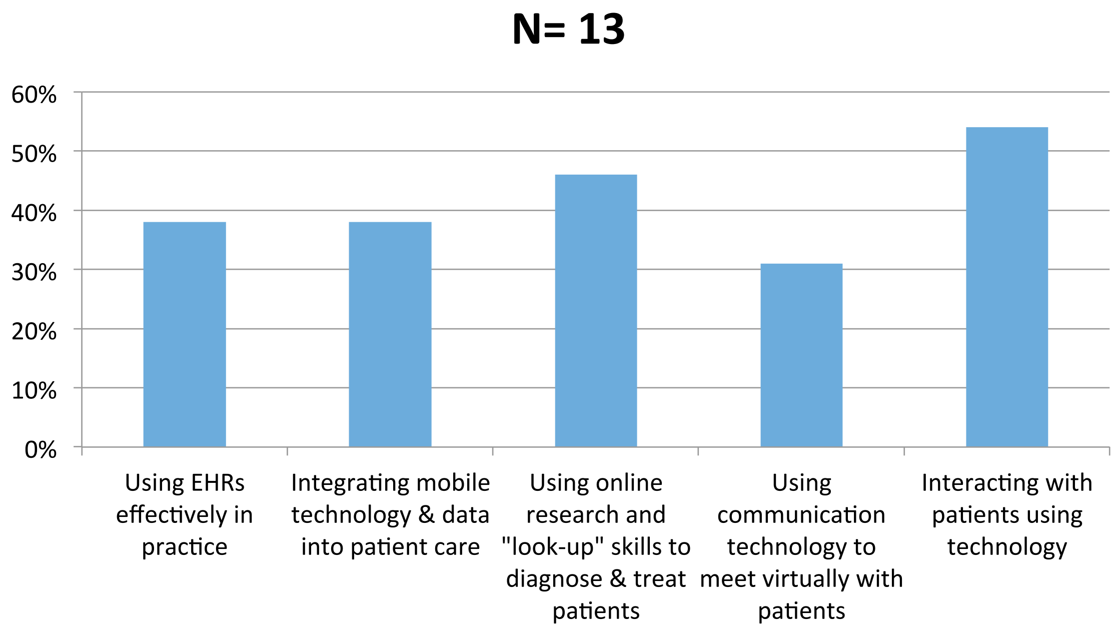
In this age of technology and information, is there still core information that physicians should know without needing to look-up through external resources?



In the age of electronic health records (EHRs) and other technology, should being able to retrieve and "look up" information efficiently and accurately be an important skill to assess for future physicians?

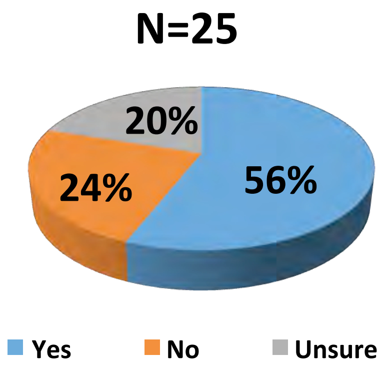


As health care technology continues to advance at a rapid pace, what knowledge should future physicians be assessed on? (check all that apply)

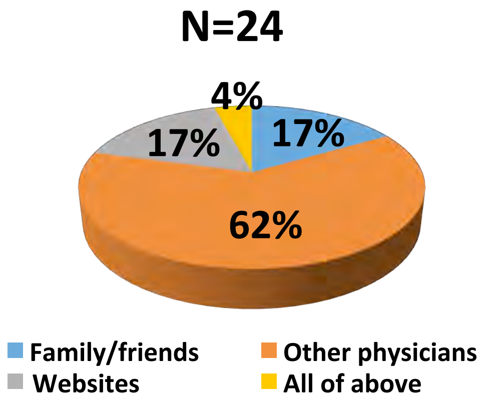


Patient Experience of Care

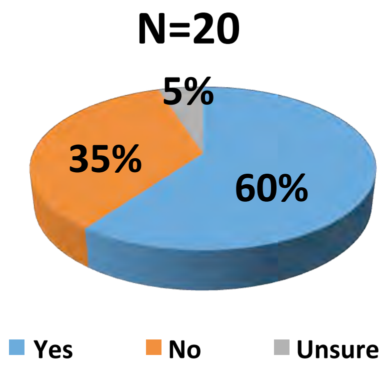
Will asking patients about their health and experience of care with a physician actually improve their health care?



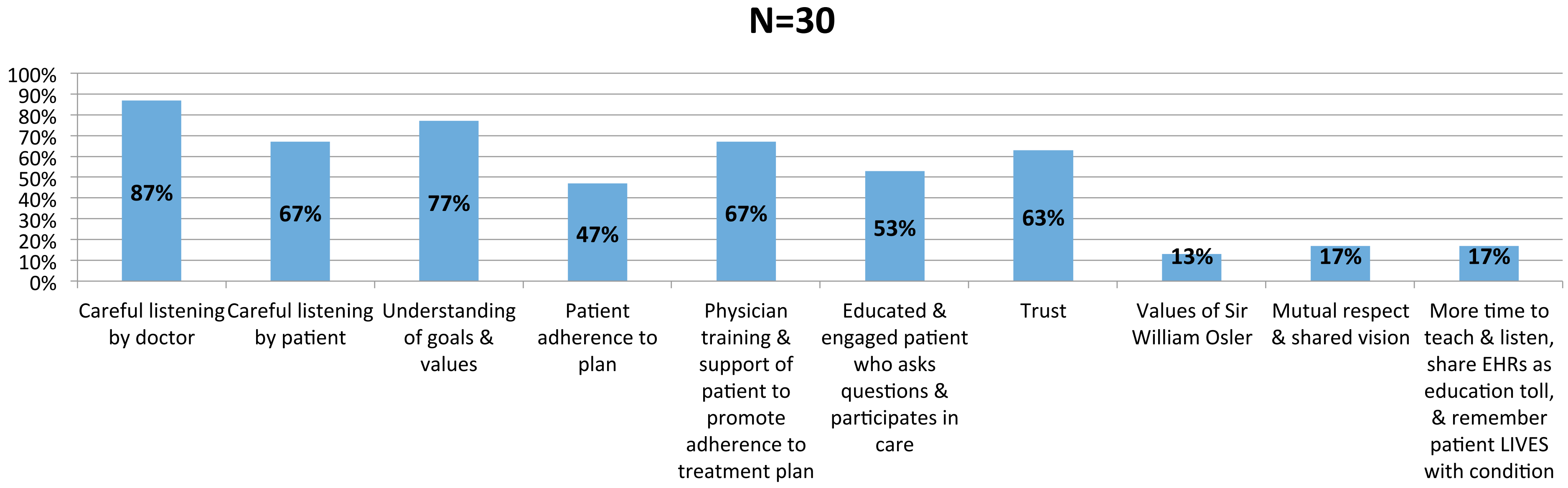
As a patient, where would you go to get important information about a physician before selecting them for your care?



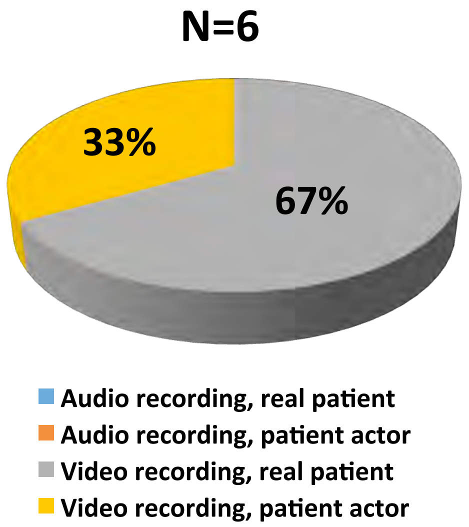
Is it possible to address patient preferences and values?



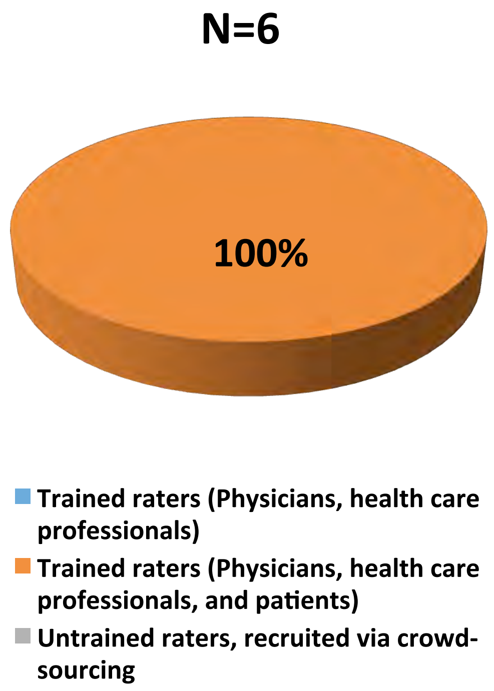
What defines a good doctor-patient partnership? (Check all that apply)



The best way to assess physicians’ skills in delivering patient-centered care would be:

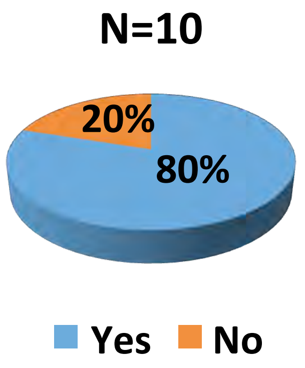


The best people to assess these skills would be:

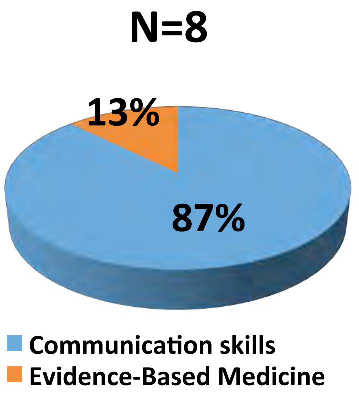


Teamwork

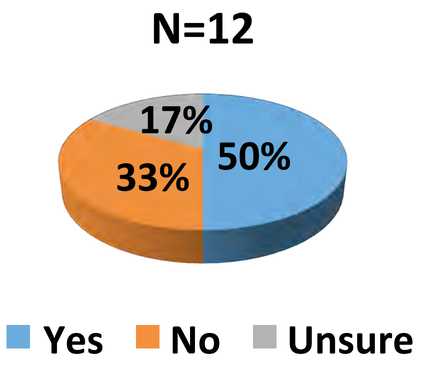
Are there skills that individual physicians need to work effectively in a multi-disciplinary team with other health care providers?



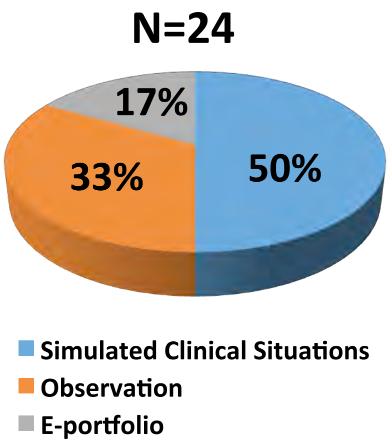
If yes, what skills?



Do you think it is possible to assess an individual physician on these skills to demonstrate competency in the area of teamwork?

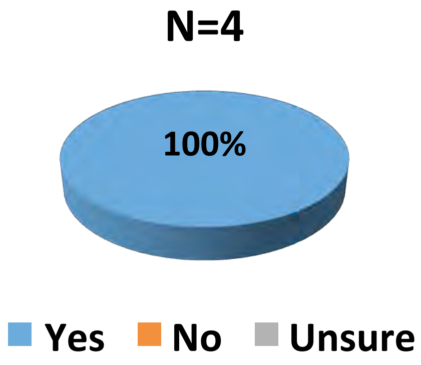


Are there downsides to patients using the Web/smartphone apps to understand their conditions?

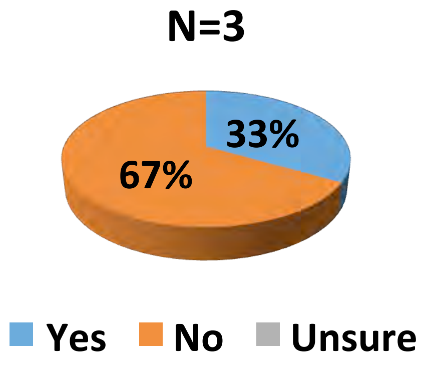


Stewardship of Resources

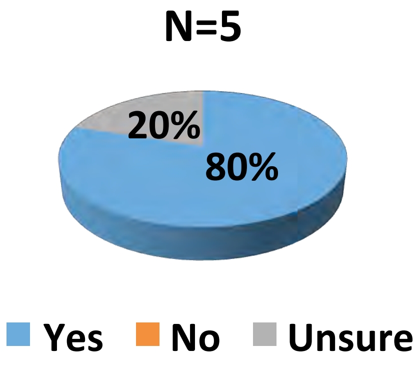
Physicians: Would you want feedback on your cost effectiveness?



Physicians: Would a case simulation on controlling costs be a helpful learning tool to improve cost effectiveness in your practice?

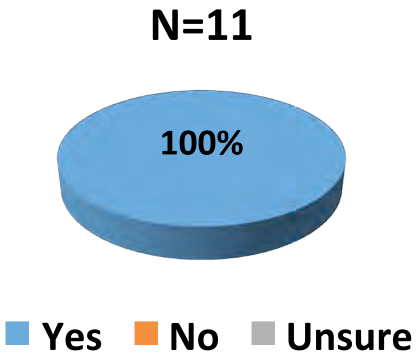


Patients: Would you want information on the cost effectiveness of your physician on clinical exam scenarios?

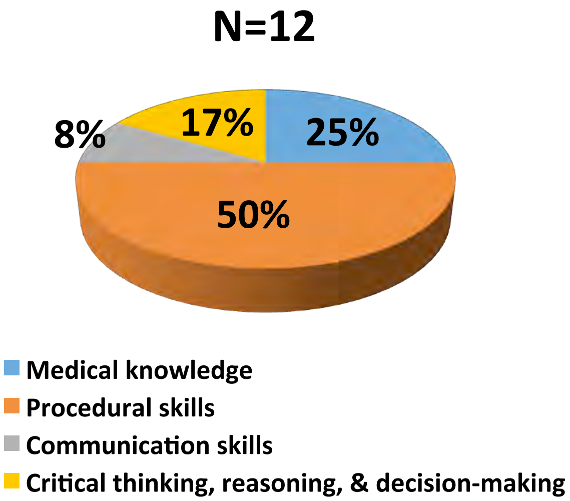


Traditional Competencies

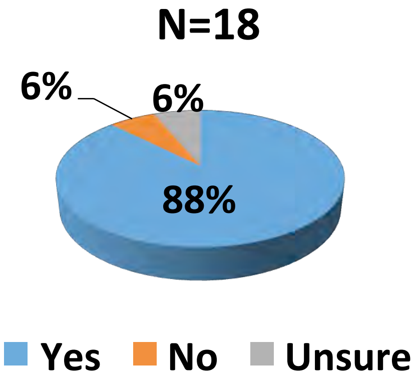
Can a physician know the right thing to do (medical knowledge), but not actually be able to perform it effectively and accurately (procedural skills)?



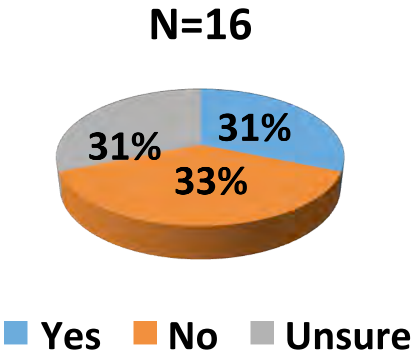
What physician skills might medical simulations be best suited to assess?



Will taking a patient's history and doing a physical exam still be important physician skills in the future?

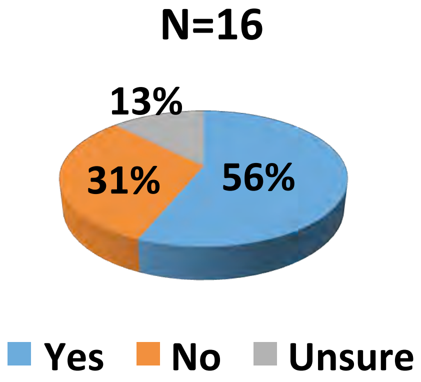


Is it possible to accurately measure a physician's history-taking and physical exam skills through assessment that is not directed observation of a real patient?

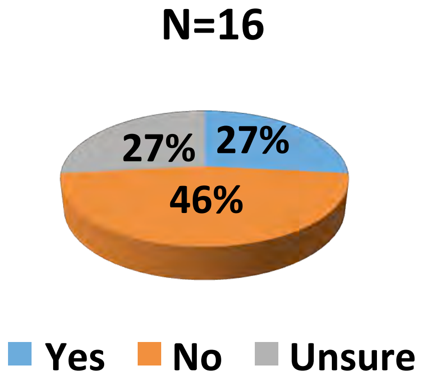


Values & Assessment Design

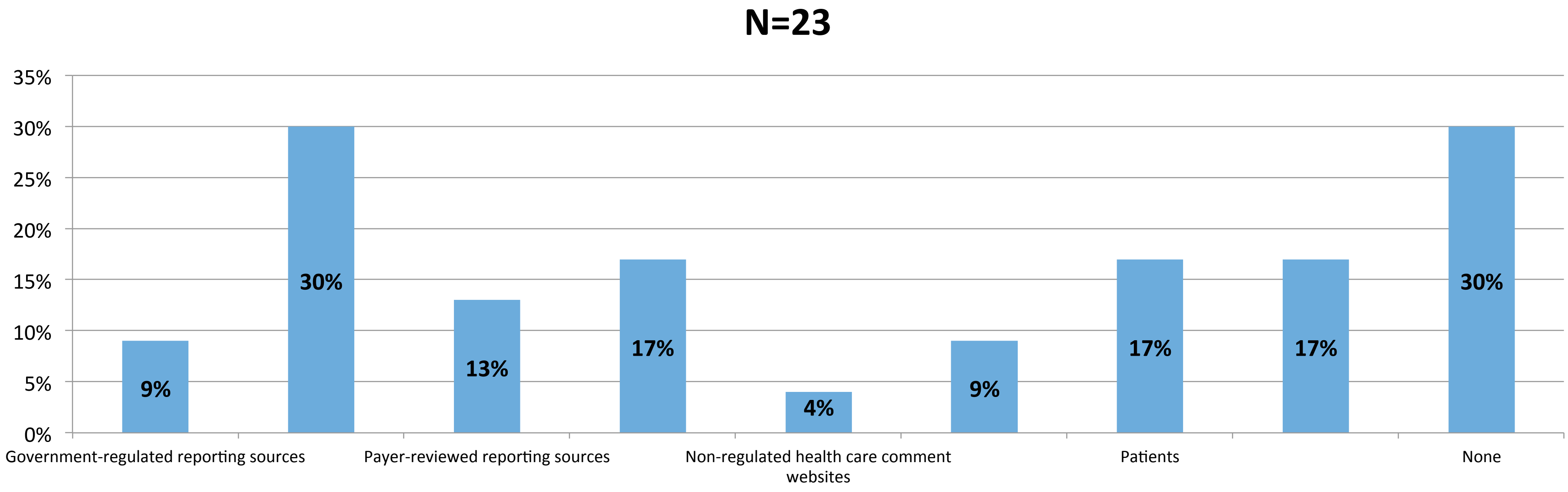
Are the benefits of adaptive testing (shorter tests for most, better precision of scores) worth the added complexity of the testing paradigm?



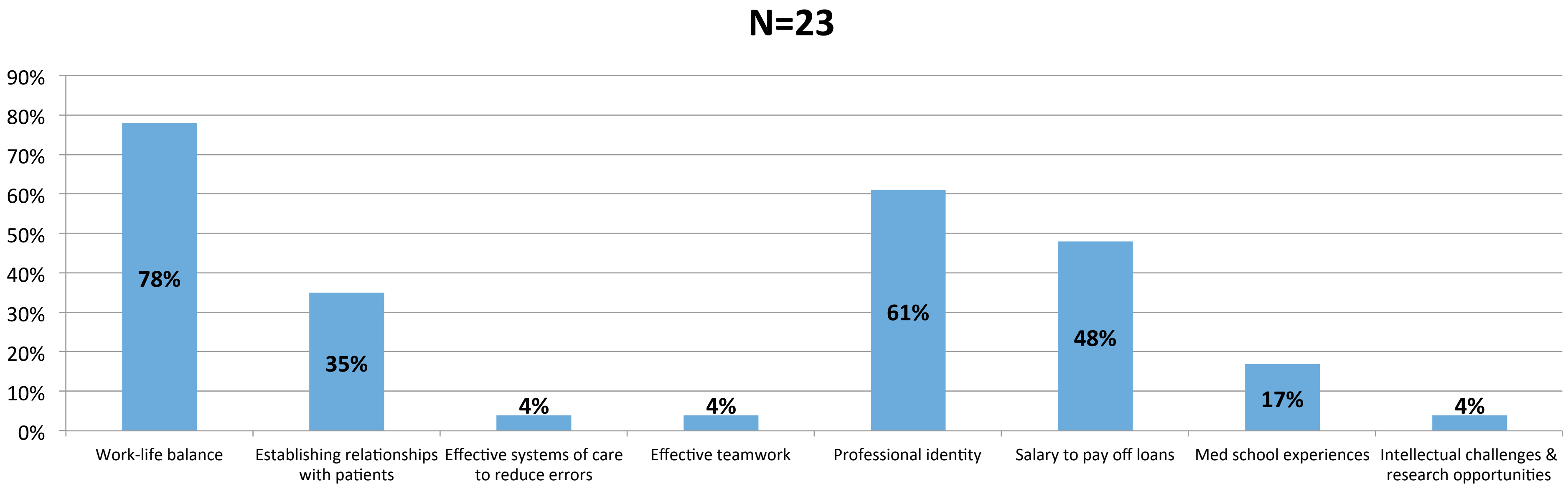
Is it fair to end the exam early for some test-takers when others may have to take a longer exam?



Who would you trust to report information about physicians to the public (check all that apply)?



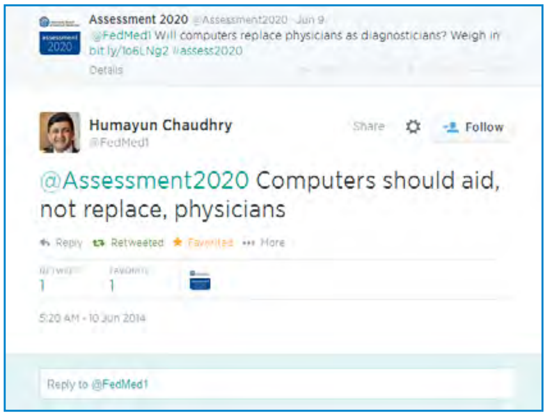
Of the following, which personal values do you believe influence younger physicians' choice in specialty (check all that apply)?



Twitter Reports

August 2014 Twitter Report for @Assessment2020

- Launch Date: May 6
- Total tweets: **251** tweets since launch with **10** tweets/wk average
- Total followers: **141** as of August 13, representing a **293% increase** since June 20
- Total impressions: **121,659** total with an estimated reach: **6,051** accounts
- **18** people joined in conversations with the account, responding to questions or providing input, averaging **9** tweets per participant



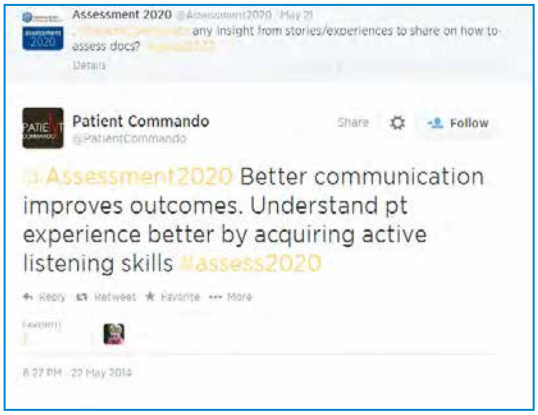
Assessment 2020 @Assessment2020 · Jun 9
@FedMed1 Will computers replace physicians as diagnosticians? Weigh in bit.ly/1o6LNg2 #assess2020
Details

Humayun Chaudhry
@FedMed1

@Assessment2020 Computers should aid, not replace, physicians

RETWEET 1 FAVORITE 1

5:20 AM - 10 Jun 2014



Assessment 2020 @Assessment2020 · May 21
[illegible] any insight from stories/experiences to share on how to assess docs? [illegible]
Details

Patient Commando
@PatientCommando

@Assessment2020 Better communication improves outcomes. Understand pt experience better by acquiring active listening skills #assess2020

FAVORITE 1

6:27 PM - 22 May 2014

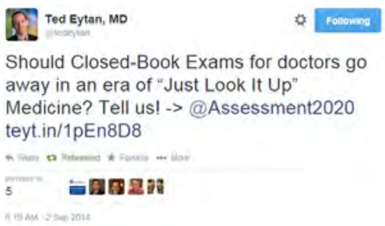
- Notable followers:
  - Institute for Patient- and Family-Centered Care **@IPFCC**
  - American Association of Colleges of Osteopathic Medicine **@AACOMmunities**
  - Official journal for Society of Hospital Medicine **@JHospMedicine**
  - American Society for Radiation Oncology **@ASTRO_org**
  - American College of Allergy, Asthma & Immunology **@ACAAI**
  - The American Gastroenterological Association **@AmerGastroAssn**
  - Steven Sternberg, Deputy Health Editor at U.S. News **@StevenSternberg**
  - Beth Toner, RWJF leadership staff **@BethTonerRN**
  - The Evidence Doc, influential epidemiologist and blogger **@TheEvidenceDoc**
  - American Educational Research Association **@AERA_EdResearch**
  - TalkAboutHealth, patient education platform **@TalkAboutHealth**

- Notable users of #assess2020 hashtag and how many times used:

@medivizor	4
@washingtonpost	3
@ameracadpeds	2
@cmsinnovates	2

September 2014 Twitter Report for @Assessment2020

- Total tweets: **319** tweets since launch with **10** tweets/wk average
- Total followers: **206** as of September 24, representing a **249% increase** since June 20
- Total impressions: **314,028** total with an estimated reach of **7,434** accounts.
    - A **258%** increase in impressions since last report
- **23** people joined in conversations with the account, responding to questions or providing input, averaging **8** tweets per participant
- Notable new followers:
    - Brown University – School of Life Sciences **@brownlifesci**
    - Alliance for Home Health Quality and Innovation **@AHHQI**
    - The Joint Commission **@TJCommission**
    - Texas Medical Association **@texmed**
    - About.com dermatology expert & journalist **@DermatologistMD**
    - National Partnership for Women & Families **@NPWF**
    - Medicine Notes, influential Twitter handle on transparency **@NotasMedicina**
    - University of Chicago Medical School **@ChicagoMedEdu**
    - Co-Founder of QuestioningMedicine Podcast **@MedQuestioning**
    - Med-Peds Hospitalist & blogger **@medpedshosp**
- Notable tweets:

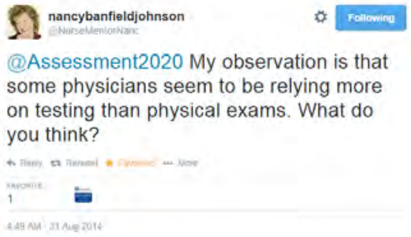


Ted Eytan, MD
@tedeytan

Should Closed-Book Exams for doctors go away in an era of "Just Look It Up" Medicine? Tell us! -> @Assessment2020 teyt.in/1pEn8D8

RETWEETS
5

8:19 AM - 2 Sep 2014

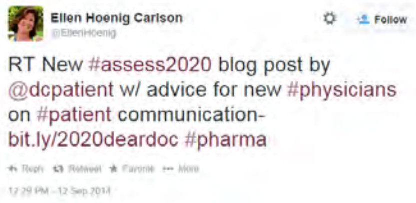


nancybanfieldjohnson
@NurseMentorNanc

@Assessment2020 My observation is that some physicians seem to be relying more on testing than physical exams. What do you think?

FAVORITE
1

4:49 AM - 21 Aug 2014

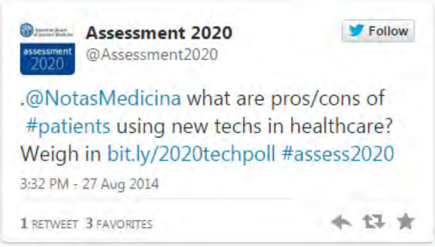


Ellen Hoenig Carlson
@EllenHoenig

RT New #assess2020 blog post by @dcpatient w/ advice for new #physicians on #patient communication- bit.ly/2020deardoc #pharma

12:29 PM - 12 Sep 2014

- Top three tweets in last 30 days (**1,089, 556** and **286** impressions):

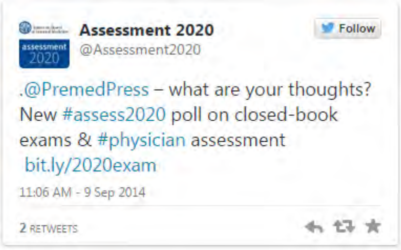


Assessment 2020
@Assessment2020

.@NotasMedicina what are pros/cons of #patients using new techs in healthcare? Weigh in bit.ly/2020techpoll #assess2020

3:32 PM - 27 Aug 2014

1 RETWEET 3 FAVORITES

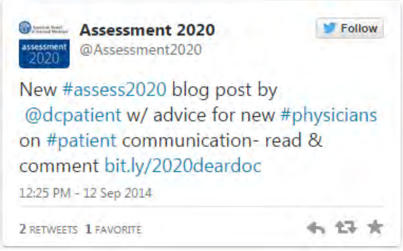


Assessment 2020
@Assessment2020

.@PremedPress – what are your thoughts? New #assess2020 poll on closed-book exams & #physician assessment bit.ly/2020exam

11:06 AM - 9 Sep 2014

2 RETWEETS



Assessment 2020
@Assessment2020

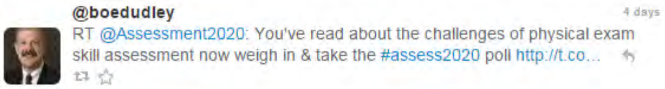
New #assess2020 blog post by @dcpatient w/ advice for new #physicians on #patient communication- read & comment bit.ly/2020deardoc

12:25 PM - 12 Sep 2014

2 RETWEETS 1 FAVORITE

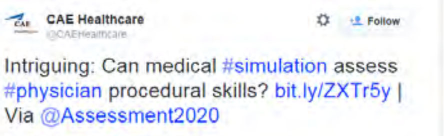
November 2014 Twitter Report for @Assessment2020

- Total tweets: **387** tweets since launch with **10** tweets/wk average
- Total followers: **248** as of November 3, a **210%** increase since June 20
- Total impressions: **344,062** total with an estimated reach of **5,900** accounts
    - An **9%** increase in impressions since last report
- **29** people joined in conversations with the account, responding to questions or providing input, averaging **9** tweets per participant
- Notable new followers:
    - Donna Cryer, Cryer LLC **@dcpatient**
    - James Legan, MD, internist with large following **@jimmie_vanagon**
    - Sandy Bauers, Philadelphia Inquirer reporter **@sbauers**
    - Dr. Kathleen Hoffman, health literacy tweet chat co-founder **@drkdhoffman**
    - Dr. Randall Oates, My HealthWare CEO **@rboates**
    - Ontario Pharmacists Association **@OntPharmacists**
    - The Center for Healthcare Engineering and Patient Safety **@UofMCHEPS**
- Notable tweets:



@boedudley 4 days

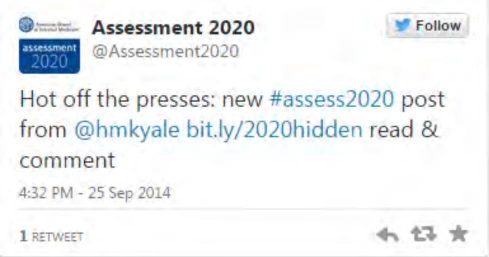
RT @Assessment2020: You've read about the challenges of physical exam skill assessment now weigh in & take the #assess2020 poll http://t.co...



CAE Healthcare @CAEHealthcare

Intriguing: Can medical #simulation assess #physician procedural skills? bit.ly/ZXTr5y | Via @Assessment2020

- Top three tweets in last 30 days (**818, 689** and **286** impressions):

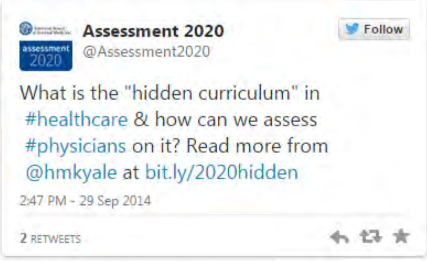


Assessment 2020 @Assessment2020

Hot off the presses: new #assess2020 post from @hmkyale bit.ly/2020hidden read & comment

4:32 PM - 25 Sep 2014

1 RETWEET

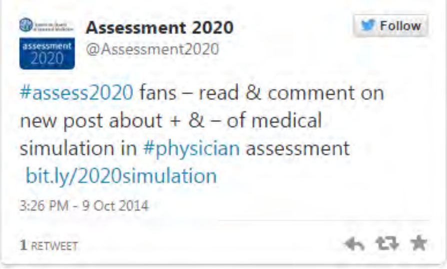


Assessment 2020 @Assessment2020

What is the "hidden curriculum" in #healthcare & how can we assess #physicians on it? Read more from @hmkyale at bit.ly/2020hidden

2:47 PM - 29 Sep 2014

2 RETWEETS



Assessment 2020 @Assessment2020

#assess2020 fans – read & comment on new post about + & – of medical simulation in #physician assessment bit.ly/2020simulation

3:26 PM - 9 Oct 2014

1 RETWEET

December 2014 Twitter Report for @Assessment2020

- Total tweets: **429** tweets since launch with **10** tweets/wk average.

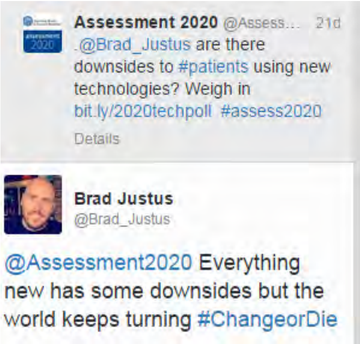
- Total followers: **259** as of December 3, a **426%** increase since June 20.

- Total impressions: **352,571** total with **3,700 impressions earned** in last 28 days.

- **29** people joined in conversations with the account, responding to questions or providing input, averaging **9** tweets per participant.

- Notable new followers:
  - American Board of Surgery **@AmBdSurg**
  - Society of Hospital Medicine **@SHMLive**
  - Dr. Carla Pugh, simulation-based assessment advocate and surgeon **@CarlaPughMDPhD**
  - Associate Chief of Emergency Medicine at Baylor College of Medicine **@bobbykapur**
  - Health Reporter for US News Health **@AKhanMedia**
  - The Health Policy Group **@healthpolicygrp**

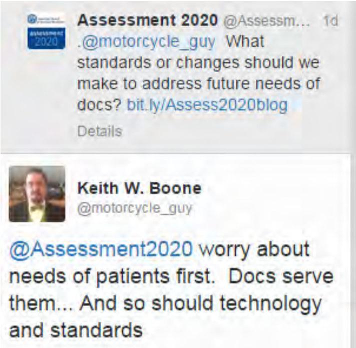
- Notable tweets:



Assessment 2020 @Assess... 21d
.@Brad_Justus are there downsides to #patients using new technologies? Weigh in bit.ly/2020techpoll #assess2020
Details

Brad Justus
@Brad_Justus

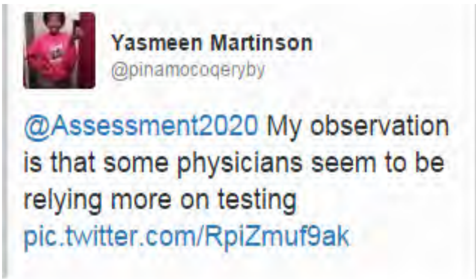
@Assessment2020 Everything new has some downsides but the world keeps turning #ChangeorDie



Assessment 2020 @Assessm... 1d
.@motorcycle_guy What standards or changes should we make to address future needs of docs? bit.ly/Assess2020blog
Details

Keith W. Boone
@motorcycle_guy

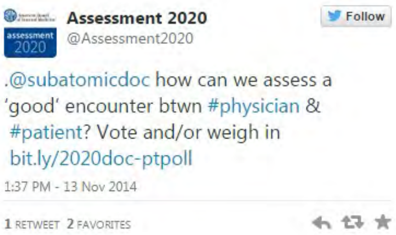
@Assessment2020 worry about needs of patients first. Docs serve them... And so should technology and standards



Yasmeen Martinson
@pinamocoqeryby

@Assessment2020 My observation is that some physicians seem to be relying more on testing pic.twitter.com/RpiZmuf9ak

- Top three tweets in last 30 days (**572, 196** and **167** impressions):

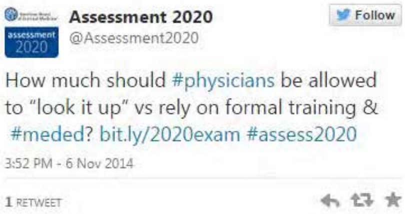


Assessment 2020 Follow
@Assessment2020

.@subatomicdoc how can we assess a 'good' encounter btwn #physician & #patient? Vote and/or weigh in bit.ly/2020doc-ptpoll

1:37 PM - 13 Nov 2014

1 RETWEET 2 FAVORITES

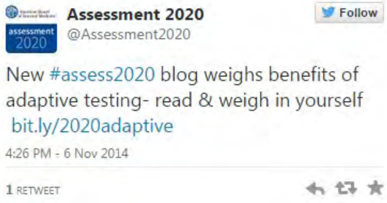


Assessment 2020 Follow
@Assessment2020

How much should #physicians be allowed to "look it up" vs rely on formal training & #meded? bit.ly/2020exam #assess2020

3:52 PM - 6 Nov 2014

1 RETWEET



Assessment 2020 Follow
@Assessment2020

New #assess2020 blog weighs benefits of adaptive testing- read & weigh in yourself bit.ly/2020adaptive

4:26 PM - 6 Nov 2014

1 RETWEET

January 2015 Twitter Report for @Assessment2020

- Total tweets: **496** tweets since launch with **10** tweets/wk average.

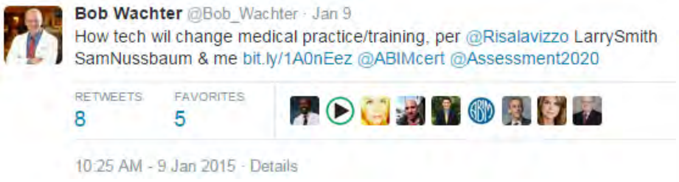
- Total followers: **300** as of January 23, a **512%** increase since June 20.

- Total impressions: **397,731** total with **5,100 impressions earned** in last 28 days, a **27%** increase since last report.

- **34** people joined in conversations with the account, responding to questions or providing input, averaging **10** tweets per participant.

- Notable new followers:
  - California Medical Association Foundation **@theCMAF**
  - Network of Ethnic Physician Organizations **@ethnicphysician**
  - National Council on Alcoholism and Drug Dependence **@NCADDNational**
  - Men's Health Network **@MensHlthNetwork**
  - Colorado Medical Society **@CoMedSoc**
  - Emergency Medicine Residents' Association **@emresidents**
  - Journal of the American Medical Association **@JAMA_Current**
  - American Board of Physical Medicine and Rehabilitation **@ABPMR**

- Notable tweets:

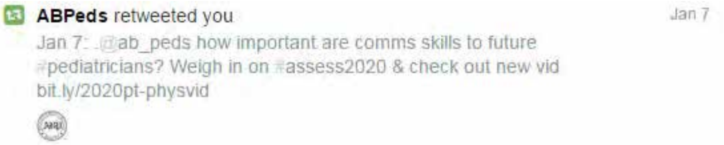


Bob Wachter @Bob_Wachter · Jan 9

How tech wil change medical practice/training, per @Risalavizzo LarrySmith SamNussbaum & me bit.ly/1A0nEez @ABIMcert @Assessment2020

RETWEETS 8 FAVORITES 5

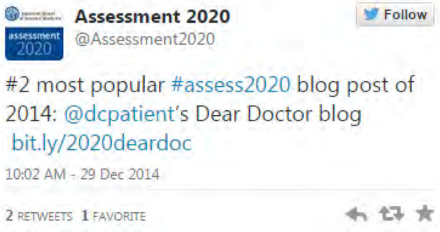
10:25 AM - 9 Jan 2015 · Details



ABPeds retweeted you — Jan 7

Jan 7: .@ab_peds how important are comms skills to future #pediatricians? Weigh in on #assess2020 & check out new vid bit.ly/2020pt-physvid

- Top three tweets in last 30 days (**363, 181** and **86** impressions):

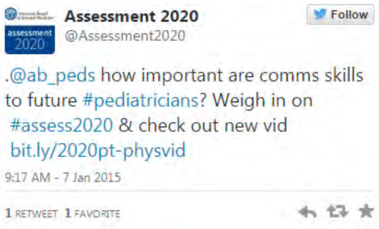


Assessment 2020 @Assessment2020

#2 most popular #assess2020 blog post of 2014: @dcpatient's Dear Doctor blog bit.ly/2020deardoc

10:02 AM - 29 Dec 2014

2 RETWEETS 1 FAVORITE

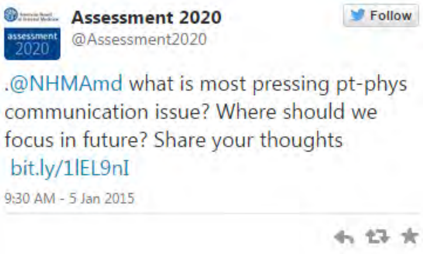


Assessment 2020 @Assessment2020

.@ab_peds how important are comms skills to future #pediatricians? Weigh in on #assess2020 & check out new vid bit.ly/2020pt-physvid

9:17 AM - 7 Jan 2015

1 RETWEET 1 FAVORITE



Assessment 2020 @Assessment2020

.@NHMAmd what is most pressing pt-phys communication issue? Where should we focus in future? Share your thoughts bit.ly/1lEL9nI

9:30 AM - 5 Jan 2015

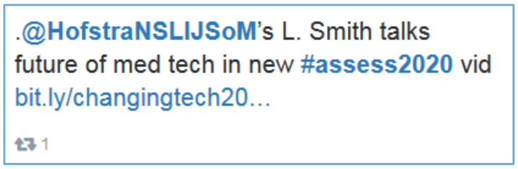
February/March 2015 Twitter Report for @Assessment2020

- Total tweets: **586** tweets since launch with **11** tweets/wk average.
- **349** followers as of April 3, a **591%** increase since June 20.
- Total impressions: **410,908** total with **13,177 impressions earned** in last two months.

February, 2015 Highlights

- Tweets 27
- Impressions 5,647
- Profile visits 144
- Mentions 10
- New followers 8

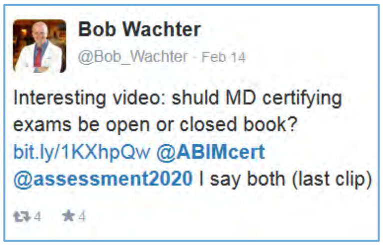
Top Tweet – earned 659 impressions



.@HofstraNSLIJSoM's L. Smith talks future of med tech in new #assess2020 vid bit.ly/changingtech20...

⟲ 1

Top mention – earned 37 impressions



Bob Wachter
@Bob_Wachter · Feb 14

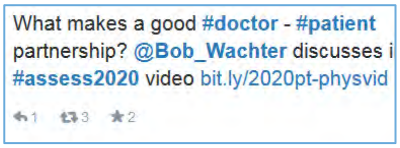
Interesting video: shuld MD certifying exams be open or closed book? bit.ly/1KXhpQw @ABIMcert @assessment2020 I say both (last clip)

⟲ 4 ★ 4

March 2015 Highlights

- Tweets 63
- Impressions 7,530
- Profile visits 171
- Mentions 9
- New followers 38

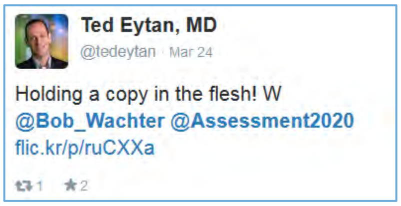
Top Tweet – earned 1,082 impressions



What makes a good #doctor - #patient partnership? @Bob_Wachter discusses i #assess2020 video bit.ly/2020pt-physvid

↩ 1 ⟲ 3 ★ 2

Top mention – earned 26 engagements



Ted Eytan, MD
@tedeytan · Mar 24

Holding a copy in the flesh! W @Bob_Wachter @Assessment2020 flic.kr/p/ruCXXa

⟲ 1 ★ 2

Group Outreach Summaries

We interviewed leaders of national public interest and consumer organizations to gather their perspectives on the general question, “What do you see as the essential features of a state-of-the-art program of physician assessment in 2020?” Half of the interviewees who represent these public interest and consumer organizations were also ABIM diplomates with a fairly sophisticated understanding of the current Maintenance of Certification (MOC) program. The other interviewees have all had at least some professional exposure to MOC and understand its purpose.

These individuals represent the following public interest and consumer groups:

- Informed Medical Decisions Foundation
- WomenHeart: The National Coalition for Women with Heart Disease
- National Patient Safety Foundation
- Consumers’ CHECKBOOK
- Office of the National Coordinator for Health Information Technology
- Patients Like Me
- Consumers Union

We explained that ABIM is actively reviewing its approach to physician assessment with an eye toward future expectations of the credential, and seeking input and feedback from a wide variety of stakeholders who represent the “consumers” of that credential – in this case, consumers and patients. We invited interviewees to consider the most important attributes of both the knowledge and practice assessment, and to tell us what areas they believe their constituencies would hope or expect to see as part of an MOC program within the next five years (2020). We also invited knowledgeable opinions about assessment techniques.

All of the individuals we spoke with appreciated the complexity of designing and applying valid, meaningful assessments to the expanding knowledge and skills that payers, the public, policymakers and employers now expect of physicians. Nevertheless, these leaders recognized that what it means to be a “competent” physician to the public is constantly changing and, as a result, ABIM will have to continuously modify its program to keep with these evolving expectations to deliver on its value proposition to the public.

Two consumer leaders, one of whom was a physician, acknowledged that ABIM should continue to aim for rigor and transparency, even in the face of daunting uncertainty, stating:

> “ABIM should think and talk about this as part and parcel of the wider evolution in accountability in health care, which is all for the good, but terribly uncertain and uncomfortable. All science moves along the continuum from ignorance to confusion to knowledge. We are in the confusion phase. But that’s better than

> being in the ignorant phase, and sooner or later we'll get to the knowledge phase. 'So hey, doc, it ain't perfect, but c'mon, you gotta jump in.'"

Values*:*

We did not ask interviewees specifically to identify the values that should guide our program, but many of them volunteered opinions that were closely aligned with values for a good assessment.

The consumer organizations (e.g., Consumers Union, Consumers' CHECKBOOK) were both emphatic that ABIM needs to continue to strive for a more rigorous and scientific approach to evaluating physicians. A rigorous assessment paired with meaningful, reliable standards are both essential to professional self-regulation. Each year, Consumers' CHECKBOOK explains the purpose of MOC to its readers and urges them to seek physicians who participate in MOC on the faith that the credential is issued following a rigorous and reliable assessment of a physician's competency and skills.

Interviewees distinguished between robust activities and assessments that genuinely challenge and engage physicians and those that are overly simplified busy work activities. They explained that physicians want both their intelligence and their time to be respected and valued, and that providing meaningful, engaging assessments is important to avoid cynicism.

Consumer organizations have long expressed frustration with the lack of publicly available, comparative information on physician quality, and trustworthiness and value of the credential to the public is important. These leaders expressed that ABIM and other specialty boards would "step up and tell us what you know," arguing that the alternative future would be judgment by HealthGrades or other publicly available ratings sources. Leaders also suggested that specialty boards are the ideal entities to help consumers make sense of and create filters for all of this information. In this vein, MOC should emphasize transparency on the knowledge and skills of physicians that are most important to the public, including clinical quality, communication, and cost.

Competencies*:*

There was wide agreement that MOC assessment should, at a minimum, include the Triple Aim priorities:

- Patient experience
- Clinical performance
- Cost

Several interviewees also suggested a "Quadruple Aim" that addresses all of the above priorities and adds patient engagement. These leaders expressed that physicians can and should learn better communication and shared decision-making skills, but that the essence of patient-centered care is something more profound and more difficult for most physicians.

These skills include learning to trust and value information generated by the patient and an evaluation of the outcome in terms of the patient's goals and expectations for their care.

Other widely shared priorities include teamwork/collaborative skills and the ability to access, metabolize, and act on population health data for improvement.

Most emphasized the need to continuously improve assessment of physician's diagnostic reasoning process and saw this as one of the most and unique and important contributions of the specialty boards. Enhancing and maintaining diagnostic reasoning capabilities was mentioned as a critical priority in conversations about consumer trust, patient safety, patient costs, system costs, and women's health.

Professionalism was another focus for the consumer representatives: "How do we create assessment on the art of medicine—compassion, respect, professionalism—not just the science? How do other disciplines (e.g., business) test this?"

Proficiency with health information technology was another area spotlighted as a necessity by several interviewees and, not surprisingly as a priority by the Office of the National Coordinator for Health Information Technology (ONC). ONC specifically noted the need to rapidly (formatively) develop physicians' skills in the use of population health data for quality improvement and longitudinal data for clinical decision support. There is recognition that this can be difficult due to the generational shift in technology, and that assessment strategies may need to develop recognition that the skill levels of physicians of different generations would be uneven. As a matter of patient safety in these areas, physicians should also be able to demonstrate proper safeguarding of patient privacy, safe e-prescribing practices, and the potential risks of patient misidentification in the electronic health record.

<u>Assessment Design</u>:

All of our interviewees want to see ABIM continuously evolve and tailor assessment design to capture and account for the desired attributes of a 21st-century physician. Leaders of consumer-facing organizations focused more on what the assessment represents to the public, and seemed uncertain of the value of formative assessment activities like quality improvement, when the only requirement for the credential is physician participation. These leaders explained that physicians should have to demonstrate that they have, in fact, improved some aspect of care if the standard is "participation in quality improvement." However, they do recognize that all physicians differ in the skills needed to effectively interpret and act on quality data, and that meaningful participation would still be of value, though not completely satisfactory.

Many of the represented organizations either have developed or are eager to co-develop tools to promote and measure physician performance on the competencies that are especially important to them. All expressed interest in continuing to work with ABIM on the development of content and methods to support physician assessment or self-assessment strategies.

Some specific suggestions related to assessment design include the following:

- Move the cognitive exam away from the written test and towards a two–part assessment of cognitive skills: 1) a series of clinical dilemmas/questions, with “policed” access to some approved set of real-world practice resources/technology; and 2) a shorter test portion without electronics that forces physicians to think on their feet and solve a clinical problem (“There will always be times when either technology support is not available and/or when used blindly, it can mislead a doctor’s thinking—i.e., the doctor’s hard drive/cognitive process is still the critical thing we should care about”).

- Test physician “self-awareness”: What/how can they know about their own performance in relation to other physicians?

- Use measurement to see if doctors are getting anything out of their improvement activities: Move away from the current practice of “active-doing” and use a methodology involving the measurement of improvements where a physician must show that they have improved what they are doing in some way.

- Patient engagement: One interviewee noted that there is a debate about the assessment of shared decision making in whether or not a physician is using tools for specific clinical scenarios (e.g., specialty-specific decision aids) or is actually actively focusing on true engagement of the patient in the process. Both are important, but the latter captures the true essence of shared decision making and is a difficult skill for most physicians. Assessment of patient engagement should also include measuring a patient’s overall experience of living with their condition each day, not just a point in time during the clinical encounter.

- Patient experience: Working with health plan directories, one of the consumer-focused groups fielded a survey based on the CAHPS framework in four metropolitan areas and found statistically significant differences in performance scores. This is the first survey of its kind on the individual physician level, and the organization strongly believes that it is methodology that can and should be adopted by all boards and other major entities.

Appendix C:

Diplomate Exam & Product Feedback

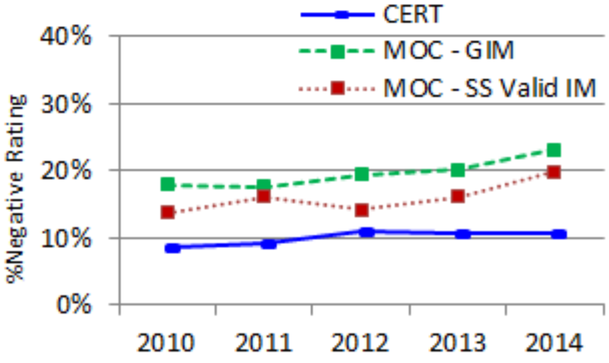
Diplomate Survey Results

Internal Medicine Exam Survey:

These results were determined based on surveys submitted by general internists and specialists with a valid Internal Medicine (IM) certificate after completion of the Internal Medicine certification or Maintenance of Certification exam in from 2010 through spring 2014. A total number of 67,081 surveys were collected from these individuals following the secure exam.

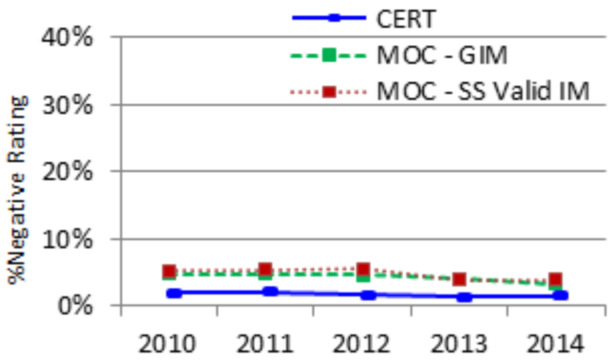
The examination was a fair assessment of clinical knowledge in this discipline

	Overall Rating		
Exam	%Negative	%Neutral	%Positive
CERT	10%	24%	66%
MOC – GIM	19%	30%	50%
MOC – SS Valid IM	15%	29%	56%



The extent of the security procedures at the test site was appropriate

	Overall Rating		
Exam	%Negative	%Neutral	%Positive
CERT	2%	3%	95%
MOC – GIM	4%	6%	90%
MOC – SS Valid IM	5%	5%	90%

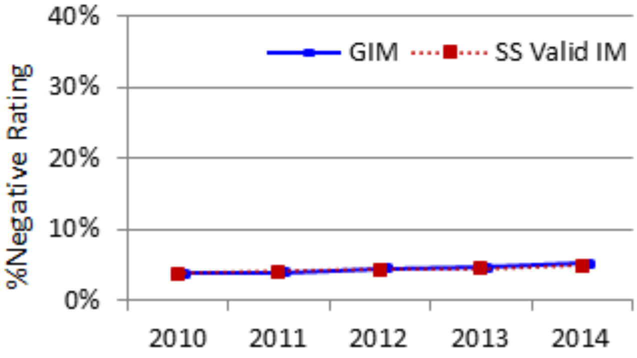


Medical Knowledge Self-Assessment Modules Survey:

Surveys were submitted by general internists and specialists with a valid IM certificate following participation in one of ABIM's Medical Knowledge Self-Assessment modules between January 2010 and September 2014. A total of 197,094 surveys were collected with 100,705 from general internists and 96,389 surveys from specialists with a valid IM certificate.

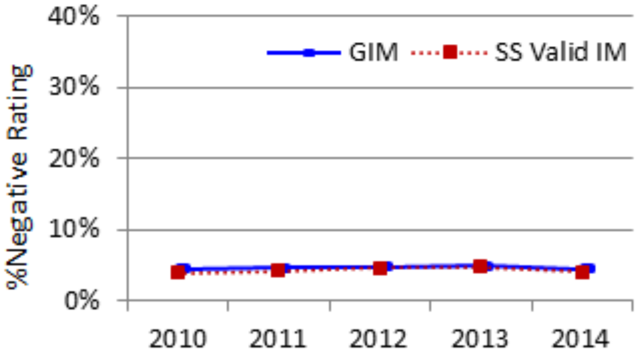
This module provided a valuable overall learning experience

	Overall Rating		
	%Negative	%Neutral	%Positive
GIM	4%	10%	85%
SS Valid IM	4%	10%	86%



I was able to complete the module without technical difficulty

	Overall Rating		
	%Negative	%Neutral	%Positive
GIM	5%	9%	86%
SS Valid IM	4%	8%	88%

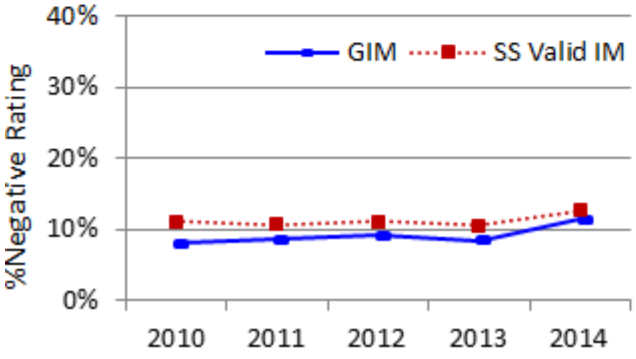


Practice Improvement Modules Surveys:

Surveys were submitted by general internists and specialists with a valid IM certificate following participation in an ABIM PIM Practice Improvement Module® between January 2010 and September 2014. A total of 26,116 surveys were obtained with 13,600 from general internists and 12,516 from specialists with a valid IM certificate.

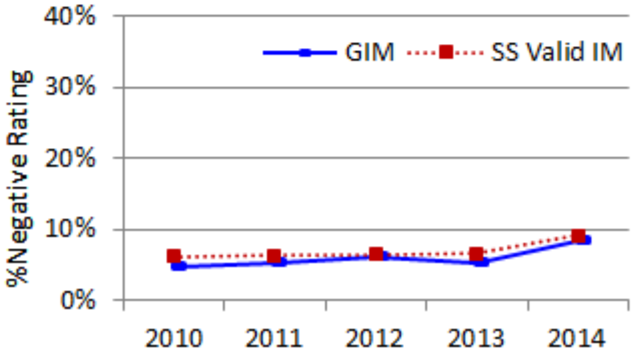
This module provided a valuable overall learning experience

	Overall Rating		
	%Negative	%Neutral	%Positive
GIM	9%	17%	74%
SS Valid IM	11%	20%	69%



Participation in this module enhanced my ability to assess current practice performance

	Overall Rating		
	%Negative	%Neutral	%Positive
GIM	6%	11%	83%
SS Valid IM	7%	13%	80%

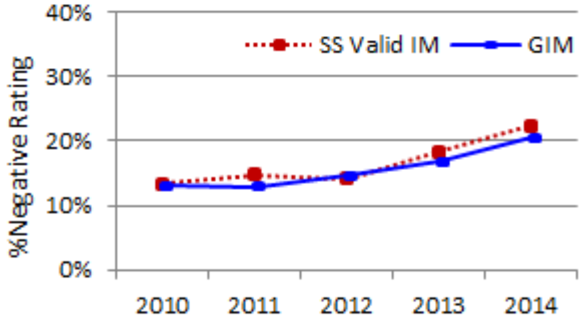


Overall Maintenance of Certification (MOC) Program Surveys:

Surveys on overall MOC program satisfaction were completed by general internists and specialists with a valid IM certificate who had completed the program between January 2010 and May 2014. A total of 13,338 surveys were received with 4,877 from general internists and 8,461 from specialists with a valid IM certificate.

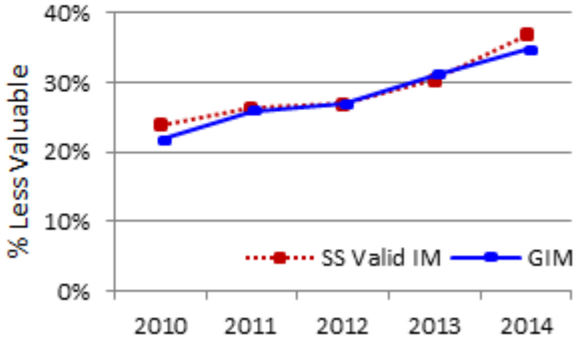
The MOC program was a valuable learning experience

	Overall Rating		
	%Negative	%Neutral	%Positive
GIM	15%	21%	63%
SS Valid IM	17%	21%	63%



How would you rate the value of particular MOC activities vs. other activities outside ABIM?

	Overall Rating		
	%Less	%Equally	%More
GIM	28%	57%	15%
SS Valid IM	29%	58%	13%



Diplomate Comment Summary

- Exams are too specific and question proportions do not follow the blueprint on the ABIM website.

- Questions on the exam are too focused at the subspecialty level and present conditions that are not common. Many diplomates view these questions as unfair.

- The exam does not align with what physicians do in practice in the content presented, the design of the exam, and the lack of available resources that physicians would use in practice every day.

- Diplomates express a desire for both the exam and other parts of the Maintenance of Certification (MOC) program to be relevant to the work they do in daily practice, including quality improvement.

- Diplomates expressed that they felt that MOC products were just busy work, laborious, and time-consuming, and were very difficult for a busy practicing physician to complete, as some of the products took months to finish. Some diplomates who had positive comments about the product itself also expressed frustration with the amount of time that it took to complete.

- Diplomates also felt that some of the products offered in the current MOC program were not a worthwhile use of their time and did not contribute any value to their practice in the end.

- Diplomates expressed a desire for opportunities to participate in activities that will be useful and engaging, and should be efficient to complete, without too much burden on a physician's time, as the current interface for the ABIM PIM Practice Improvement Modules® was described to be too slow with delays in entering certain fields and technical glitches that were difficult to manage.

- Diplomates find that current activities involve too much duplicative data collection and data entry, and finding ways to transfer data from existing sources, like electronic health records, would improve the process significantly.

- Diplomates value parts of the program that allowed them to recognize weaknesses in their practice and make changes that would improve their care for patients and patient outcomes.

- Program surveys of diplomates have increasingly shown that diplomates feel that there are too many questions on the secure, closed-book exam that focus on infrequently seen, but important, illnesses and uncommon patient populations. They would frequently consult external resources in practice to treat these patients, but that information is not currently available to them on the exam.

- Many diplomates expressed that single-best answer items might appear to have a single correct response, but that in practice, physicians often access additional resources that would help them make a good judgment call, or choosing the correct response was dependent on other things they knew about the patient.

- Some diplomates expressed issues with the testing environment itself, including: testing experience was uncomfortable and anxiety-producing; the software was clunky and computer malfunctions sometimes hurt the test-taking experience; and the security procedures were unnecessary and excessive.

- Diplomates expressed dissatisfaction with the quality and clarity of both audio components and images on the exam. Audio components were difficult to hear and too short in length, making answering audio questions difficult. Images were also poor quality, grainy, and out-of-date and the lack of zoom feature was a significant barrier in answering image questions.

Appendix D:

Clinical Diagnostic Reasoning

Assessment 2020: Clinical Reasoning Workgroup

May 20, 2014

Charge*: To provide a two-page concept paper describing a future approach for assessing clinical reasoning with the aspiration of improving this physician skill.*

This document provides an executive summary regarding the main ideas generated from three conference calls. The references that we used are listed in ***Appendix 1***.

Per the charge, this document will outline definitions, models, and suggested assessment approaches for this construct. Our group narrowed the construct to ***diagnostic* reasoning** (for the reasons listed below). All workgroup members endorsed the importance of assessing *clinical reasoning* (both diagnostic and therapeutic) as it is central to what an internist does in practice, and good clinical reasoning has been cited as being likely to reduce errors and improve patient morbidity and mortality, efficiency, and cost of care.

Definitions:

Clinical reasoning is a multi-step, often iterative, process by which medical professionals make decisions about patient care. This includes the steps up to and including making decisions about diagnosis and treatment. Clinical reasoning depends on the specifics of the situation, including attributes and preferences of the patient, other health care team members, the system where care is provided, and the physician him or herself. Some scenarios call for quick recognition and action while others require careful deliberation. Effective clinical reasoning requires thinking both quickly (e.g., pattern recognition) and slowly (e.g., deliberately) and, most importantly, maintaining an appropriate balance between these two reasoning processes. Good clinical reasoning is believed to result in efficient, effective, safe, and cost-conscious care. Diagnosis and therapy can be thought of as two parts of the overall clinical reasoning process. We agreed upon a social-cognitive model for portraying clinical reasoning (Durning, 2010) as a whole. It emphasizes the participants in the encounter (e.g., patient and physician), the setting or environment, and their interactions. In doing so, the model is consistent with our current understanding of clinical reasoning portrayed in the above definition. ***Appendix 2*** displays this inclusive model in more detail. Physician factors include their knowledge and prior experience, patient factors include acuity of illness, and system factors are things such as time for the appointment and the presence of an EHR. For purposes of this document and our current recommendations, we will focus on physician factors (e.g., a cognitive vs. socio-cognitive approach) as they are most readily assessed through ABIM assessment venues.

Diagnostic reasoning focuses on the steps up to and including establishing the diagnosis and stops short of treatment. *The group decided to focus our task on diagnostic reasoning for several reasons*:

- Narrowing the scope allows us to build a framework for assessment that has fewer moving parts. The field's understanding of diagnostic reasoning is currently far more robust than that of therapeutic reasoning.

- Appropriate therapeutic reasoning is dependent, at least at times and in part, on successful diagnostic reasoning.

- The "patient experience of care" workgroup is charged with examining the doctor-patient dyad in decision-making around therapeutic treatment.

Model of diagnostic reasoning:

We used a modification of Bowen's model (2006) of *Key Elements of the Clinical Diagnostic Reasoning Process* for honing our focus on diagnostic reasoning. The model is made up of several components that may be applied iteratively depending on knowledge, context, experience, and other factors. Research suggests that several of these components may be actively engaged simultaneously and that they are not sequential, each influencing the other. So, while drawn in sequence, they should not be thought of as steps in a process necessarily. The literature has also found that a high degree of mental flexibility and adaptability is also required for successful clinical reasoning.

1. *Patient's story.* The patient's story is, basically, how the patient initially presents to the doctor. Whatever information is gathered prior to the interaction with the doctor is part of this story.

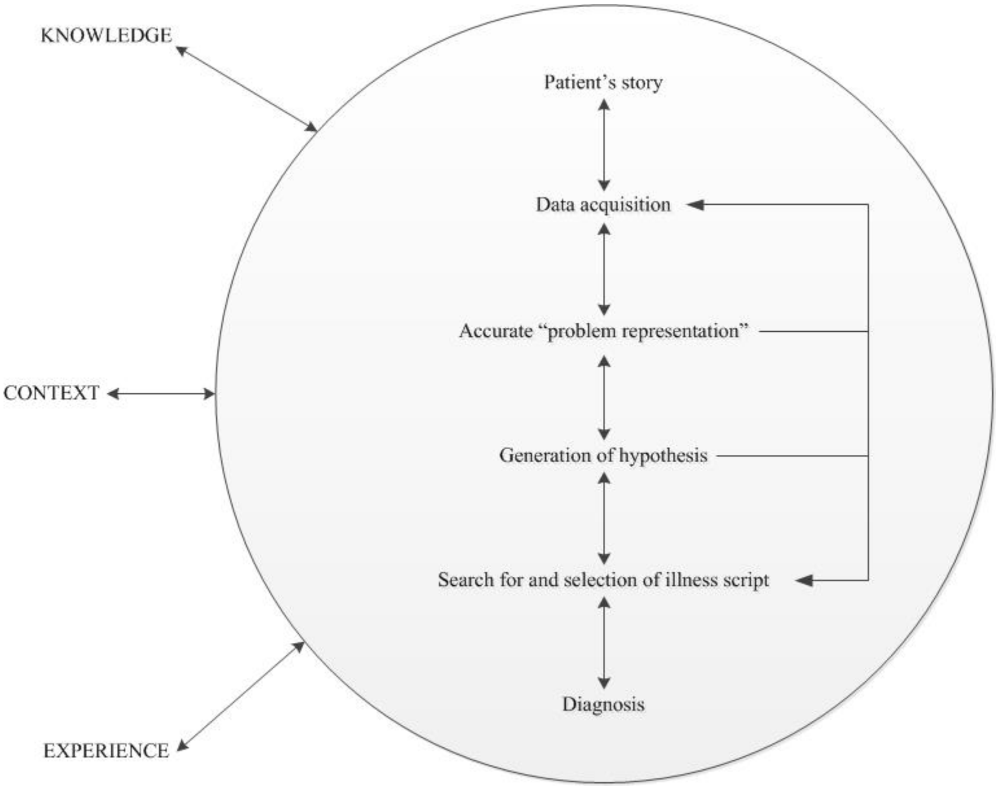
2. *Data acquisition*. Data acquisition consists of gathering additional information—history, physical exam, laboratory or radiographic tests.

3. *Problem representation*. Problem representation is the assembly of the pertinent features in the data that define a patient's condition.

4. *Hypothesis generation*. Problem representation suggests certain illnesses and may constrain the illnesses considered just as readily as the hypotheses one holds might alter data acquisition, problem representation, or illness script comparison and selection.

5. *Illness script selection*. Illness script selection involves taking knowledge of how illnesses present, comparing them to the features in the present patient, and looking for a match. This matching is a combination of analytic (e.g., deliberate application of decision rules or search for features) and non-analytic (e.g., pattern recognition) processes. It ends with selecting what is perceived to be the most likely diagnosis.

A revised version of Bowen's model is shown below.



Suggested approaches to assessing diagnostic reasoning:

We reviewed the current literature on assessment approaches, commonly used tools, measurement evidence, as well as cost and benefit. A table depicting this work is found in ***Appendix 3.*** The group endorsed the following principles: a) assess more than the single best answer, b) evaluate different components of the process, c) use current evidence and theory, and d) acknowledge diplomates' desire for authenticity in the assessment strategies. Assessing diagnostic reasoning will require a portfolio of feasible and valid assessment tools since one assessment is unlikely to capture what we are hoping to model and measure.

Since the goal of measuring diagnostic reasoning is focused on assessing the process of arriving at the correct diagnosis, the group recommends the following course of action for ABIM.

ABIM should conduct research on several new options of assessment such as virtual patients, and use the key features approach to determine whether they can measure the process of diagnostic reasoning and whether they add value above and beyond what is being measured through multiple-choice questions. Concurrently, ABIM should enhance its current multiple choice questions. Our recommendations include:

1. Given the robust psychometric evidence of multiple-choice question (MCQ) examinations and their feasibility, consider enhancing the current MCQ approach, through:

a. Fine-tuning options in MCQs by considering common misconceptions/undesirable actions.
b. Using sequences and combinations of items in selected response formats, including MCQs, that create compact mini-performance tasks that build on one another.
c. Using short constructed response tasks (i.e., short answers) that can be scored for content relatively easily through modern natural language processing (NLP) techniques.
d. Using the principles of key features approach (i.e., focusing on the most critical aspects of a patient encounter for any given case under consideration).

2. Develop and validate virtual patients (VP). This approach allows for assessing the different steps in the process in our model in a more authentic way than MCQs and offers the potential for robust individualized feedback and a partial credit scoring model. We acknowledged the high resource requirements to develop this modality, but consensus was reached that this assessment should be considered for the future. With this, we might be able to make the secure exam a combination of MCQs and VPs.

3. Develop and validate the key features exam approach, which can assess more than a single best answer and the steps in the process through a variety of examination formats. Key feature exams have been used in the Canadian assessment systems and appear to be more authentic than MCQs. With this, we might be able to make the secure exam a combination of MCQs and key features.

4. Explore some of the other assessment tools as additional possibilities to develop a portfolio of assessments that add value to certification and maintenance of certification in the area of diagnostic reasoning.

Appendix 1: References used in the work of the Clinical Reasoning Workgroup

1. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. *The New England Journal of Medicine.* 2006(21):2217.

2. Durning SJ, Artino AR, Holmboe E, Beckman TJ, van der Vleuten C, Schuwirth L. Aging and cognitive performance: Challenges and implications for physicians practicing in the 21st century. *Journal of Continuing Education in the Health Professions.* 2010;30(3):153-160.

3. Eva KW. What every teacher needs to know about clinical reasoning. *Medical Education.* 2005;39(1):98-106.

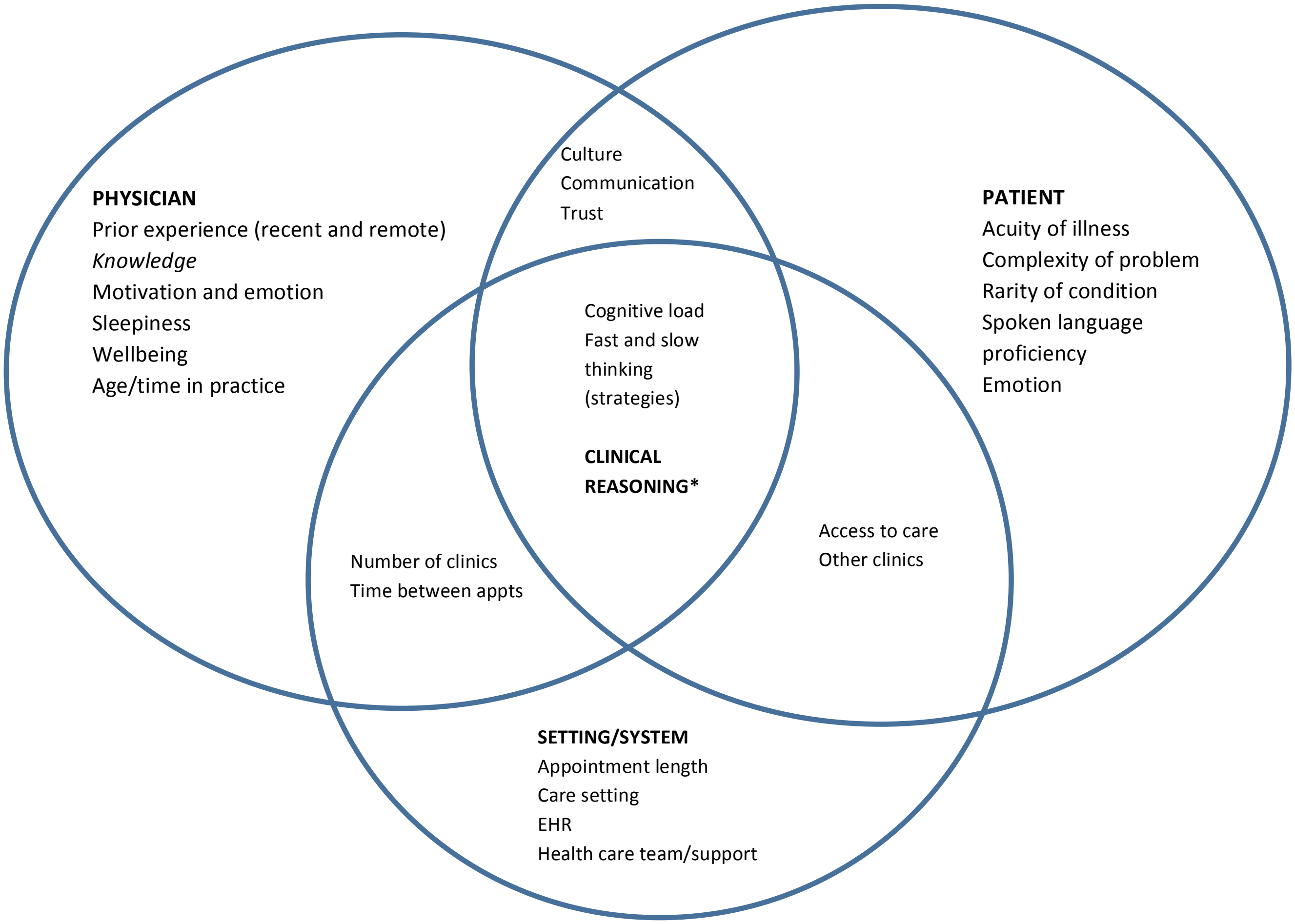
4. Ilgen JS, Humbert AJ, Kuhn G, et al. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. *Academic Emergency Medicine: Official Journal Of The Society For Academic Emergency Medicine.* 2012;19(12):1454-1461.

5. van Bruggen L, Manrique-van Woudenbergh M, Spierenburg E, Vos J. Preferred question types for computer-based assessment of clinical reasoning: a literature study. *Perspectives On Medical Education.* 2012;1(4):162-171.

6. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Medical Education.* 2005;39(3):309-317.

Appendix 2: An inclusive model for the clinical reasoning process

This model includes both diagnostic and therapeutic reasoning. Circles represent the physician, the patient, and the setting. Where the circles overlap represent interactions between these three entities. Not all the listed factors are relevant for each encounter and the model is also not inclusive of all possible factors. We believe that this model could inform other small group taskforce discussions.



* Clinical reasoning emerges from the above elements and interactions and is believed to follow the steps proposed by Bowen's model (for diagnostic reasoning). As we will focus on diagnostic reasoning, we will use the Bowen model, with modifications, as informed by our above model for the task of assessing diagnostic reasoning.

Appendix 3 – Attributes for Clinical Reasoning Assessments

	Tool	Short Description	Reliability and Validity Evidence (Very Good/ Good/Fair/Poor)	Pros	Cons	Example	Comments About Steps in Model
1	**Standard ABIM MCQ**	Standard clinical vignette followed by typically 5 options.	Very good reliability evidence. Good validity evidence (reliability of .8 in two hours).	Reliability and validity evidence. Broad content can be sampled.	Low authenticity (though vignettes constructed to reflect clinical situations), cueing (A-E options). Intermediates (those right out of residency) do better than individuals in practice. **Research**: assess obtaining and representing problem (e.g., intermediate steps, per model).	A 67-year-old man with hypertension, diabetes, and high cholesterol comes to the physician with chest pain that radiates to his left arm. Vitals are blood pressure 160/80 mm Hg, pulse 98/minute, respirations 18/min and temperature 99F. The patient is diaphoretic. Examination reveals a systolic murmur. ECG demonstrates 2mm ST segment elevations in leads II, III, and AVF. Which of the following is the most likely diagnosis? a) myocardial infarction; b) pulmonary embolus; c) aortic dissection; d) gastroesophageal reflux disease; e) pericarditis.	***MCQs in their various formats could be used to assess intermediate steps in our model*. Example (data acquisition)**: A 67-year-old man comes to the physician with chest pain that radiates to his left arm. Which of the following would be most important to address in the history? a. character of pain b. duration of pain c. intensity of pain d. history of abdominal aneurism in family e. alcohol consumption. **Example (problem representation):** A 67-year-old man comes to the physician with chest pain that radiates to his left arm. Which of the following best characterizes his presentation based on the data provided? a. unstable angina b. typical chest pain c. stable angina d. pyrosis e. pleuritic. **Example (hypothesis generation)**: A 67-year-old man comes to the physician with chest pain that radiates to his left arm. Which of the following diagnoses should be considered (select all that apply)? a. myocardial infarction b. pulmonary embolism c. pericarditis d. aortic dissection e. COPD. **Example (illness script selection)**: A 67-year-old man comes to the physician with chest pain that radiates to his left arm. Which of the following is the most appropriate next step? a. troponin b. treadmill test c. CBC d. Chest CT e. GI cocktail.
2	**MCQ with audio/videos, images, and/or link to external resources**	Standard clinical vignette but descriptions also include audio and/or video files as well as images.	Good reliability and validity evidence.	Reliability and validity evidence. Broad content can be sampled. More authentic than typical MCQ.	Improved but still not fully authentic, cuing, all information is given. **Research:** use of external resources and how may impact reliability, validity, etc.	A 67-year-old man with hypertension, diabetes, and high cholesterol comes to the physician with chest pain that radiates to his left arm. Vitals are blood pressure 160/80 mm Hg, pulse 98/minute, respirations 18/min and temperature 99F. The patient is diaphoretic. Examination of the heart reveals (click for audio). ECG is shown. Which of the following is the most likely diagnosis? a) myocardial infarction b) pulmonary embolus c) aortic dissection d) gastroesophageal reflux disease e) pericarditis.	
3	**Short answer**	Standard clinical vignette and type in answer in computer -- short answer.	May have lower reliability (need more time for testing), validity evidence unknown. Prior work with reliability of .8 in 2 hours.	Broad content can be sampled. No cuing (must generate hypotheses/scripts).	Minimal validity evidence. Takes more time per item (impacts reliability) and more expensive to score if use either machine or human rating. **Research**: pilot machine rating and gather reliability and validity evidence.	A 67-year-old man with hypertension, diabetes, and high cholesterol comes to the physician with chest pain that radiates to his left arm. Vitals are blood pressure 160/80 mm Hg, pulse 98/minute, respirations 18/min and temperature 99F. The patient is diaphoretic. Examination of the heart reveals (click for audio). ECG is shown. What is the most likely diagnosis?	
4	**Script Concordance Testing (SCT)**	Based on script theory. A standard vignette is given. In a diagnostic SCT, the next step is providing the diagnostic script. Another piece of evidence is then given that must be "weighed" in light of the proposed script and vignette (typically on a much more likely through much less likely scale: 3- or 5-option scale that includes neutral category). These are studies assessing the viability of the SCT. They are not necessarily in formal examination settings.	Fair reliability, little validity. SCT shows some differences between levels of expertise but conclusions are somewhat tenuous. Problematic that partial credit is given and determining who are the experts for scoring the examination. Scoring is problematic in that raters must be in some agreement in order for the scoring to be plausible and this rarely occurs.	Designed specifically to mirror cognitive tasks for clinical reasoning. Items are as easy or easier to develop compared to MCQs. Can sample more items over a unit of time based on prior work.	Expert judgment in building scoring system which compromises reliability and validity. **Research**: could score based on complete consensus and give a free test justification. Could also test intermediate steps.	A 67-year-old man with hypertension, diabetes, and high cholesterol comes to the physician with chest pain. If you were thinking of a diagnosis of myocardial infarction and you learned that the chest pain radiated to his left arm this hypothesis would become (-2/much less likely, -1/less likely, 0/neutral, +1/more likely, +2/much more likely).	***SCT can be designed to assess various steps in our model***. Give the vignette. Assign a diagnosis (If you were considering a diagnosis of acute myocardial infarction and you found (Hx detail or PE finding), this diagnosis becomes (tests data acquisition and script selection) or modify lead in. Give vignette ask if you were considering an acute coronary syndrome (problem representation) and you found (insert finding).

5	**Key Features Examination (KFE)**	The key features approach was originally designed to replace longer cases known as patient management problems. Key features are defined as the critical steps in identifying or resolving a clinical problem. The longer cases were distilled down to the key branch points in clinical decision making, where identifying a critical piece of information would lead to a correct diagnosis or a crucial step in the management would determine the patient's outcome. The trigger is a case vignette, and the learner may be asked 2-3 questions about the case. Key features exams are used for the high stakes Canadian Qualifying Exam at the end of medical school and in other lower stakes settings, but are seldom used in the U.S.	Reliability evidence suggests that roughly twice the amount of time is needed to reach MCQ reliability (reliability of .8 with 8 hours of testing). Performance on KFEs is associated with performance in practice like prior MCQ literature.	Reflects actual practice. Can assess multiple correct answers.	Disadvantages of using key features exams are the extensive training required to develop the exams, and the lack of familiarity of U.S. trainees with the format. The literature provides some examples/guidelines for design and writing of KFEs: 1. Farmer EA, Gordon Page G. A practical guide to assessing clinical decision- making skills using the key features approach. Medical Education 2005;39:1188-94. 2. Schuwirth LWT, Blackmore DB, Mom E, Van de Wildenberg F, Stoffers H, Van der Vleuten CPM. How to write short cases for assessing problem-solving skills. Medical Teacher 1999;21(2):144 - 50).	A 67-year-old man with hypertension, diabetes, and high cholesterol comes to the physician with chest pain that radiates to his left arm. Vitals are blood pressure 160/80 mm Hg, pulse 98/minute, respirations 18/min and temperature 99F. The patient is diaphoretic. Examination reveals a systolic murmur. ECG demonstrates 2mm ST segment elevations in leads II, III, and AVF. 1. What clinical problems would you focus on in your immediate management of this patient? List up to three. 2. How should you treat this patient at this time? Select up to three (provide long list). 3. After management of the patient's acute condition, what additional measures, if any, would you take? Select up to four or select None, if none is indicated.	KFEs can assess the different steps in our model like MCQs or SCT above.
6	**Oral examination**	The oral exam is given by an expert assessor who gives some prompts that explore diagnostic reasoning.	High inter-rater reliability IF use dichotomous acceptable/ unacceptable decision. Reliability of .8 in 4 hours for dichotomous decision.	Can directly target multiple aspects of clinical reasoning. Can approximate authentic encounters in practice.	Expense, anxiety, feasibility, and other challenges of administering and rating of these examinations. This was abandoned in past due to concerns about feasibility, bias, rater training and psychometric properties.	A 67-year-old man with hypertension, diabetes, and high cholesterol comes to the physician with chest pain that radiates to his left arm. Vitals are blood pressure 160/80 mm Hg, pulse 98/minute, respirations 18/min and temperature 99F. The patient is diaphoretic. Examination reveals a systolic murmur. ECG demonstrates 2mm ST segment elevations in leads II, III, and AVF.1. What clinical problems would you focus on in your immediate management of this patient?	
7	**Clinical/ Comprehensive Integrative Puzzle (CIP)**	The CIP is like a clinical crossword puzzle. It asks examinees to compare and contrast a group of related diagnoses (typically 4-7) on a variety of domains such as history, physical examination, laboratories, etc. For each domain, descriptions are provided which fit a given diagnosis. The learner matches the appropriate domain description to the diagnosis (e.g., episodic, chest pressure with exertion relieved by rest would be matched with stable angina). It assesses the learner's ability to determine the key elements, or features, that discriminate one diagnosis from another.	Additional research needed to determine reliability and validity of this assessment. Early work suggests time needed for an item is similar to MCQ item.	Can target intermediate steps in clinical reasoning process. Can construct items that fall within domains of Bowen model for clinical diagnostic reasoning process and obtain scores within a diagnosis and within a "Bowen model" step.	Unknown reliability and validity. Like KFEs, will need to train faculty and develop exams and lack of familiarity with U.S. trainees with this format. The literature does provide an example/guideline for designing a CIP: Ber R. The CIP (comprehensive integrative puzzle) assessment method. Med Teach. 2003 Mar;25(2):171-6. **Research**: construct a series of CIPs on common diagnoses in internal medicine. Determine feasibility as well as psychometric features.	*COLUMNS ARE:* Medical history; Physical exam; Chest x-ray and ECG; Laboratory and other tests; Treatment and follow-up; Pathology. ROWS ARE: Unstable angina; Myocardial infarction; Rheumatic mitral stenosis; Acute pericarditis; Infective endocarditis; Hypertrophic cardiomyopathy. *Example from Ber 2003:* Selected matching items for inserting into table: Medical History a. A 28-year-old woman, in her third month of pregnancy, arrived at the emergency room because of severe shortness of breath (dyspnea).She complains of exertional fatigue from the beginning of her pregnancy, and increasing shortness of breath during the last week. b. A 25-year-old man complains of shortness of breath and dizziness on exertion. Both his grandfather and elder brother died suddenly at the age of 32 years.	Columns for CIP can address our model.
8	**Virtual Patients - on Computer**	Work through a scenario with different "paths" based on how steps are answered. High variability in how material is presented, how interactive the simulated case is, and feedback generated.	USMLE Step 3 evidence shows reasonable reliability and construct validity.	Scenarios can be designed to assess many aspects/all steps in clinical reasoning process (in theory), especially if use short cases.	Expense, more validity data needed, cannot sample content broadly if long case, unfamiliarity with testing format and development like KFEs. **Research**: currently constructing pilot simulation with Lifecom and exploring other options as well.	See USMLE Step 3 example video of patient with chest pain (http://download.usmle.org/PrimumTutorial/Primum_00_START.htm) and Annals Virtual Patient (http://vp.acponline.org/virtualpatients/Product/index) Could also review breathless case from our prior calls.	VPs, OSCEs, oral exam, and WBA could potentially target different steps in the process/our model with explicit questions.
9	**Objective Structured Clinical Exam (e.g., Diagnostic Reasoning Assessment (DRA)**	Subject of a 2006 dissertation, later published as a book, this method uses standardized patients as the basis for assessing clinical reasoning.	Fair: g-study showed reasonable variance due to examinee. Reliability of .8 in 4 hours.	Scenarios with SPs designed specifically to highlight clinical reasoning skills.	Challenges of having raters evaluate examinee performance. Costs, feasibility, sampling. Cannot sample content broadly given the amount of time. **Research**: pilot study in local diplomates to determine above issues.		
10	**Workplace-Based Assessments (WBA) (e.g., Mini-CEX or other direct observation)**	Directly observe individual in a health care setting. Could be done in real time or via (e.g., video). Following the encounter, ask additional questions, as needed, to assess clinical reasoning process.	Reliability of .8 in two hours.	Based on authentic encounters.	Unknown validity evidence. Resource intensive and need for sampling. Confidentiality issues of concern for both physician and patients. **Research**: patient wears pocket camera and physician is recorded and receives feedback on clinical reasoning (and/or other domains).		

Appendix E:

Patient Experience of Care

Assessment 2020: Patient Experience of Care Workgroup

May 6, 2014

Charge: *To provide a two-page concept paper defining the specific behaviors that physicians should exhibit to enhance the patient's experience of care and delineate approaches to assessing these behaviors with the aspiration of improving them.*

We summarize the main ideas generated from two conference calls with the workgroup and outline the rationale, definition, and suggested assessment approaches for this construct. All workgroup members endorsed the importance of all aspects of patient experience of care (Appendix 1) but, in particular, focused on the relational rather than the functional aspects of patient care, as these were skills not necessarily taught in medical training and not emphasized currently by the ABIM. Evidence was provided that showed some aspects of the patient experience of care are positively associated with clinical effectiveness and patient safety.

Rationale

The group considered both patient experience of care and patient-reported outcome measures, and evaluated examples of the domains being measured by each (Appendix 2). We concluded that, at least at this point in time, patient experience of care measures have been more validated (from an assessment perspective) and more directly focus on the patient and physician dyad; patient-reported outcomes focus more on functional aspects such as pain management and health status. Historically, the ABIM has emphasized the science behind being a good doctor while the patient, their values and individuality, were not given as much attention in ABIM programs. Thus, asking physicians to incorporate the values and preferences of patients and their families in medical treatment was deemed to be an important skill set for physicians to possess.

Definition

One component in the Doyle framework (Appendix 1) resonated with the workgroup and we adopted that focus for patient experience of care. This component is the second relational aspect -"Participation of patient in decisions and respect and understanding for beliefs, values, concerns, preferences and their understanding of their condition."

It was observed that if physicians could perform well on this component, the other relational aspects, including emotional and psychological support, involvement of family and caregivers in decisions, clear information, and transparency when something goes wrong, would follow. This definition felt very similar to that of shared decision-making (i.e., SDM is a process of a patient and physician collaboration to decide on a course of treatment from acceptable options). The best choice is based on medical evidence and meets the patient's needs, values, and preferences), the group felt that SDM might be too narrow since more general communication skills are not included in its definition.

Principles

Patient experience of care is a legitimate value and goal in its own right and should not be eclipsed by the notion that the construct is correlated positively with a variety of patient outcomes. We do not want physicians assuming that this is another way of measuring the patient outcomes – or the "real thing" – and, therefore, reacting negatively to additional oversight.

The methods we use to measure a physician's skill with respect to patient experience of care need to be considered carefully and the outcome of the measurement should be meaningful and fair. The outcome should reflect the goal of the physician meeting their patient's needs in a diplomatic and caring manner, regardless of their health status.

The public message about introducing Patient Experience of Care as a construct to be measured in ABIM's programs should be supported by the Assessment 2020 Taskforce and must be thoughtfully addressed as it will not be seen as straightforward or valuable to many physicians.

Assessments

The group reviewed the current literature on assessment approaches, the available tools, measurement evidence, as well as costs and benefits of each (Appendix 3). We considered a variety of assessments such as patient surveys (e.g., Consumer Assessment of Healthcare Providers and Systems (CAHPS) or Physician Achievement Review (PARs)), objective structured clinical examinations (OSCEs), and patient-physician interaction through live encounters or audio or video recordings.

Good patient experience of care surveys have the advantage that they are feasible to do, are based on a large, representative sample of patients from a broad sampling of interactions, and have been validated through research studies. The disadvantages are that they lack the rich context of specific patient-physician interactions. OSCEs have the advantages of a richer context with some structure over the patient's presentation and research exists informing the construction of reliable OSCEs. Their disadvantages are that they are more costly and less feasible to implement (e.g., encounters need to be arranged with either actors or live patients and the judges need to be trained), cannot sample broadly among types of interactions due to time constraints, and there are few research validity studies. Encounters through audio and video recordings show some promise in that they are quite authentic but have similar disadvantages to OSCEs. The group thought that crowdsourcing (a method to obtain ratings by soliciting contributions from an online community, used in modern learning environments) might be one approach to more feasibly generate ratings based on multiple stakeholders and multiple samples of behavior but research on this approach is needed.

Since the goal of measuring the patient's experience of care is for the physician to hear the patient's voice, the group recommends the following course of action for ABIM.

ABIM should conduct research by pilot testing several of the options, including crowdsourcing, to better understand the suitability of the approaches for ABIM. The purpose of this research will be to determine:

1. Who are the appropriate judges for rating patient experience of care?
2. How do we train the judges to provide fair and accurate ratings?
3. Do we need to rate the raters?
4. If some raters are consistently low or high, should their ratings get less weight?
5. How does ABIM ensure that personal information (of either physician or patient) is protected?
6. How does ABIM ensure that the assessment reflects the performance of the individual physician rather than the health system in which the physician works?
7. How many judges are needed in order to achieve an assessment that is reliable, valid, and fair?
8. How do we score the assessment and what feedback do we provide to the physician?
9. Will physicians in ABIM programs engage willingly in this and perceive it as fair?
10. Is there evidence to support the validity of whatever approach is chosen?

Appendix 1: Doyle (2013) Framework for Patient Experience of Care

	Relational Aspects	Functional Aspects
1	Emotional and psychological support, relieving fear and anxiety, treated with respect, kindness, dignity, compassion, understanding	Effective treatment delivered by trusted professionals
2	Participation of patient in decisions and respect and understanding for beliefs, values, concerns, preferences and their understanding of their condition	Timely, tailored and expert management of physical symptoms
3	Involvement of and support for family and care givers in decisions	Attention to physical support needs and environmental needs (e.g. clean, safe, comfortable environment)
4	Clear, comprehensible information and communication tailored to patients needs to support informed decisions (awareness of available options, risks and benefits of treatment) and enable self-care	Coordination and continuity of care, smooth transitions from one setting to another
5	Transparency, honesty, disclosure when something goes wrong	

Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. BMJ Open. 2013:3:e001570.doi:10.1136/bmjopen-2012-001570.

Appendix 2: Patient Experience of Care Survey Instruments

Consumer Assessment of Healthcare Providers and Systems - Group and Clinician (CG-CAHPS), Patient Satisfaction Questionnaire (PSQ III), Visit-Specific Satisfaction Questionnaire (VSQ), Physician Achievement Review (PAR), Consumer Satisfaction Survey (CSS), Ambulatory Care Experiences Survey (ACES), General Practice Assessment Survey (GPAS), VA Survey of Healthcare Experiences of Patients (SHEP), Components of Primary Care Instrument (CPCI), Primary Care Assessment Survey (PCAS)

Framework		CG-CAHPS	PSQ III	VSQ	PAR	CSS	ACES	GPAS	SHEP	CPCI	PCAS
Relational Aspects											
1	Emotional and psychological support, relieving fear and anxiety, treated with respect, kindness, dignity, compassion, understanding	X	X	X	X	X	X		X	X	X
2	Participation of patient in decisions and respect and understanding for beliefs, values, concerns, preferences and their understanding of their condition	X	X		X	X	X		X	X	X
3	Involvement of and support for family and caregivers in decisions								X		
4	Clear, comprehensible information and communication tailored to patients needs to support informed decisions (awareness of available options, risks and benefits of treatment) and enable self-care	X	X		X	X	X		X	X	X
5	Transparency, honesty, disclosure when something goes wrong										
Functional Aspects											
1	Effective treatment delivered by trusted professionals	X	X	X	X	X		X	X		X
2	Timely, tailored and expert management of physical symptoms	X	X	X	X	X	X	X	X		X
3	Attention to physical support needs and environmental needs (e.g., clean, safe, comfortable environment)	X	X	X	X	X	X	X	X		X
4	Coordination and continuity of care, smooth transitions from one setting to another	X	X	X	X	X	X		X	X	X

Appendix 3: Attributes of Patient Experience of Care Assessments

Name of Tool	Short Description	Type	Reference	Reliability and Content/ Construct Validity Evidence (Very Good/ Good/ Fair/Poor)	Pros	Cons
CG-CAHPS (Consumer Assessment of Healthcare Providers and Systems - Group and Clinician)	Asks patients in ambulatory care about their recent experiences with clinicians and their staff. Questions can be included that are directly about patient preferences such as "Did you and this provider talk about reasons you might NOT want to have the surgery or procedure?"	Patient survey via mail or phone	Dyer N, Sorra JS, Smith SA, Cleary P, Hays R. Psychometric Properties of the Consumer Assessment of Healthcare Providers and Systems (CAHPS®) Clinician and Group Adult Visit Survey. Medical care. 2012;50(Suppl):S28-S34.	Good	Generalizable – can be used in primary care and specialty care settings, acceptable reliability at the individual physician level.	To get reasonable reliability and validity, need a minimum of 45 surveys per provider, a strict sampling approach and a third party to administer it to patients; currently low fidelity to practice since relying on patient's memory of encounter.
PSQ III (Patient Satisfaction Questionnaire)	Measures several dimensions of patient satisfaction with medical care including interpersonal manner, communication, and time spent with doctor, among others.	Physician self-administered patient survey via mail	https://www.rand.org/content/dam/rand/www/external/health/surveys_tools/psq/psq3_scoring.pdf	Very Good	In-depth – the survey includes 7 subscales (e.g., general satisfaction, interpersonal care, etc.), strong psychometric properties (validity and reliability).	Lengthy (50 items), not administered online; low fidelity to practice since relying on patient's memory of encounter.

PAR (Physician Achievement Review)	A program utilizing a variety of questionnaires, including a patient questionnaire, to provide physicians with information about their medical practice from the viewpoint of people they serve.	Questionnaires administered via mail	Hall W, Violato C, Lewkonia R, et al. Assessment of physician performance in Alberta: the Physician Achievement Review. CMAJ: Canadian Medical Association Journal. 1999;161(1):52-57. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1232653/pdf/cmaj_161_1_52.pdf	Very Good	The program seeks feedback from different perspectives: the physician, the patient, medical colleagues, and non-physician health care workers, broad audience (it was developed for all Alberta physicians (office vs. hospital practice) and specialty groups can modify it to meet their needs), strong psychometric properties (validity and reliability).	Time-consuming – the physician must fill out a 26 question survey, and then distribute questionnaires to 25 patients, 8 medical colleagues, and 6 non-physician health care coworkers; costly - administration is about $200 per physician (in 1999 dollars), not administered online; low fidelity to practice since relying on people's memory of experiences.
OPTION (Observing Patient Involvement)	The OPTION scale assesses whether physicians include patients in decision making during a medical consult.	Trained observers use the scale to rate audiotaped consultations	Elwyn G1, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. Qual Saf Health Care. 2003 Apr;12(2):93-9. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1743691/pdf/v012p00093.pdf	Good	Generalizable to all types of medicine, acceptable psychometric properties, higher fidelity to practice since assessing an audiotape.	Requires trained observers as raters and difficult to calibrate the rater training; need more encounters to strengthen measurement properties.
PTS (Patient Trust Scale)	The PTS measures patient trust and the relationship with physician payment method.	Telephone survey	http://academicdepartments.musc.edu/family_medicine/rcmar/pts.htm	Fair	Strong internal reliability, short/concise (10-item questionnaire).	Not tested in elderly or minority populations, no information on external validity; low fidelity to practice since relying on patient's memory of encounter.

Common Ground Instrument	Examines patient-centered communication skills in office visits.	Four-station OSCE	http://www.stfm.org/fmhub/fm2004/march/forrest189.pdf	Fair	High fidelity to practice since observing encounters with standardized patients.	Studied in med students; fair intra-rater reliability; time-consuming.
Primary Care Assessment Survey	Measures the defining characteristics of primary care posited by the Institute of Medicine Committee on the Future of Primary Care. It measures seven features of primary care through 11 summary scales: access (financial and organizational), continuity (relationship duration and visit-based continuity), comprehensiveness (“whole-person” knowledge of the patient and preventive risk counseling), integration of care, quality of the clinician–patient interaction (clinician–patient communication and thoroughness of physical examinations), interpersonal treatment, and patient trust.	Patient questionnaire	http://journals.lww.com/lww-medicalcare/Abstract/1998/05000/The_Primary_Care_Assessment_Survey__Tests__of_Data.12.aspx	Good	Favored by patients over other similar surveys; good reliability; reasonable validity -associated with adherence and health status.	Lengthy (51 items); trust subscale was shown to be confusing for patients; low fidelity to practice since relying on patient's memory of encounter.
Components of Primary Care Instrument	It was developed for use in the Direct Observation of Primary Care study and designed to measure the processes of primary care rather than its structural or systemic aspects such as access to care. It has 43 questions and measures 8 domains of primary care.	Patient questionnaire	http://psycnet.apa.org/psycinfo/1997-06895-003	Good	Reasonable reliability and validity -- associated with screening, health habit counselling, and immunization.	Semantic differential response scale not well liked by patients; low fidelity to practice since relying on patient's memory of encounter.

VA Survey of Healthcare Experiences of Patients	Managed by the VHA Office of Quality and Performance (OQP), regularly solicits patient responses related to a specific and most recent episode of either outpatient or inpatient care.	Mail survey administered to random samples of VA patients on a monthly basis	http://annals.org/article.aspx?articleid=718025&issueno=12&atab=10	Fair In 2011, VHA replaced it with CAHPS.	Based on Pickering Institutes surveys, shows moderate reliability values, some scales are predictors on one-year mortality, post discharge.	Several items, relies on probability sample methods and being replaced by CAHPS; low fidelity to practice since relying on patient's memory of encounter.
Ambulatory Care Experiences Survey	Produces 11 summary measures of patients' experiences across two domains: quality of physician-patient interactions and organizational features of care.	Mail survey	http://onlinelibrary.wiley.com/doi/10.1111/j.1525-1497.2005.00311.x/full	Very Good	High reliability, shows correlations with other quality measures measured at the physician level, comprehensive domain of care domains, measures associated with preventive care processes.	Patient sample sizes must be 45 higher and based on probability sampling methods so data collection costs are high; low fidelity to practice since relying on patient's memory of encounter.
Primary Care Assessment Tool	Includes information on the focus of the health care facility, patient characteristics, services available onsite, and patient-, provider-, and facility-related perspectives on the experiences of care received and care provided.	Several modes: Face-to-face, telephone, or mail survey	http://www.jfponline.com/index.php?id=22143&tx_ttnews[tt_news]=167894	Very Good	Comprehensive assessment of a primary care system, good reliability on scales, factor loading on scales items show interpretable constructs.	Relies on probability sampling procedures, large sample sizes costs, several items and seven domain scales; low fidelity to practice since relying on patient's memory of encounter.
Crowdsourcing (Cordar and Lok)	Using Virtual Patients for Patient-Clinician Communication Training This study investigated how virtual patients and crowdsourcing could be used to train medical students on having empathy in patient interactions.	Crowdsource ratings of interactions with virtual patients	http://care.cs.columbia.edu/chi2013health/CRPapers/Cordar.pdf	None	Virtual patients less costly than standardized patients; crowdsourcing inexpensive, allows for large number of respondents and modern in its use of social media.	Studied in med students.

Always Use Teach-back	Sponsored by the Pickering Institute, the Iowa Health System will develop and implement an "Always Use Teach-back!" toolkit for the three care settings encountered by patients being discharged from the hospital: hospital discharge, primary care follow-up, and home health support. The toolkit will include training modules with videos demonstrating effective use of teach-back. IHS will train and coach physicians and nurses on the "Always Use Teach-back!" approach.	Web-administered training for caregivers including videos; trainees receive feedback from other team members who complete an observation form	http://www.teachbacktraining.org/	None	Pickering Institute tool, practice-based.	No patient input; system has little evidence of effectiveness; developed as teaching not assessment tool.
USMLE Step 2 Clinical Skills exam	Joint program of the Federation of State Medical Boards (FSMB) and the National Board of Medical Examiners to test doctor-patient communication skills.	Standardized patient simulations	Whelan GP, McKinley DW, Boulet JR, Macrae J, Kamholz S. Validation of the doctor-patient communication component of the ECFMG Clinical Skills Assessment. Med Educ 2001;35:757-61.	Good	Reasonable reliability and some evidence of construct validity (with ABIM's program director ratings of communication skills); higher fidelity to practice.	Expensive to administer.

SCOPE (Studying Communication in Oncologist–Patient Encounters)	The SCOPE at Duke University, Durham VA and University of Pittsburgh was a trial where clinic visits between participating oncologists and their patients with advanced cancer were audio recorded. Surveys evaluated patients' trust in their oncologists and perceptions of their oncologists' communication.	Audio recordings and patient surveys; trained raters evaluate audio recordings	Tulsky JA, Arnold RM, Alexander SC, Olsen MK, et al. Enhancing communication between oncologists and patients with a computer-based training program: a randomized trial. Ann Intern Med. 2011;155:593-601.	None	High fidelity to practice – physicians recording their actual patients encounters so most meaningful.	Technology may be problematic; need several raters per recording and several patients. Measurement properties of instrument not studied. Need consent from patients.

Appendix 4: References used in the Patient Experience of Care workgroup

1. Instruments Available for Use in Assessment Center. 2013; http://www.assessmentcenter.net/documents/InstrumentLibrary.pdf. Accessed May 20, 2014.

2. Appendix A: Measures for the Adult 12 Month-Survey. 2013; https://cahps.ahrq.gov/surveys-guidance/cg/cgkit/1309_CG_Measures.pdf. Accessed May 20, 2014.

3. Black N. Patient reported outcome measures could help transform healthcare. *BMJ (Clinical Research Ed.).* 2013;346:f167-f167.

4. Boulet JR, Murray D. Review article: assessment in anesthesiology education. *Canadian Journal Of Anaesthesia = Journal Canadien D'anesthésie.* 2012;59(2):182-192.

5. Doyle C, Lennox L, Bell D. A systematic review of evidence on the links between patient experience and clinical safety and effectiveness. *BMJ Open.* 2013;3(1).

6. Elwyn G, Edwards A, Wensing M, Hood K, Atwell C, Grol R. Shared decision making: developing the OPTION scale for measuring patient involvement. (Original Article): British Medical Association; 2003:93.

7. Gravel K, Légaré F, Graham ID. Barriers and facilitators to implementing shared decision-making in clinical practice: a systematic review of health professionals' perceptions. *Implementation Science: IS.* 2006;1:16-16.

8. Lin GA, Fagerlin A. Shared decision making: state of the science. *Circulation. Cardiovascular Quality And Outcomes.* 2014;7(2):328-334.

9. Manary MP, Boulding W, Staelin R, Glickman SW. The patient experience and health outcomes. *The New England Journal of Medicine.* 2013;368(3):201-203.

10. Ting HH, Brito JP, Montori VM. Shared decision making: science and action. *Circulation. Cardiovascular Quality And Outcomes.* 2014;7(2):323-327.

11. Valderas JM, Kotzeva A, Espallargues M, et al. The impact of measuring patient-reported outcomes in clinical practice: a systematic review of the literature. *Quality of Life Research.* 2008;17(2):179-193.

Appendix F:

Teamwork

Assessment 2020: Teamwork Workgroup

Fall 2014

Charge: *To provide a concept paper that identifies essential components of teamwork in health care and approaches to assessing teamwork in physicians.*

This paper summarizes overarching themes identified over the course of a month-long e-mail discussion among members of the teamwork workgroup. The document also provides a working definition of teamwork and describes approaches to assessing teaming while highlighting several of the specific tools currently available.

In the discussion, workgroup members acknowledged the growing recognition of teamwork's importance in health care. However, there was some uncertainty around how to best measure an individual's ability on a team and around the most effective way to provide meaningful feedback to physicians. There was a general consensus that assessment of teamwork is more appropriate as a voluntary activity than a mandated part of Maintenance of Certification (MOC) and that assessment tools need to account for a variety of practice settings. Additionally, evidence was presented demonstrating the impact of effective teamwork on health outcomes.

Rationale

Evidence in the literature supports the notion that teamwork is an important factor in a patient's safety and health outcomes, especially as it relates to adverse events. Through qualitative interviews, thought leaders in health care identified teamwork as a crucial skill for physicians of the future.

There is a significant body of research on the subject but we have chosen to feature some key articles that provide rationale for the importance of teamwork in health care settings:

- *Teamwork and patient safety in dynamic domains of health care: a review of the literature* (1)
  A systematic review on the methods used to study teamwork and the facets of teamwork that are relevant to quality of care and patient safety.

- *Team-training in health care: a narrative synthesis of the literature* (2)
  Another systematic review, this article emphasizes the impact of team training in health care settings, focusing on teamwork behaviors, knowledge and attitudes. This review also provides evidence to demonstrate the effect of team-training on clinical outcomes, and quality and safety indices.

- *Does training in obstetric emergencies improve neonatal outcome?* (3)
  An example of how multi-professional training on obstetric emergencies for all team members of obstetric medical staff resulted in a drop in low Apgar scores and hypoxic-ischaemic encephalopathy. This study was unique in that all staff members, from midwives to obstetricians and anesthetists, were required to take part in the training.

Definition

The following describes the state of the field of teamwork:

Teamwork refers to two related things: the cooperative or coordinated effort of a group of individuals acting together to achieve a common goal, and the efficiency and effectiveness of the work of a group or team.

Assessments

There are three main approaches used to assess teamwork in health care, each with its own strengths and limitations.

The three kinds of teamwork assessment are:

1. Multi-source feedback, in which members of a team assess the work they do together
2. Observation of teams in context, in which an outside observer looks at how a team performs, often using a defined rubric or checklist of teamwork behaviors/practices
3. Simulation, in which a team performs its work out of context so that it can be more carefully observed and analyzed

Each of these can be used to assess a team as a whole, or individuals within the team.

For example, the American Board of Internal Medicine (ABIM) Teamwork Effectiveness Assessment Module is a multi-source feedback assessment that focuses on individual physicians within a team; by contrast, the Relational Coordination Survey and the Team Strategies and Tools to Enhance Performance and Patient Safety (TeamSTEPPS) Team Performance Questionnaire use a multi-source feedback approach to assess the quality of teamwork of an entire team.

Instruments like the Anaesthetists' Non-Technical Skills assessment or the Interprofessional Professionalism assessment are checklists of behaviors that can be used by a teacher/supervisor or other kinds of observers to evaluate the teamwork skills/behaviors of a single team member as they perform their work. The TeamSTEPPS Team Performance Observation tool provides a framework for an observer to look at an entire team.

Simulations can be constructed to assess individuals on a team, or to assess the entire team – as when an entire unit or hospital uses "in situ" simulation to evaluate how well the organization as a whole responds to a staged crisis event.

Each of these approaches has its own strengths and limitations. Multi-source feedback is excellent for providing information on how the members of a team evaluate the team's work, but does not provide data suitable for summative evaluation or comparison with others. Observation by an outside observer can provide a more objective view of a team's work, but not all healthcare teams work together directly (making them hard to observe), and an outside observer can miss important nuances that affect a team's work. Simulation isolates an individual or group, allowing their work to be more closely analyzed, but the logistics are demanding and not all teams do work that is amenable to being simulated.

It has been suggested that we look to other means of analysis as well as other fields for assessing teamwork. Some of the most promising work on measuring teamwork and team effectiveness comes from studies using social network analysis (SNA). Stephen J. Lurie, MD, PhD, and others have used SNA to study team function (including a family member) in an intensive care unit. SNA can be used to create a "snapshot" of team behavior but it is not yet clear how to use the method for evaluation without a credible "gold standard," that is, a scientific approach to assessing and comparing the quality of the networks that the analysis defines. (4)

There were a number of recurring themes that emerged from the discussion about ABIM's role in assessing teamwork. The group recommends the following course of action for ABIM:

- Assessment of teamwork should be voluntary, not a mandatory part of MOC (at least at this point in time).
- Teamwork assessment tools should be sensitive to practice setting.
- Longitudinal assessments are as important as one-time assessments.
- There's a difference between assessing individuals within a team and the function of a team as a unit; ABIM may be best suited to measuring an individual in team-oriented situations.
- The goal of assessment should be credible feedback physicians can use to assess and improve their own practice and performance. Assessments should include multi-source feedback, with a diverse and large, informed sample of perspectives.
- There is evidence of the importance of teamwork on health outcomes (e.g., impact of training of obstetric teams on neonatal outcomes).
- We should look to other professions for paradigms of effective teamwork assessment (e.g., teams that run U.S. nuclear power plants; military).

Appendix 1

Manser (2009) Model of Teamwork in a Health Care Setting

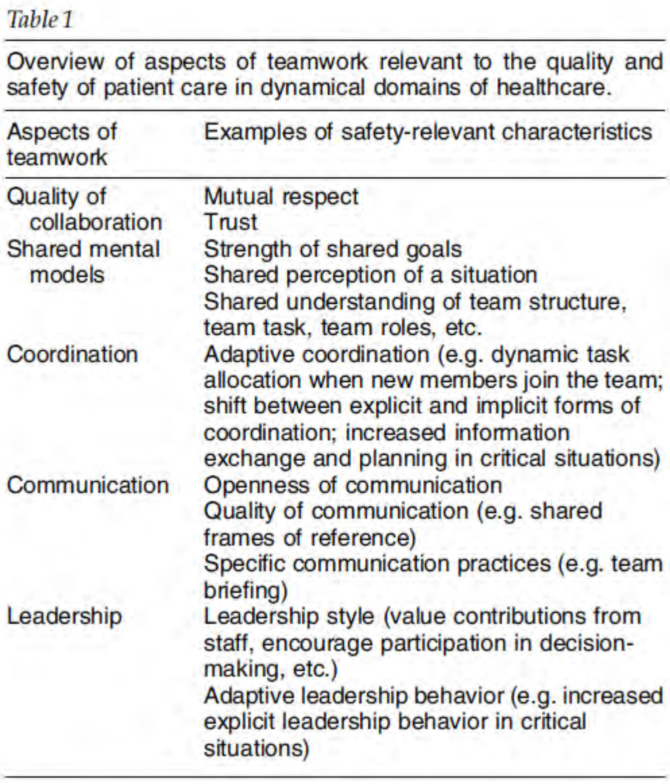


Table 1

Overview of aspects of teamwork relevant to the quality and safety of patient care in dynamical domains of healthcare.

Aspects of teamwork	Examples of safety-relevant characteristics
Quality of collaboration	Mutual respect Trust
Shared mental models	Strength of shared goals Shared perception of a situation Shared understanding of team structure, team task, team roles, etc.
Coordination	Adaptive coordination (e.g. dynamic task allocation when new members join the team; shift between explicit and implicit forms of coordination; increased information exchange and planning in critical situations)
Communication	Openness of communication Quality of communication (e.g. shared frames of reference) Specific communication practices (e.g. team briefing)
Leadership	Leadership style (value contributions from staff, encourage participation in decision-making, etc.) Adaptive leadership behavior (e.g. increased explicit leadership behavior in critical situations)

Appendix 2

References Used in the Teamwork Workgroup

1. Manser T. Teamwork and patient safety in dynamic domains of healthcare: a review of the literature. Acta Anaesthesiol Scand. 2009 Feb;53(2):143-51.
2. Weaver SJ, Dy SM, Rosen MA. Team-training in healthcare: a narrative synthesis of the literature. BMJ Qual Saf. 2014 May;23(5):359-72.
3. Draycott T, Sibanda T, Owen L, Akande V, Winter C, Reading S, Whitelaw A. Does training in obstetric emergencies improve neonatal outcome? BJOG. 2006 Feb;113(2):177-82.
4. Lurie SJ, Fogg TT, Dozier AM. Social network analysis as a method of assessing institutional culture: three case studies. Acad Med. 2009 Aug;84(8):1029-35.

Appendix G:

External Resources Manuscript

Comparing Open- and Closed-Book Examinations: A Systematic Review

Steven J Durning, MD, PhD

S.J. Durning is Professor of Medicine and Pathology, Uniformed Services University of the Health Sciences (USUHS), Bethesda, MD

Ting Dong, PhD

T. Dong is Assistant Professor of Medicine, USUHS

Temple Ratcliffe, MD

T. Ratcliffe is Assistant Professor of Medicine, USUHS

Lambert Schuwirth, MD, PhD

L. Schuwirth is Professor of Medical Education, Flinders University, Australia

Professor for innovative assessment, Maastricht University, The Netherlands

Anthony R Artino, Jr, PhD

A.R. Artino is Associate Professor of Medicine, USUHS

John R Boulet, PhD

J.R. Boulet is Vice President, Research and Evaluation, Foundation for Advancement of International Medical Education and Research

Kevin Eva, PhD

K. Eva is Senior Scientist, Centre for Health Education Scholarship, University of British Columbia

Word count: 4554

Corresponding author: Steven J Durning, MD, PhD.

Department of Medicine (ICR), Uniformed Services University
4301 Jones Bridge Road
Bethesda, MD 20814
Tel: 301-295-3603
Fax: 301-295-3557
Email: steven.durning@usuhs.edu

Background: The rapid expansion of knowledge and the emergence of technology that enables physicians to access an unprecedented amount of information raise fundamental questions about examination practices.

Objective: To compare the relative effectiveness and relevance of open-book examinations (OBEs) and closed-book examinations (CBEs).

Methods: Systematic review of peer-reviewed articles retrieved from MEDLINE, ERIC, EMBASE and PsycINFO (1962-June 2014).

Results: From 4192 studies, 37 were included. There was a fair amount of diversity, both in terms of level of the learner and subject studied. The frequency with which outcomes were identified was as follows: (1) exam preparation (n=22, 59%), (2) test anxiety (n=14, 38%), (3) exam performance (n=28, 76%), (4) psychometrics and logistics (n=8, 22%), (5) testing effects (n=24, 65%) and (6) public perception (n=5, 14%). With respect to pre-examination outcomes, findings were equivocal, but if there is an impact it favors the argument that students prepare more extensively for CBEs. For during-examination outcomes, it appears examinees take longer to complete OBEs. Studies addressing examination performance favored CBE, particularly when preparation for CBE was greater than for OBE. With respect to post-examination outcomes, the evidence suggests little difference in testing effects or public perception.

Conclusions: Given the data available to date, there does not appear to be sufficient evidence for exclusively using one or the other testing format. As such, we believe that a combined approach could become a more significant part of health professional testing protocols as licensing, certification and recertification bodies seek ways to assess critical competencies other than the maintenance of medical knowledge.

The rapid expansion of knowledge and the emergence of technology that enables healthcare practitioners to access an unprecedented amount of information raise fundamental questions about the adequacy of the closed-book examination (CBE) practices commonly used by the health professions. Some scholars argue that, in a world of exponential knowledge growth, any examination of relevance must assess the examinee's ability to find, understand, evaluate, and use external resources. Such proponents of the open-book examination (OBE) argue that these exams are more authentic to real-world practice and carry the message that success is not about "rote memorization"[1, 2, 3]. Since professionals of the future will not be able to "know" all the information needed for competent performance[4], meaningful assessment of medical practice, the argument goes, should include provisions that allow individuals to look up information to arrive at the correct answer.

Other scholars who defend the status quo of CBEs cite the expertise literature that has consistently found expert performance to be closely tied to rich and well-organized content knowledge of a given subject. For example, a number of studies have found that high performance on CBEs is associated with better practice outcomes.[5, 6] In many professional situations a physician's ability to look up unknown information is restricted by constraints such as time and internet access. As such, well-organized, content-specific knowledge remains the primary prerequisite for expert performance. From this perspective, merely putting a vast amount of information at a physician's "fingertips" is not likely to result in improved care because, to be effective, the physician needs knowledge to guide their search and needs to be able to integrate any new information with their existing knowledge and experience. Stated another way, reliance on information technology has the potential to detrimentally increase cognitive load (e.g., mental effort), decrease learning and critical appraisal of information, and ultimately harm patient care.[7]

Within this debate, it is clear that views on what defines a competent healthcare professional are changing. Where formerly the focus laid almost entirely on the possession of knowledge, currently physicians are expected to be able to use external point-of-care knowledge. For modern assessment to be aligned to the changing notion of medical competence, it is important to better understand the various pros and cons of OBE and CBE assessment approaches. This is true both in terms of promoting assessment-for-learning and in high stakes contexts such as credentialing and licensing assessment.

To inform this important issue, which impacts the examination of physicians across the continuum of their professional careers, we conducted a systematic review of the literature comparing the two assessment strategies. Our fundamental questions were: (1) what is the evidence regarding the comparative effectiveness of OBEs and CBEs? and (2) how might these findings inform current examination practices and future research in health professional education? To be inclusive, we broadly defined OBEs as a test or assessment that allows the use of any resource such as the internet, a textbook, course notes, or journals, and we searched for studies in all educational fields, not just medical education.

Methods

We began this systematic review with a scoping search of the topic by two of the study authors (SJD and TD). No prior systematic reviews on the topic had been performed, and a scoping search was thought necessary to better understand the breadth and depth of the relevant literature. This initial search included MEDLINE and ERIC and was conducted in the spring of 2013. The following terms were used: open book examinations, closed book examinations, comparing open and closed book examinations, resources with examinations, multiple choice examinations, examination format, examination type, open book tests, and closed book tests. A third investigator, who is a research librarian, conducted a separate scoping search using the same data sources. The three investigators compiled articles from these sources resulting in 488 citations and the titles and abstracts from these citations were reviewed for inclusion. Articles were excluded only if they were deemed to be unrelated to our review, available only in abstract form, not available in English, or represented textbooks. This process resulted in 78 citations that were discussed and underwent further review. We then iteratively generated a list of themes that could be used as preliminary outcome categories for a more comprehensive systematic review and used this step to further refine our inclusion and exclusion criteria and our search strategy and terms (Appendix 1) within the various databases included in our review.

For our systematic review, we followed established PRISMA Guidelines[8] and specific guidelines provided in the medical education literature.[9] We limited our search to full-length published peer-reviewed journal articles involving learners in either descriptive reports or educational interventions, using any study design to address our research questions. We further limited the papers reviewed to those that compared (either directly or indirectly) open- and closed-book examinations, and for which an English version of the paper was available.

Relevant studies were identified by searching three databases (during the summer of 2013 and included no date restrictions (e.g. what was available until date searched): 1) Ovid Medline (June 2013), 2) Ovid Embase (July 2013), and 3) ERIC (June 2013) We included the following keywords or their combinations: open book exam, open book examination, open book test, closed book exam, closed book examination, closed book test, computer aided test, web based examination. To identify additional studies, we searched the bibliographies of those studies found by our electronic search, contacted experts in the field, and conducted an open web search using Google Scholar and PsycInfo using the search terms listed above. Appendix 4 displays the terms used for the systematic search.

A data collection form was then used to rate each paper. This form was constructed based upon the findings of our scoping review and refined through a series of conference calls between the authors. The form included details on the study type, setting, participant demographics, outcome measures, study quality, limitations, and additional comments. The form was pilot tested and revised by having each member of the investigative team review two articles. We subsequently reviewed the form and discussed additional articles by conference call until consensus on the form was achieved. The final version of the data collection form is included in Appendix 3.

Three authors (SJD, TD, TR) independently reviewed the titles and abstracts of the retrieved publications. Each was initially categorized as *include*, *exclude*, or *uncertain*. All *include* and *uncertain* titles and abstracts were reviewed in the subsequent stage (i.e., review of the full text version of the papers; see Figure 1). Authors disagreed regarding inclusion for 44 of the 4192 titles and abstracts (see Figure 1), all of which were subsequently included in the full paper review. After this review, 299 articles remained. The same three study authors then reviewed the full text of all 299 articles (see Figure 1) using the same categorization framework (*include*, *exclude*, *uncertain*). In doing so, 193 were deemed beyond the scope of this review. The remaining 106 full text papers underwent a more detailed review and coding by the larger study team with each paper having at least two reviewers. Sixty-nine articles were excluded following this additional round of review, which included a series of conference calls and detailed coding using the data extraction form. Ultimately, 37 papers were included in our review.

We used the categorization framework from our scoping search to structure the outcome categories. We report them here in the sequence in which they would occur in the testing process: (1) examination preparation, (2) test anxiety, (3) exam performance, (4) psychometrics and logistics, (5) testing effects, and (6) public perception. Any article could have multiple outcomes and was reviewed for relevant themes by two of the study authors. Following review and coding, conference calls were held between all coders until complete agreement was achieved for the coding of every article. A third coder was needed to resolve conflicts for 3 of the 37 papers.

Finally, the extent to which the research found was fit for purpose was evaluated by having each reviewer code the manuscript for the presence of clear research questions and hypotheses and by recording judgments of quality (using a 5-point rating scale). These latter judgments were made in relation to the degree to which each study effectively addressed a research question comparing the relative benefits of OBE vs. CBE.

Results

We retrieved 4192 citations from the literature (all search engines included all dates to present time of search), which resulted in 37 articles being included in our review (see Figure 1 and Appendix 1). The frequency with which outcomes were identified was as follows: (1) exam preparation (n=22, 59%), (2) test anxiety (n=14, 38%), (3) exam performance (n=28, 76%), (4) psychometrics and logistics (n=8, 22%), (5) testing effects (n=24, 65%) and (6) public perception (n=5, 14%). We first report findings for study quality and context and then discuss each outcome category individually.

Study quality

Overall, the quality of the papers included in our review was deemed to be adequate. Table 1 presents descriptive statistics outlining the quality of the research found. Explicit research questions were presented in 31 papers (84%), hypotheses were stated in 14 (38%) and hypotheses were justified in 10 (27%). Conceptual and/or theoretical frameworks were described in 7 papers (19%).

[Insert Table 1 here]

Study context

Thirty-four investigations (92%) were single institution studies. Nearly half were performed in the US (n=18, 49%). Other locations included the Netherlands (n= 5, 14%), the United Kingdom (n= 4, 11%), Greece (n=3, 8%), and Australia (n=2, 5%) and one study (3%) was included from each of the following countries: Canada, Denmark, Norway, Africa, and Israel. The majority of studies pertained to college-level students (n=24, 65%), two studies investigated high school students (5%), eight investigated medical students (22%; 2 of these were multi-institutional), two investigated other post-college instructional settings (5%), and one study (3%) included physicians in practice. For the majority of the students, the stakes of the examination were rated as medium (n=21, 57%) in that the assessments were generally end-of-course examinations. Two (5%) were considered high stakes, being equivalent to national licensing type examinations. A minority of studies included a formal incentive (n=6, 16%) to participants beyond a course grade and those incentives generally took the form of extra credit or a small payment for participation.

Only a minority of studies reported enrolling participants with significant prior experience with OBE (n=7, 19%). A few papers reported that participants had some experience with OBE (n=4, 11%), but most articles either reported that participants had no prior experience or prior experience was not mentioned (n=26, 70%). Because the findings did not appear to differ based on learner (e.g., high school, undergraduate, graduate, or physicians in practice), we describe the findings in each theme as a whole, unless otherwise stated. Appendix 1 provides detailed results for each paper by themes outlined below. Some papers were coded under more than one theme.

(1) Exam Preparation

Of concern in the OBE vs. CBE debate is that the format of the exams will have a fundamental influence on test preparation (and, hence, learning). As alluded to previously, some argue that CBEs may promote more superficial learning by requiring students to memorize large amounts of material whereas OBEs may focus learners on the application of what they have learned. Others argue that CBEs, compared to OBEs, prompt students to study more as they will not be able to look things up during the exam. Appendix 1 lists the studies comparing OBE and CBE that investigated exam preparation.

In terms of preparation time, findings were inconsistent across studies, but in sum appear to favor CBEs. Some showed that students reported more preparation time for CBEs than OBEs [10, 11, 12](Appendix 1) or attended class less often if the test was an OBE[12]. Others, however, reported that students prepared for OBEs and CBEs similarly[13, 14]; no studies reported more preparation time for OBEs than CBEs. Of note, preparation time is not always a good proxy measure of learning or test performance; an increase in preparation time could just as easily indicate insufficient prior engagement with the material rather than being a precursor to improved performance.[15]

Reviewing the articles that examined preparation strategy revealed some studies[16, 17] that reported students did not change study tactics for OBEs vs. CBEs, and no correlation between test format and deep vs. surface learning approaches were found.[17] Others did find differences[2, 18], but it was unclear exactly what was being measured by these outcomes.

In summary, research exploring exam preparation was equivocal with respect to whether students prepare differently (or at greater length) for CBEs or OBEs. When differences did exist they tended to show that participants at all levels of education studied more when they expected a CBE versus an OBE.

(2) Test Anxiety

Emotions have long been known to affect cognitive performance.[19] Although negative emotions were once thought to have exclusively deleterious effects on test performance, contemporary theories of emotion suggest that such an assumption is overly simplistic.[20] For example, a negative emotion such as anxiety might actually motivate a student to study for a CBE, which could result in superior performance when compared to an unstressed student preparing for an OBE. Regardless, reducing test anxiety is often reported to be one of the main motivations for considering OBEs. Unfortunately, our findings indicate that anxiety effects were typically examined as a secondary issue relative to a study's primary purpose (see Appendix 1), and all studies that assessed emotions lacked a theoretical grounding. In particular, of the 14 studies with emotion-related outcomes, none employed a theory of emotion to help frame the study or explain the findings.

What evidence does exist suggests that students may overestimate the impact that OBEs or partial OBEs (i.e., exams in which students can bring some prepared material like a "cheat sheet" rather than having access to anything desired) have on reducing their anxiety. Several studies suggest that students associate OBEs with less anxiety[16, 21, 22], but that only a minority of students actually report lower anxiety.[22, 23] Similarly, Baillie and Toohey[23] conducted a study in an engineering context and found that the anxiety associated with taking OBEs was not reduced as much as the authors had expected, with a large proportion of students (45%) reporting being just as stressed with OBEs as CBEs. It has been suggested that certain aspects of OBEs, such as the belief that examiners will choose questions of greater difficulty, can be anxiety-provoking for students.[24] It remains to be seen whether students overestimate the impact that OBEs or partial OBEs have on reducing their anxiety because they lack familiarity with the test format.

Taken together, findings from the limited research on anxiety and its relationship to OBE and CBE formats suggest that students may overestimate the impact that OBEs or partial OBEs have on reducing their anxiety and, by extension, potentially improving their performance. In addition to incomplete reporting of methods and analyses, and the fact that anxiety effects are often explored as an afterthought, another major problem is lack of theoretical grounding.

(3) Exam performance

The most common outcome explored was examination performance, defined as a comparison of student achievement on OBE relative to CBE formats (see Appendix 1). Intuitively, one might expect that examinees would perform better on OBE because they have the capacity to look up answers, leading proponents of CBE to argue that score inflation will result in an inappropriately high number of candidates passing the examination. Opponents suggest that the OBE format does not inherently lessen difficulty, but frees the examiner to focus questions on the ability to apply knowledge (i.e., testing an aspect of ability that cannot simply be "looked up"). Further, the time required to look up information can increase difficulty by creating pressures towards efficiency. Two caveats are particularly noteworthy when considering exam performance: (1) in most studies, students had little to no experience with OBEs - only one study[25] that explicitly addressed examination performance reported that students had such prior experience; and (2) exam performance is a particularly difficult outcome to grapple with because the difficulty of an exam depends on the questions asked and some proponents of OBE argue that the main advantage of this testing strategy is enabling a different style/focus of questions to be addressed. Different questions across different examination formats may, therefore, be required to enable the advantages of OBEs to be recognized while simultaneously preventing one from making direct comparisons about the strength of one's performance from one format to the next.

The majority of the examinations were MCQ format, but some were also essay and/or short answer (Appendix 1). Typically, no significant difference in examinee performance was found [26, 27, 28] or performance favored CBEs (Appendix 1). In investigations favoring CBEs, when explored, the authors generally suggested that the reason for the difference related to examination preparation. Some studies did show better performance on OBEs immediately after learning, but even those differences did not persist over time (i.e., OBE and CBE performance was equivalent or CBE performance was superior on a subsequent delayed test; Appendix 1).

In terms of the relationship between test preparation and exam performance, an investigation by Block is particularly useful. In the first experiment, CBE expectancy led to a 10% increase on a delayed transfer test over OBE. In a second experiment demonstrating improved performance with CBEs, subjects reported spending less time studying (i.e., less exam preparation) when expecting an OBE. In a different study performed by Carrier, (3) students scored significantly lower when expecting an OBE versus when expecting a CBE for their final examination. The author suggested that this may be due to deepening examinees' approach to learning (defined as studying lecture notes, making chapter notes, highlighting and/or underlining, and coming to office hours, activities that correlated with higher exam scores). In one study, students commented that they were less prepared for the final examination because they expected to be able to find the answers in the book during the examination with an OBE. To counter the notion that lower performance is due to being unable to find material in a resource during an OBE, three studies reported that the preparation of OBE materials (e.g., cheat sheets or note cards) was not sufficient to improve performance on a CBE[29, 30, 31], suggesting that the difference may be due to differences in learning. Finally, in an investigation comparing OBE and CBE with a CBE final examination, students in the experimental section scored lower and recalled significantly less about topics that were covered on OBEs than those covered by CBEs. These results suggest that OBEs may impede long-term learning of material (at least in the context of an introductory biology course).

In sum, studies comparing exam performance appear to favor CBE. However, the combination of relatively little experience with OBE and the differences in exam preparation noted in several investigations highlighted in this section leave open the possibility that OBE performance could be improved through instructing students about OBEs or providing practice tests. On this point, three sets of authors indicated that students need to have the right expectation for what it takes to do well in OBE.[23, 24, 25]

(4) Psychometrics and Logistics

Despite many strongly held intuitions about the intrinsic value of different testing formats (e.g., multiple choice inducing recognition memory vs. short answer format requiring recall), research has generally shown that the validity of a test is determined more by the content of the questions included than by the examination format.[32, 33, 34]

Only two studies were found that directly examined the impact of the exam format on the psychometric utility of the assessment. One comparison was limited because test content and number of questions were confounded with assessment format[3], while the second study concluded that a suitably constructed set of questions could be used to discriminate student abilities in either an OBE or CBE environment.[35]

It may not be realistic in real world settings to compare reliability across test format while keeping the number of items constant. Three studies that compared CBEs to OBEs with respect to their influences on amount of time required to take the test found that students took 10-60% longer to complete OBE tests relative to CBE tests.[10, 28, 35] In other words, if one controls for amount of testing time it is likely that fewer questions would be asked in OBE format and, hence, the reliability of the equivalent CBE formatted exam can be anticipated to be higher.

(5) Testing effects

Testing effects arise when taking an exam improves subsequent performance. Such benefits can arise in indirect ways (e.g., being prompted to study, as outlined in the exam preparation section) or from direct effects of the material becoming more memorable when participants are tested on it than when they simply study for a test.[36] Most commonly, direct testing effects are demonstrated by separating research participants into two groups, one of which is asked to study material and then take an intervention test, while the other group is asked to study only (multiple times to equate the time participants are exposed to the material across groups). The testing effect is demonstrated when the tested group outperforms the study group on a subsequent outcome exam. This testing effect (or test-enhanced learning) has been well-documented in multiple fields.[37]. The most commonly supported hypothesis is that the act of testing creates a desirable difficulty that requires one to retrieve knowledge from memory, thereby making that information more retrievable in the future[37].

Whether the explicit act of memory retrieval is required (as it is in CBE) and/or it is the act of struggling with the information that matters (as occurs in both CBE and OBE) remains to be determined as authorities continue to debate the relative merits of learning with OBE vs. CBE. Proponents of CBE argue that learning requires active construction of memory that is less likely to occur when one relies on external resources to answer test questions. OBE proponents, in contrast, argue that such examinations may enhance the ability to apply knowledge because rote memorization is not the emphasis.

According to the empirical evidence, both OBE and CBE demonstrate testing effects (Appendix 2). Four of the six studies comparing OBEs and CBEs demonstrated testing effects that were roughly equivalent between exam formats[10, 13, 38, 39],(Appendix 2). The testing effect of CBEs was found to be superior to that of OBEs in one study.[12] These researchers demonstrated that during a summative CBE participants performed worse on material covered by an OBE intervention relative to a CBE intervention.[12] In one investigation, CBE without feedback yielded lesser testing effects than CBE with immediate feedback or OBEs; subsequent experiments, however, found OBE and CBE without feedback to be equivalent.[10, 38]

Overall, it appears that testing effects are observed regardless of whether OBEs or CBEs are used. Consistent with prior studies of testing effects, students' collective self-perceptions ran counter to the finding that testing effects occur; students felt that studying alone was more effective preparation than taking either an OBE or CBE as the intervention test.[38]

(6) Public Perception

Public perception was viewed through two lenses—the learner's perspective and the perspective of teachers. No studies incorporating the views of patients were found. From the perspective of the learner, some studies suggest that OBEs were seen to have several advantages over CBEs.[2, 17, 24] On the other hand, students also commented in several studies that OBE questions were more difficult and that they desired additional practice or training for the OBE format.[17]

Teachers' views often challenged the implementation of OBEs. Teachers familiar with administering CBE expressed concerns over the increased resources associated with preparing OBEs as well as the perception that additional time was required for learners to take OBEs.

Discussion

Overall, the empirical literature comparing OBEs and CBEs is fairly limited. Of the studies that do exist, there is a fair amount of diversity, both in terms of learner level and the subjects studied (see Appendix 1). While it can be challenging to generalize these findings from diverse learner groups and academic subjects to the field of medicine, this diversity is potentially beneficial when attempting to gain a general picture of the influence of exam format.

Despite these challenges, we identified major outcomes and considered their susceptibility to changes in exam format. While the data were limited, the studies were generally of good quality for the questions addressed and we did not identify any systematic differences in the use of OBE vs. CBE by the field studied (e.g., medical education vs. education vs. social sciences) or level of content (e.g., graduate vs. undergraduate student). Prior to the examination, findings were equivocal, but if there is an impact it favors the argument that people prepare more extensively for CBEs than for OBEs. This may be related to the finding that students anticipate lessened anxiety with OBE even though that anticipation does not appear to translate to actual experiences of lessened anxiety. During the examination, it appears to take examinees longer to complete OBEs relative to CBEs, which could either influence the test's reliability, if testing time is kept constant, or influence the length of time that must be offered candidates to complete the exam. Studies addressing examination performance favored CBEs, particularly when preparation for CBEs was greater than for OBEs. With respect to post-examination outcomes, we did not find robust evidence for differences in testing effects or public perception. That said, one might imagine concerned patients who express the sentiment "how can you be an expert if you need to look things up on the internet?"[41]

Putting all of this together, the decision regarding which type of examination to use might need to be based less on learning and performance outcomes and more on logistical limitations as well as the desire to authentically represent what individuals are expected to do in practice. Given that we found evidence of the testing effect under both conditions and the fact that participants' perceptions of testing effects run counter to empirical findings, a related question is how often an individual should be examined to maximize the testing effects. While not the subject of this review, data from this field support increasing the frequency of examinations to improve learning; one examination each decade, as is practiced with many certifying bodies, does not maximize the potential impact of testing effects. A further exploration of contemporary learning theories might provide a useful lens for understanding and interpreting how environmental factors and personal factors interact in dynamic ways to impact examination performance and the pedagogical value of testing.[43]

In terms of feasibility, it's challenging for high stakes testing organizations that value test security to allow access to the entire internet. [44] At the same time, choosing a limited number of Web-based external resources erodes the authenticity of the experience, could disadvantage examinees who are less familiar with the chosen tools, and has the potential to impact the fairness of the process if technical difficulties arise during an examination. Additional feasibility issues include the cost of allowing access to Web-based resources and the greater amount of time required to achieve the same reliability with OBE relative to CBE. Issues such as cost and fairness have not been addressed in prior investigations and represent some next steps for continued research.

In terms of authenticity, it must first be noted that the studies conducted to date have rarely looked at "high-stakes" assessment. While there is good reason to argue that an important skill to maintain is a physician's ability to find information and to not barge ahead if she is uncertain, there can be a perception that OBEs are easier than CBEs. Indeed, although studies are lacking, an excerpt from the American Board of Ophthalmology, regarding changes to their recertification examination, captures the sentiment of many:

> "The decision to change from an open-book, take-home examination to a closed-book, computerized proctored examination was based primarily on the recognition of the value of the certificate within the public domain … state medical licensing boards are increasingly asking for a proctored examination. It is of utmost importance to assure the public of the rigor and validity of our certificates. From this standpoint, changing to an OBE poses unintended risks that are not easily mitigated without a wealth of data and advocacy to the contrary." (http://abop.org/faqs/maintenance-of-certification/#intent)

Given our study objectives, we identified several limitations in the existing literature. First, a minority of studies reported that learners had significant prior experience with OBEs.

Providing learner training and making OBEs more prevalent could greatly alter perceptions of OBEs. Second, very few of the investigations reviewed included electronic resources (e.g., the internet) as a parameter for OBEs as many studies were conducted before the internet was widely used. With today's widespread use of informatics and the internet in medicine, we suspect that enabling the use of the internet in an examination might facilitate knowledge application. Third, few investigations have involved practicing physicians. Such investigations are needed to enable deeper understanding of the public's view of using OBEs for certification and/or recertification purposes because the practice of medicine entails a promise of expertise and public trust. Fourth, the majority of studies were conducted within a single institution, which limits their generalizability and a minority of studies included a conceptual and/or theoretical framework, which can make interpretation difficult.

As the volume of medical knowledge is rapidly expanding, education and assessment cannot be designed such that graduation will ensure sufficient knowledge for future independent practice. Instead, education and assessment will have to instill within trainees the motivation and learning strategies needed to become lifelong, self-regulated, learners. Traditional behaviorist approaches to testing cannot fulfill this requirement simply because the behaviour desired usually fades away when the reinforcement (in this case, the testing) is discontinued. Following logically from this argument, we wish to point out that the outcomes used in the studies reviewed here did not capture elements deemed to be essential by the current assessment-for-learning discourse. For example, no study looked at whether the incorporation of CBEs or OBEs yielded differences in reflection-on-action or receptivity to feedback when examinees formulated learning goals or were presented with external data. We believe that pursuing these avenues of research would be particularly informative for future developments in assessment.

We believe that both OBEs and CBEs can contribute to an assessment program in part due to their complementary pros and cons. OBEs should not be thought of as an alternative to CBEs, but their value may be in expanding beyond what is measured by CBEs. For example, exploring the "skill" or efficacy of looking up information on the internet seems unlikely to be accomplished through a CBE. A strategy for testing agencies, therefore, could be coupling OBEs with standard CBEs to explore these different "skills" without losing the reliability that comes from asking a larger number of questions in a short period of time. Furthermore, testing effects are not currently being optimized given the infrequency of examinations and the research findings that suggest the magnitude of testing effects increases with repeated testing. A series of mandatory, but ungraded (i.e., lower stakes) OBEs might help to improve aspects of these processes, such as capitalizing on the testing effect while not dramatically increasing learner anxiety. One examination each decade, as is practiced by many certifying bodies, is unlikely to maximize the educational impact of testing or induce the habits of continuous developmental efforts that the profession seeks to encourage. Further, by including some OBE items, the opportunity for improving authenticity and reducing the stigma that may align with the need to look things up could be leveraged. Any such benefits, however, may only be realized by recognizing the need identified by several authors that training in OBE is necessary for both students and examiners. Expectations need to be established regarding the types of questions used, the need for preparation, and how much time can be used to search for information that is not readily known.

Conclusion

Given the data that have been collected to date, there does not appear to be sufficient evidence for relying on OBE or a CBE formats. As such, we believe that a combined approach could become a more significant part of testing programs including physician certification or recertification.

Acknowledgements: none

Disclaimers: The views expressed herein are those of the authors and do not necessarily reflect the views of the Department of Defense or other federal agencies.

Ethical approval: Not applicable

Funding: None

Previous presentations: None

References

1. Feldhusen JF. An evaluation of college students' reactions to open book examinations. Educ Psychol Meas. 1961;XXI:637-46.
2. Theophilides C, Dionysiou O. The major functions of the open-book examination at the university level: a factor analytic study. Studies in Educational Evaluation. 1996;22:157-70.
3. Heijne-Penninga M, Kuks JB, Hofman WHA, Cohen-Schotanus J. Influence of open- and closed-book tests on medial students' leaning approaches. Med Educ. 2008;42:967-74.
4. Adair JG and Vohra N. The explosion of knowledge, references, and citations: Psychology's unique response to a crisis. Am Psychol. 2003;58:15-23.
5. Ramsey PG, Carline JD, Inui TS, Larson EB, LoGerfo JP, Wenrich MD. Predictive validity of certification by the American Board of Internal Medicine. Ann Intern Med.1989;110:719-26.
6. Tamblyn R, Abrahamowicz M, Dauphinee D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. JAMA. 2007;298:993-1001.
7. Young JQ, van Merrienboer J, Durning SJ, and ten Cate O. Cognitive Load Theory: implications for medical education: AMEE Guide No. 86.Med Teach. 2014;36:371-84.
8. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med. 2009;151:264-9.
9. Cook DA, West CP. Conducting systematic reviews in medical education: a stepwise approach. Med Educ. 2012;46:943-52.
10. Agarwal PK, Roediger HL. Expectancy of an open-book test decreases performance on a delayed closed-book test. Memory. 2011;19:836-52.
11. Boniface D. Candidates' use of notes and textbooks during an open-book examination. Educ Res. 1985;27:201-9.
12. Moore R, Jensen PA. Do open-book exams impede long-term learning in introductory biology course? Journal of College Science Teaching. 2007;36:46-9.
13. Gharib A, Phillips W, Mathew N. Cheat sheet or open-book? A comparison of the effects of exam types on performance, retentions, and anxiety. Psychology Research. 2012;2:469-78.
14. Betts LR, Elder TJ, Hartley J, Trueman M. Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? Assessment & Evaluation in Higher Education. 2009;34:1-15.
15. Heijne-Penninga M, Kuks JBM, Hofman WHA, Cohen-Schotanus J. Influences of deep learning, need for cognition and preparation time on open- and closed-book test performance. Med Educ. 2010;44:884-91.
16. Broyles IL, Cyr PR, Korsen N. Open book tests: assessment of academic learning in clerkships. Med Teach. 2005;27:456-62.
17. Dale VHM, Wieland B, Pirkelbauer B, Nevel A. Value and benefits of open-book examinations as assessment for deep learning in a post-graduate animal health course. Curriculum and Assessment. 2009;36:403-10.
18. Carrier LM. College students' choices of study strategies. Percept Motor Skill. 2003;96:54-6.
19. Schutz PA & Pekrun R. Emotion in education. 2007. Elsevier.

20. McConnell MM, Eva KW. The role of emotion in the learning and transfer of clinical skills and knowledge. Acad Med. 2012;87:1316-22.
21. Ben-Chaim D, Zoller U. Examination-type preferences of secondary school students and their teachers in the science disciplines. Instr Sci. 1997;25:347-67.
22. Dickson KL, Miller MD. Authorized crib cards do not improve exam performance. Teach Psychol. 2005;32:230-3.
23. Baillie C, Toohey S. The 'power test': its impact on student learning in a materials science course for engineering. Assessment and Evaluation in Higher Education. 1997;22:33-48.
24. Eilertsen TV, Valdermo O. Open-book assessment: a contribution to improved learning? Studies in Educational Evaluation. 2000;26:91-103.
25. Rakes GC. The effects of open book testing on student performance in online learning environments. Accessed online 7/2013.
26. Schumacher CF, Butzin DW, Finberg L, Burg FD. The effect of open- vs. closed-book testing on performance on a multiple-choice examination in pediatrics. Pediatrics. 1978;61:256-61.
27. Ioannidou MK. Testing and life-long learning: open-book and closed-book examination in a university course. Studies in Educational Evaluation. 1997;23:131-9.
28. Weber LJ, McBee JK, Krebs JE. Take home tests: an experimental study. Res High Educ. 1983;18:473-83.
29. Block RM. A discussion of the effect of open-book and closed-book exams on student achievement in an introductory statistics course. PRIMUS. 2012;22:228-38.
30. Wachsman Y. Should cheat sheets be used as study aides in economics tests? Economics Billetin. 2002;1:1:1-11.
31. Dickson KL, Bauer JJ. Do students learn course material during crib sheet construction? Teach Psychol. 2008;35:117-20.
32. Norman GR, Smith EKM, Powles AC, Rooney PJ, Henry NL, Dodd PE. Factors underlying performance on written tests of knowledge. Med Educ.1987;21:297-304.
33. Schuwirth LWT, Van der Vleuten CPM, Donkers HHLM. A closer look at cueing effects in multiple-choice questions. Med Educ. 1996;30:44-9.
34. Ward WC. A comparison of free-response and multiple-choice forms of verbal aptitude tests. Appl Psych Meas.1982;6:1-11.
35. Brightwell R, Daniel J, Stewart A. Evaluation: is an open book examination easier? Bioscience Education. 2004;3
36. Roediger HL and Butler AC. The critical role of retrieval practice in long-term retention. Trends Cogn Sci. 2011 Jan;15(1):20-7
37. Larsen D, Butler A, Roediger H. Test-enhanced learning in medical education. Med Educ. 2008;42:959-66.
38. Agarwal PK, Karpicke JD, Kang SK, Roediger HL, McDermott KB. Examining the testing effect with open- and closed-book test. Appl Cognitive Psych. 2008;22:861-76.
39. Pauker JD. Effect of open book examinations on test performance in an undergraduate child psychology course. Teach Psychol. 1974;1:71-3.
40. Heijne-Penninga M, Kuks JBM, Hofman WHA, Muijtjens AMM, Cohen-Schotanus J. Influence of PBL with open-book tests on knowledge retention measured with progress tests. Adv in Health Sci Educ. 2013; 18:485-95.
41. KevinMD.com (accessed 7/2013)

42. Theophilides C, Koutselini M. Educational research and evaluation: an international journal on theory and practice. 2000;6:379-93.
43. Pekrun R. The Control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. Educ Psychol Rev. 2006; 18:315-341.
44. Lipner RS, Lucey CR. Putting the secure examination to the test. JAMA. 2010; 304(12):1379-1380.
45. Heijne-Penninga M, Kuks JBM, Hofman WHA, Cohen-Schotanus J. Directing students to profound open-book test preparation: the relationship between deep learning and open-book test time. Med Teach. 2011;33:e16-21.
46. Jehu D, Picton CJ, Futcher S. The use of notes in examinations. British Journal of Educaitonal Psychology. 1970; 40(3):335-37.
47. Kalish RA. An experimental evaluation of the open book examination. J Educ Psychol. 1958;49:200-4.
48. Krarup N, Naeraa N, Olsen C. Open-book tests in a university course. High Educ. 1974;3:157-64.
49. Ronald SL, Lola A. The relationship between testing condition and student test scores. Journal of Instructional Psychology. 2004;31:304-13.
50. Shine S, Kiravu C, Astley J. In defence of open-book engineering degree examinations. International Journal of Mechanical Engineering Education. 2004;32:197-211.
51. Whitley BE. Does ‘cheating’ help? The effect of using authorized crib notes during examinations. College Student Journal. 1996;30:489-93.

Table 1: Study quality

	Mean	Min	Max	# (%) of studies with a rating of 4 or 5
Trustworthiness	4.0	2	5	27 (73%)
Rigor	3.8	2	5	20 (54%)
Implementation	3.8	2	5	21 (57%)
Analysis	3.7	2	5	18 (49%)

Note. The response scale ranged from 1 = strongly disagree to 5 = strongly agree

Appendix 1: Coding results for included articles

Exam Preparation

	Authors	# of participants	Participants' prior experience with OBE	Level of the participants'investigated	Country	Outcomes	How measured?	Key findings
10	Agarwal & Roediger	108	Not stated	Undergraduate students	US	Study time	Timed by computer	Participants studied for less time when expecting an OBE.
23	Baillie & Toohey	Not clear	Had orientation in classroom	Undergraduate students	Australia	The effect of teaching preparation for OBE	Interview	The change of the teaching method of the course has helped the students to be prepared for OBE.
14	Betts, Elder, Hartley, & Trueman	116	Implied--OBE seems to be part of the curriculum	Undergraduate students	UK	Preparation efforts	Self-reported questionnaire	Students reported preparing for the OBE and CBE to a similar extent. However, there was a marginally significant interaction between examination condition and gender. Females prepared more for CBE than for the OBE. There was no significant difference in the amount that males prepared for the OBE and CBE.
29	Block	938	Not stated	Undergraduate students	US	Preparation efforts	Instructors' observation	The students came to the OBE not fully prepared and expected to find the needed answers in the book. CBE with notecards led to better preparation.
11	Boniface	30	Not stated	Undergraduate students	UK	Preparation efforts	Self-reported questionnaire	Students thought they would do more preparation for CBE than OBE.
16	Broyles, Cyr, & Korsen	174 (only 18 were interviewed after the exams)	Not stated	Medical students (MS-3)	US	Preparation tactics	Interview	More than 60% of the students did not change study tactics for OBE (those who did change studied less); 25% waited until last week of 4 week rotation to study.
18	Carrier	58	Not stated	College students	US	Preparation tactics	Self-reported questionnaire	For CBE, surface studying was mostly done, but deep approach (studying lecture notes,

								making chapter notes, highlighting and/or underlining, and coming to office hours) correlated with high scores. For OBE similar proportions of students used surface and deep approach, nothing correlated with examination performance.
17	Dale, Wieland, Pirkelbauel, & Nevel	14	2/14 had prior experience with OBE	Graduate, professional education	UK	Preparation tactics	Self-reported questionnaire and interview	No statistically significant correlation was found between perceptions of different assessment methods (OBEs, essays, SAQs, MCQs) and deep vs surface learning approach scores. Most students felt prepared for the OBEs.
31	Dickson & Bauer	53	Not stated	Undergraduate students	US	Preparation for cheat-sheet exam (partial OBE)	Interview	Preparing crib sheets does not enhance learning, but use of crib sheets enhanced test performance.
24	Eirlertsen &Valdermo	350	Not stated	High school students	Norway	Training for OBE	Action research: survey, interview, class observation	The problem of not preparing thoroughly for OBE declines when OBE is applied over a period of time and as part of a broader approach aimed at strengthening students' understanding of learning and knowledge. Many students need to learn at an early stage that they have to be equally well if not better prepared for OBE vs. CBE.
1	Feldhusen	90	Not stated	Undergraduate students	US	Preparation tactics	Self-reported questionnaire	The students felt that it is useless to "cram" for exams and that the open-book exam reduces memorization of factual material in preparation for OBE.
13	Gharib, Phillips, & Mathew	387	Not stated	Undergraduate students	US	Preparation time	Self-reported questionnaire	The actual reported study time for OBE vs CBE did not differ, although the students believed that they would study most for CBE.
15	Heijne-Penninga, Kuks, Hofman, Cohen-Shotanus	239	Yes	Medical students (MS-2)	Netherlands	Preparation time	Self-reported questionnaire	The students reported more preparation time for CBE than OBE.
45	Heijne-Penninga, Kuks, Hofman, Cohen-Shotanus	663	Yes	Medical students (MS-2 & MS-3)	Netherlands	Preparation time and tactics	Self-reported questionnaire	Second and third year college students differed significantly in OBE prep time (t(662)=2.25, p<0.01). Third year students spent less prep time and prepared more deeply.
3	Heijne-	570				Exam	Deep	Counter to the hypothesis and prevailing

	Penninga, Kuks, Hofman, Cohen-Shotanus		Yes	Medical students (MS-2 & MS-3)	Netherlands	preparation and testing effects	Information Processing questionnaire	wisdom, CBE preparation and not OBE preparation was associated with deep learning.
12	Moore & Jensen	351	Not stated	Undergraduate students	US	Preparation efforts; Class attendance	Informal discussion and class observation	Students seemed to prepare for OBE in the same way that they prepared for CBE. However, some students' class attendance dropped significantly when the upcoming exam was an OBE.
25	Rakes	49	Participants took a practice test to familiarize with OBE.	Graduate students	US	Training for OBE	Test performance; training intervention	In online learning environment, the administration of OBE may adversely affect students' exam performance because they do not necessarily understand the requirements of OBE. Training may mitigate the inclination not to study for OBE.
2	Theophilides & Dionysiou	173	Not stated	Undergraduate students	Greece	Preparation tactics	Self-reported questionnaire	The perceived functions of OBE include a factor of exam preparation. The items within this factor are: When preparing for exam, Compares and contrasts information obtained; studies various resources; interrelates information acquired and conclusions drawn; reconstructs course content and integrates knowledge gained; practices study skills (note taking, textbook studying).
42	Theophilides, Koutselini	201/276 respondents to survey	Not stated	Undergraduate students	Greece	Preparation tactics	Self-reported questionnaire	When students expect an OBE, they are more attentive throughout the semester and engage in more study activities that promote deep learning of the course subject matter.
30	Wachsman	299	Not stated	Undergraduate students	US	Preparation for cheat-sheet exam (partial OBE)	Test performance	The combined effect of preparing and using cheat sheets is positively associated with students' test performance, even when controlling for preparation time.
				Test Anxiety				
23	Baillie & Toohey	55	Had orientation in classroom	Undergraduate students	Australia	Anxiety	Focus groups, interviews	Anxiety associated with taking OBE was not reduced as much as investigators expected. A large portion of students (45%) were just as stressed with OBE as CBE.

21	Ben-Chiam & Zoller	236	Not stated	High school students	Israel	Preferred exam types, anxiety, stress	Self-reported questionnaire, interviews	OBEs, particularly in the form of take-home exams without strict time limits, reduce the level of student anxiety.
14	Betts, Elder, Hartley, & Trueman	116	Implied-- OBE seems to be part of the curriculum	Undergraduate students	UK	Anxiety	Self-reported questionnaire	Students reported feeling more anxious when a correction for guess was included in a CBE than when the correction was used in an OBE.
16	Broyles, Cyr, & Korsen	18	Not stated	Medical students (MS-3)	US	Anxiety, stress reduction	Interviews	Most students (80%) noted that they were less anxious and less stressed when taking OBEs.
17	Dale, Wieland, Pirkelbauel, & Nevel	14	2/14 had prior experience with OBE	Graduate, professional education	UK	Enjoyment, stress	Self-reported questionnaire, interviews	Limited qualitative results suggest that OBEs were thought to be less stressful than traditional CBEs.
31	Dickson & Bauer	53	Not stated	Undergraduate students	US	Anxiety	Self-reported questionnaire	The vast majority of students (80%) reported that making a crib sheet reduced their stress during the exam.
22	Dickson & Miller	52	Most students had experience with OBE	Undergraduate students	US	Anxiety	Self-reported questionnaire	Despite high expectations by 79% of students that using a “cheat sheet” would lower anxiety during a test, only 41% reported that using the cheat sheet during the test actually lowered anxiety.
24	Eirlertsen &Valdermo	350	Not stated	High school students	Norway	Anxiety	Interviews, self-reported questionnaires	OBE may reduce anxiety for some, but at the same time, some aspects of OBE, such as unfamiliar assignments or shortage of time to make use of the available materials, can also be anxiety-provoking.
1	Feldhusen	90	Not stated	Undergraduate students	US	Worry, tension	Self-reported questionnaire	When comparing OBE and CBE), students reported less worry and tension with OBE.
13	Gharib, Phillips, & Mathew	396	Not stated	Undergraduate students	US	Anxiety	Self-reported questionnaire	Students reported higher anxiety in a cheat sheet exam relative to OBE (CBE anxiety not examined). Test anxiety measured right before the exam (cheat sheet and OBE) was negatively correlated with scores on the exams.
46	Jehu, Picton, & Futcher	29	Not stated	Undergraduate students	UK	Anxiety	Self-reported questionnaire	The availability of notes (cheat sheet) reduced anxiety *during* an exam but did not reduce anxiety *before* the exam.

2	Theophilides & Dionysiou	173	Not stated	Undergraduate students	Greece	Anxiety	Self-reported questionnaire	Of the five factors that comprised a questionnaire about the characteristics of OBEs, none of the factors were related to exam anxiety. Further, the five factors did not vary by students' exam anxiety level or expected graduation grade.
42	Theophilides & Koutselini	181	Not stated	Undergraduate students	Greece	Perceived differences in exam types	Self-reported questionnaire	Limited qualitative results suggest that the OBE alternative reduced exam tension and stress. Participants stated that they approached OBEs with greater optimism and worked out their answers in a more relaxed way.
28	Weber, McBee, & Krebs	64	Not stated	Undergraduate students	US	Anxiety	Self-reported questionnaire	Students believed that OBEs and take-home exams caused much less stress than CBEs.
					Exam Performance			
38	Agarwal, Karpicke, et al	36 (experiment 1); 48 (experiment 2)	Not stated	Undergraduate students	US	Immediate and Delayed performance (I and D respectively)	MCQs	Initial: OBE superior performance than CBE; both superior to studying alone. Delayed: OBE and CBE effect equivalent. CBE with feedback resulted in greater final performance than without it.
10	Agarwal & Roediger	108	Not stated	Undergraduate students	US	I and D	Short answer	OBE lead to better initial performance but similar results after 2 day delay (delayed test). Participants studied for less time when expecting an OBE and correspondingly performed worse (10% improved performance with expecting CBE over OBE).
14	Betts, Elder, Hartley, & Trueman	116	Implied--OBE seems to be part of the curriculum	Undergraduate students	UK	I and Correction for guessing	MCQ	Scored higher and left fewer questions unanswered when no correction for guessing. Students favor the use of correction for guessing for OBE but not CBE.
29	Block	938	Not stated	Undergraduate students	US	D	MCQ + short answer	Better delayed performance when prior exam CBE vs OBE format.
11	Boniface	30	Not stated	Undergraduate students	UK	D	MCQ+ short answer	There were large negative correlations between the amount of time devoted to consulting notes and texts and exam score (R= -.44 for both, .-.39 for notes, -.13 for texts).

								The correlations between the amount of time devoted looking things up and scores on previous assessments were also negative (-.33 for CBE)
35	Brightwell, Daniel, & Stewart	196	Not stated	Undergraduate students	Australia	D	On-line MCQ examination	No differences in mean, minimum or maximum scores
16	Broyles, Cyr, & Korsen	174	Not stated	Medical students (MS-3)	US	I	MCQ	OBE 88.2 vs. CBE 84 (ANOVA p=.03). Small statistically significant improvement with OBE vs CBE.
18	Carrier	58	Not stated	College students	US	D and study preparation approach (surface or deep)	Self-reported questionnaire + MCQ, exam	For CBE, deep approach (studying lecture notes, making chapter notes, highlighting and/or underlining, and coming to office hours) correlated with higher performance. For OBE deep or surface approach did not correlate with exam performance. Similar proportions of students used surface and deep approach for OBE and CBE
31	Dickson & Bauer	53	Not stated	Undergraduate students	US	Cheat-sheet exam (partial OBE) vs. CBE, I	Interview +MCQ	Preparing crib sheets does not enhance performance, but use of crib sheets enhanced test performance.
22	Dickson & Miller	54	Most students had experience with OBE	Undergraduate students	US	Cheat-sheet exam (partial OBE) vs. CBE, I	Questionnaire + MCQ	Using crib card did not lead to higher exam scores on either low or high order MCQs despite students' belief that it would
13	Gharib, Phillips, & Mathew	387	Not stated	Undergraduate students	US	OBE, CBE, partial OBE (cheat sheet), I	MCQ	OBE and partial OBE> CBE score.
15	Heijne-Penninga, Kuks, Hofman, Cohen-Shotanus	239	Yes	Medical students (MS-2)	Netherlands	D+ preparation time and need for cognition	MCQ	Need for cognition or the tendency of an individual to engage in effortful cognitive activities and to enjoy thinking positively influenced both OBE and CBE performance; deep learning and time for preparation did not correlate with performance on either OBE or CBE.
45	Heijne-Penninga, Kuks, Hofman, Cohen-Shotanus	663	Yes	Medical students (MS-2 & MS-3)	Netherlands	D	Self-reported questionnaire + MCQ	No difference in general; performed worse on CBE when tested on material previously received in OBE (or assumed that it would be tested by OBE)
3	Heijne-	570				I + learning	Deep	Students scored significantly higher when

	Penninga, Kuks, Hofman, Cohen-Shotanus		Yes	Medical students (MS-2 & MS-3)	Netherlands	approach survey and	information processing survey + MCQ	preparing for CBE. Counter to the hypothesis and prevailing wisdom, CBE preparation and not OBE preparation was associated with deep learning.
40	Heijne-Penninga, Kuks, Hofman, Muitijens, Cohen-Shotanus	1648	Students at the PBLob had prior experience	UME -- 5th and 6th year students	Netherlands	D + PBL vs traditional learning (TL) approach	Not stated	OBE and CBE for PBL students higher than TL students.
27	Ioannidou	72	Not stated	Undergraduate students	Cyprus	D	MCQ + essay	No significant difference in scores on OBE vs CBE
46	Jehu, Picton, & Futcher	29	Not stated	Undergraduate students	UK	Partial OBE (note sheet) vs. CBE	Essay	No significant difference in scores
12	Moore & Jensen	351	Not stated	Undergraduate students	US	I + D	MCQ	Average grades on Exam 1 in the control and experimental sections were not significantly different. However, grades on Exams 2 and 3 were significantly higher for OBE. On the final exam, grades were significantly higher for CBE. Some students' class attendance dropped significantly when the upcoming exam was an OBE.
47	Kalish	158	Not stated	Undergraduate students	US	I + D	MCQ	The results have indicated that, although under the conditions of this experiment the group average scores are not affected by OBE vs CBE, the two types of examinations appear to measure different abilities (based on correlation between test 1 and test 2 being higher when both tests were CBE relative to when second was OBE).
48	Krarup et al	108	Yes	Medical students (6th term)	Denmark	I + D	MCQ	No differences in scores overall; recall items (15%) showed higher OBE performance
39	Pauker	96	Not stated	Undergraduate students	Canada	I +D	MCQ	No difference in scores by examination format
25	Rakes	49	Participants took a practice test to familiarize with OBE.	Graduate students	US	D+ additional arm of OBE training	MCQ	CBE performance superior to OBE; OBE scores improved with explicit training. In online learning environment, the administration of OBE may adversely affect students' exam performance because they do not necessarily understand the requirements of OBE. Training may mitigate the inclination

								not to study for OBE. Partial OBE helped students who received C and D grades on prior CBE.
26	Schumacher et al	196 students, 96 practicing pediatricians	Not stated	Practice group (practicing pediatricians) and student group (3rd and 4th year medical students)	US	I	MCQ	For both tests, practicing MDs performed better. Medical students did better on OBE than CBE; practicing MDs performance on OBE and CBE not different.
50	Shine	131	Not stated	Undergraduate students	Botswana, New Zealand and UK	I + D	MCQ + essay + short answer	Averages of test performance same or lower on OBE than CBE. OBE takes examiners more thought and skill in designing tests.
30	Wachsman	299	Not stated	Undergraduate students	US	I + D (cheat sheet vs CBE)	MCQ	The combined effect of preparing and using cheat sheets is positively associated with students' test performance, even when controlling for preparation time.
28	Weber, McBee, & Krebs	64	Not stated	Undergraduate students	US	I (Take home test vs OBE vs CBE)	MCQ	No significant differences existed between take home (64.9%) and open book (61.5%) exams or between open book and closed book exams (57.9%).
51	Whitley	136	Not stated	Undergraduate students	US	Partial OBE (crib sheet) vs. CBE	MCQ and short answer	Equivocal effect: using notes improved performance in one of two examination sessions
				Testing Effects				
47	Kalish	158	Not stated	Undergraduate students	US	Self-perceived judgment of learning	Single item question dealing with "degree of help" with OBE format	No relationship between perceived degree of assistance with OBE format and performance.
1	Feldhusen	90	Not stated	Undergraduate students	US	Self-perceived judgment of learning	13-item questionnaire administered at end of semester	Students felt that OBE were superior to CBE with promoting learning during testing.
24	Eirlertsen &		Not stated	High school students	Norway	Attitudes towards OBE	Mixed methods with	OBE format helpful in getting the students and teachers to understand the

	Valdermo	350					questionnaires and focus group interviews	nature of knowledge and process of learning.
31	Dickson & Bauer	53	Not stated	Undergraduate students	US	Self-report measure of helpfulness of crib sheet with respect to perceived learning	End of semester questionnaire	91.8% students felt using a “crib sheet” was helpful for learning. However, crib sheet preparation not associated with improved exam performance (when crib sheet not used).
17	Dale, Wieland, Pirkelbauel, & Nevel	14	2/14 had prior experience with OBE	Graduate, professional education	UK	Deep vs. surface learning	Questionnaire including Approaches and Study Skills Inventory for Students, interviews	A small association was observed between a reported preference for a Deep learning approach and favorable responses regarding OBE.
10	Agarwal & Roediger	108	Not stated	Undergraduate students	US	Self-perceived judgment of learning	0-100% self-perception of how well a passage would be remembered in 1 week	Students erroneously predicted that studying (rather than taking a test) would lead to better retention.
2	Theophilides & Dionysiou	173	Not stated	Undergraduate students	Greece	Deep learning	38 statements with 5 point Likert scale	OBE perceived by students to have five advantages including allowing students to creatively use the knowledge they gained in the course and encouraging students to apply deep approaches to studying.
22	Dickson & Miller	54	Most students had experience with OBE	Undergraduate students	US	Self-perceived judgment of learning	Pre and post-semester questionnaire assessing perceived effect on exam performance and learning (post only)	Perceived learning benefits from crib sheets mixed. Students’ perceived effect on exam performance was more positive before the semester than at the end.

Testing Effects (also see appendix 2 that explicitly used testing effects as a theoretical framework for the investigation

Appendix 2: Selected articles that explicitly used the testing effect as a theoretical framework.

	Author(s)	Theoretical framework	Participants	Format	Quantitative Findings	Key finding(s)
38A *	Agarawal et al	Testing effect	36 undergraduate psychology students	One week delayed CBE (short answer responses to passages)	CBE M 0.59 CBE w/ feedback M 0.68* OBE M 0.65* *Significant difference (M1⁄4.68) was greater than performance in the closed-book test without feedback condition (M 1⁄4 .59), t(35) 1⁄4 2.58, d 1⁄4 .57, prep 1⁄4 .94	Testing effect demonstrated in all exam types. Taking OBE or CBE with feedback prior to final exam superior to CBE without feedback.
38B	Agarwal et al	Testing effect	48 undergraduate psychology students	One week delayed CBE (short answer responses to passages)	CBE M 0.55 CBE w/ feedback M 0.66* OBE M 0.66* *Significant difference	Similar to 1A, OBE or CBE with feedback prior to final exam superior to CBE without feedback.
10A	Agarwal et al	Testing effect, Barnett and Ceci's taxonomy of transfer, Bloom's taxonomy	72 undergraduate students (Dept of Psych subject pool)	2 day delay GRE style passages and MCQs	Study-only M 0.49 OBE M 0.63* CBE w/ feedback M 0.61* *Significant difference	Equivalent testing effect demonstrated for OBE and CBE (w/ feedback) compared with study only condition.
10B	Agarwal et al	Testing effect	108 undergraduate students (Dept of Psych subject pool)	2 day delay GRE style passages and MCQs	Delayed Fact CBE M 0.22 (.01) OBE M 0.20 (.01) Delayed Comprehension CBE M 0.63 (.01) OBE M 0.66 (.02) Delayed Transfer CBE M 0.40 (.02) OBE w/ feedback M 0.42 (.02)	Similar to 2A, replicated similar performance on 2 day delayed CBE regardless of prior OBE or CBE (this time, practice CBE without feedback).
31	Dickson, Bauer	Coding hypothesis (improved learning with crib sheet) and Dependency hypothesis (crib sheet changes study habits)	53 undergraduate psychology students	15 item CBE pretest MCQ exams prior to pOBE (partial OBE)	1st CB pretest M 55.7 (SD 15.8) 1st pOBE M 74.7 SD = 13.9 $p < .001$ 2nd CB pretest M 62.6 SD 16.8 2nd pOBE M 69.1 SD 17.8 $p < .01$	Improvements in performance seen in pOBE were not seen on closed-book pretests of identical questions calling improved learning through crib sheet creation into question.

13A	Gharib, Phillips, Mathew	Testing effect, partial OBE	297undergraduate, Intro to Pysch (only 2 of 5 sections took retention quizzes – 191 quizzes)	2 week, unannounced closed book 10 –item retention quiz, MCQ	OBE M 6.41 (1.94) pOBE M 6.44 (1.88) CBE M 6.38 (1.94)	No differences in testing effects demonstrated between OBE, pOBE or CBE formats in an undergraduate introduction to psychology class.
13B	Gharib, Phillips, Mathew	Testing effect, partial OBE	99 undergraduate, Intro to Stats (343 retention quizzes analyzed)	2 week, unannounced closed book 10-item retention quiz , MCQ	OBE M 6.18 (1.75) pOBE M 6.31 (1.83)	Similar to 20A, no differences in testing effects noted in undergraduate statistics class.
40	Heijne-Penninga, Kuks, Hofman, Muitijens, Cohen-Shotanus	Testing effect, information processing theory, knowledge organization, PBL curriculum	1648 medical students (Denmark)	200 CBE MCQ progress tests (4 tests annually, core and back-up knowledge)	OBE curriculum core knowledge was significantly higher vs CBE curriculum on 4 of 8 tests ($p < 0.0021$ one sided t test) (no difference with back-up knowledge)	A small benefit in “core knowledge” retention was observed in a cohort of medical students whose PBL curriculum included OBE assessment.
12	Jensen, Moore	Testing effect	351 undergrate Intro to Bio students (US)	70 MCQ CB final exam	Overall CBE M 74 + 4, OBE M 63+5 * Matched content (2nd and 3rd exams) CBE M 75 + 4 OBE M 57 + 7 * CBE M 74 + 3 OBE 61 + 6 *	Testing effect of CBE superior to OBE (further evidenced by Further, the content they scored lower on matched the content covered by the OBEs.
39	Pauker	Testing Effect	96 undergraduate students (Canada)	75 MCQ CBE final	OBE M 52.8 (SD 8.3) CBE M 54.8 (SD 5.8)	Testing effect identical except for lowest tertile of students who performed significantly better if all prior tests CBE.

**A and B refer to papers where there were more than one investigation reported in the manuscript*

Appendix 3: Article coding sheet

Reviewer Name:
Date:
1st reviewer □ 2nd reviewer □ 3rd reviewer □

Section A. Basic information

Reference Number: ______________________________
Authors (First, second, third author, et al):
____
____
Title:
____
____
Publication: ____________________________________Year____Vol_____Issue___Pages____
____
Search Method:
□ Electronic search
□ Hand search (e.g. bibliography of electronic search, recommended paper)

1. Citation type
   □ Original research paper/peer reviewed empirical study
   □ Review paper
   □ Commentary/opinion
   □ Conf. paper/proceedings
   □ Book
   □ Thesis
   □ Other (List__________________________)

2. Description of OBE

3. Description of CBE

4. Research question
   □ Stated (list____)
   □ Implied (list____)

5. Stated hypothesis
   □ Stated (list____)
   □ Implied (list____)
   □ Not clear

6. Hypothesis justification

   □ Stated (list____)

   □ Not clear

7. Conceptual framework

   □ Stated (list____)

   □ Not clear

Section B. Outcomes considered

□ Exam performance

□ Exam preparation

□ Logistics

□ Public perception

□ Other (list___________)

□ Anxiety or enjoyment

□ Psychometrics

□ Testing effects

□ Judgment of learning

Section C. Study context

1. Country ____________

2. Student level

   □ College students

   □ High school students

   □ Medical students

   □ Other post-college settings

   □ Others (list__________________________)

3. Test stakes

   □ High □ Medium □ Low

4. Type of material tested (list_________________________________)

5. Test format (list_____________________________)

6. Number of questions _______

7. Delay between learning and test

   □ Yes □ No

8. Prior experience with OBE

   □ Yes □ No

9. Incentive provided

   □ Yes   □ No

Section D. Study design

1. Description ____________________________
2. Within vs. Between subjects design

   □ Within   □ Between
3. Condition 1 ______________________________
4. Condition 2 ______________________________
5. Condition 3 ______________________________
6. Condition 4 ______________________________
7. Comparability of exam context

   □ Comparable   □ Not comparable
8. Comparability of exam format

   □ Comparable   □ Not comparable
9. Outcome 1 ______________________________
10. Outcome 2 ______________________________
11. Outcome 3 ______________________________
12. Outcome 4 ______________________________
13. Outcome 5 ______________________________
14. Scoring

    □ Objective   □ Subjective

Section E. Sample characteristics

1. Total N ______
2. OBE N ______
3. CBE N ______
4. Age ________
5. Other demographics noted ___________________

Section F. Findings (include statistics)

1. Outcome 1 ____

2. Outcome 2 ____

3. Outcome 3 ____

4. Outcome 4 ____

5. Outcome 5 ____

6. Main conclusion ____

Section G. Evaluation

1. Trustworthiness
   □5 □4 □3 □2 □1
2. Rigour
   □5 □4 □3 □2 □1
3. Implementation
   □5 □4 □3 □2 □1
4. Analysis
   □5 □4 □3 □2 □1

5. □Include in the review □Exclude from the review

6. Limitations ____

7. Additional references to check ______________________________________

8. Notes ____

Appendix 4: Specific search terms and strategy

Example search strategy:

MEDLINE (OVID) :

Search Strategy 1:
((computer or web or paper or internet) adj (aided or based) adj3 (assess$ or test$ or exam$ or curriculum or learning or evaluat$ or teaching)).ti,ab.
AND
(exp education, professional/ or ("professional education" or "medical education" or "nursing education" or "graduate education" or clerkship$ or residen& or student$).mp.)

Search Strategy 2:
((open or closed) adj1 (book or web) adj4 (assess$ or test$ or exam$)).ti,ab.

Searches OR together and limited to English Language

MEDLINE (Ovid)

MEDLINE (OVID) Search Strategy (June 7, 2013):
Notes: Retrieved 1340, 73 duplicate records = 1267 Total

Search Strategy 1:
((computer or web or paper or internet) adj (aided or based) adj3 (assess$ or test$ or exam$ or curriculum or learning or evaluat$ or teaching)).ti,ab.
AND
(exp education, professional/ or ("professional education" or "medical education" or "nursing education" or "graduate education" or clerkship$ or residen& or student$).mp.)

Search Strategy 2:
((open or closed) adj1 (book or web) adj4 (assess$ or test$ or exam$)).ti,ab.

Searches OR together and limited to English Language

ERIC Datatabase

ERIC Search (June 7, 2013):

Search Strategy 1: 46 records
All Fields: ("open book" OR "closed book" OR "open web") AND (assess* OR test* OR exam*)
Limits: Peer Review

Search Strategy 2: 557 records
Thesaurus Descriptors:"Tests"
AND
Keywords:computer OR web OR paper OR internet
AND
(Keywords:assess* OR test* OR exam* OR curriculum OR learning OR evaluat* OR teaching
AND Limits: Peer Review

Duplicates removed: 34
EMBASE

July 1 2013 (2,619 articles)

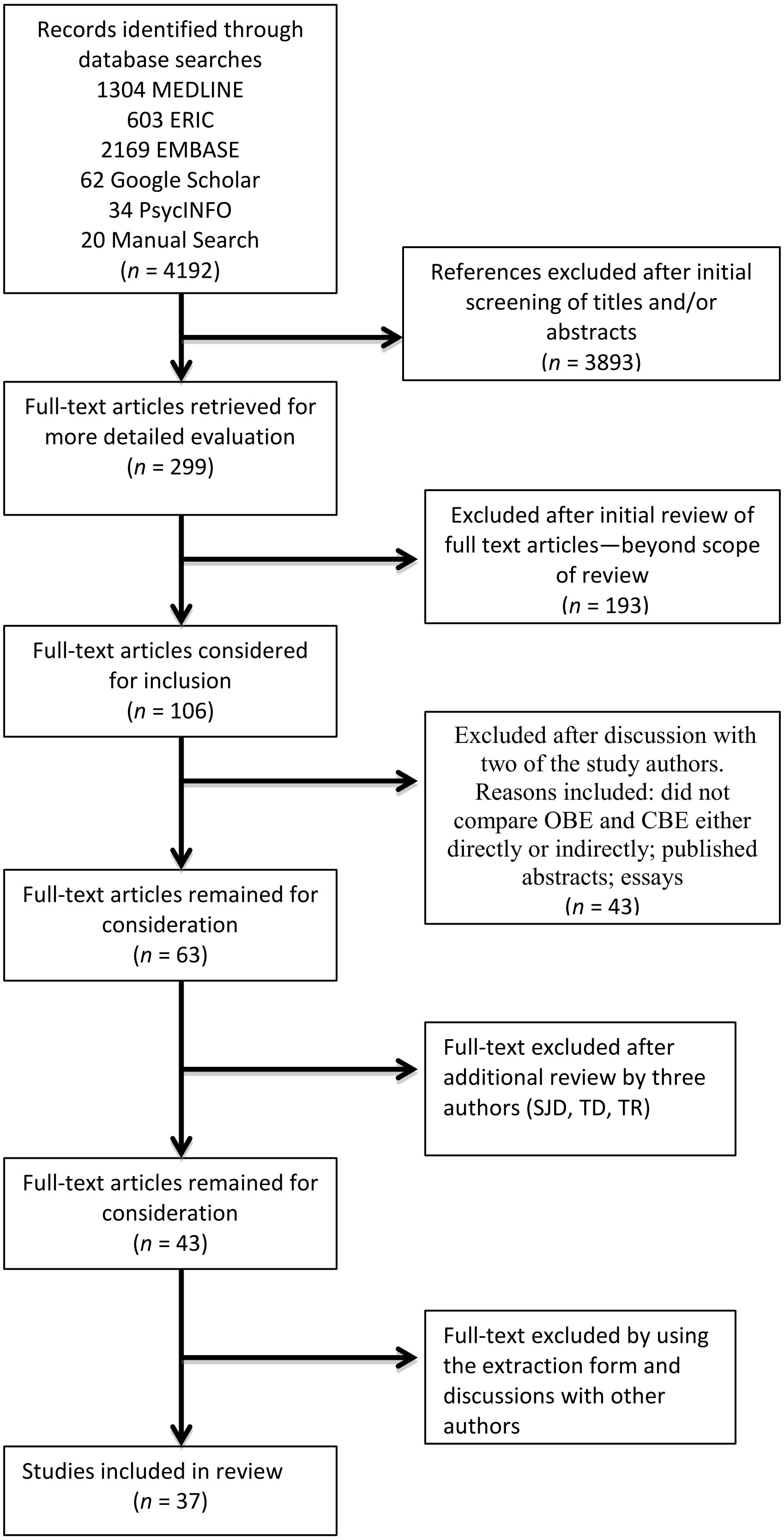
Search 1:
'computer aided':ab,ti OR 'computer based':ab,ti OR 'web based':ab,ti OR 'paper based':ab,ti OR 'internet based':ab,ti
AND
assess*:ab,ti OR test*:ab,ti OR exam*:ab,ti OR curriculum:ab,ti OR learning:ab,ti OR evaluat*:ab,ti OR teaching:ab,ti
AND
'medical student'/exp OR 'resident'/exp OR residency:ab,ti OR resident*:ab,ti OR 'clinical education'/exp OR clerkship* OR 'nursing student'/exp OR student*:ab,ti

Search 2:
'open book' OR 'open web' OR 'closed book' AND (assess* OR test* OR exam*)

Limited to English AND EMBASE

Duplicates removed: 8

Figure 1. Flow-chart of article selection


Records identified through database searches
1304 MEDLINE
603 ERIC
2169 EMBASE
62 Google Scholar
34 PsycINFO
20 Manual Search
(n = 4192)
References excluded after initial screening of titles and/or abstracts
(n = 3893)
Full-text articles retrieved for more detailed evaluation
(n = 299)
Excluded after initial review of full text articles—beyond scope of review
(n = 193)
Full-text articles considered for inclusion
(n = 106)
Excluded after discussion with two of the study authors. Reasons included: did not compare OBE and CBE either directly or indirectly; published abstracts; essays
(n = 43)
Full-text articles remained for consideration
(n = 63)
Full-text excluded after additional review by three authors (SJD, TD, TR)
Full-text articles remained for consideration
(n = 43)
Full-text excluded by using the extraction form and discussions with other authors
Studies included in review
(n = 37)

Appendix H:

Assessment Design & Performance in Practice

Potential Ideas for Changes to MOC Assessment

November 13, 2015

American Board of Internal Medicine

Evaluation, Research & Development

Executive Summary

When planning a certification assessment, important issues to consider are: (1) the purpose, (2) the design and (3) the policy space of the assessment. This white paper proposes six innovative solutions to redesign ABIM Maintenance of Certification (MOC) assessments by altering features of our current assessments with respect to these three issues. There are likely many more ideas that may evolve so please consider these only as a starting point for brainstorming new ideas.

The first idea **(“Diffusing the Bright Line”**) proposes changes to MOC policy that institutes a three-category classification system based on results of the secure exam. Candidates scoring within one standard error below the cut score would have an *indeterminate* decision instead of a forced fail decision. Candidates with an indeterminate decision would retain their certificate, but would be required to re-take the exam each year until successful.

A second idea **(“Using Prior Test Information to Determine Exam Cycles”**) proposes changes to MOC policy and design that are dictated by individual candidate performance on the secure examination. The specific requirements may take the form of altering the number of years between re-takes of the exam.

A third idea **(“Shorter Check-up Exams Using Prior Test Information”)** proposes changes to MOC policy and design so that rather than ask MOC candidates to sit for the secure exam every ten years, we instead ask them to complete short check-up exams every two or three years in a secure setting (either test center or remote proctoring). Most physicians would never need to take a longer exam. Only physicians for whom we are less confident in their ability would need to take a longer exam.

A fourth idea **(“Frequent Low-Stakes Assessments Resulting in a High Stakes Decision”)** proposes changes to MOC policy and design by administering small formative assessments every six months on demand (like jury duty) in lieu of the larger summative assessment every ten years. The small assessments would be administered at-home with remote proctoring and open-book and over 5 years the physician would need to maintain a certain score to keep one’s certification.

A fifth idea (**“Diagnosing Mastery of the Breadth of the Discipline”**) proposes changes to MOC policy by creating a powerful, Part II assessment that could diagnose strengths and weaknesses in preparation for the Part III pass-fail exam, still only once every 10 years. The web-based, adaptive diagnostic assessment would point out—through immediate feedback—areas of the exam blueprint where “brush-up” study would be advisable before sitting for the high-stakes Part III pass-fail exam. The Part II assessment would prepare physicians better so they have a greater chance of passing the high stakes exam.

A sixth idea (“**Practice Evaluation Using Population Health Measures**”) proposes changes to MOC policy in that physicians could opt for a performance assessment in lieu of passing the secure exam by demonstrating in their practice the capacity to maintain control of physician-sensitive measures of their patient population.

A Brief Note: The Parameters That Can Be Manipulated in Re-thinking Assessment

Regarding the **purpose** of the assessment, any of the proposed ideas might consider how we may change the purpose of our assessment.

We might first ask "*What* will it measure?" Will the assessment measure a single medical knowledge construct, as the current Part III exam does? Or should it instead measure more than one competency- such as communication skills or quality improvement - with each one reported on an independent scale? How many disciplines should be represented by exams in the ABIM program—the current number or does modern practice feature more specialized (i.e., "focused") sub-disciplines, which the ABIM program should reflect? For each construct, will the assessment report candidate ability level, as the current Part III exam does? Or will it instead report growth since last measurement? Or will it thoroughly identify gaps in learning (i.e., diagnostic assessment)?

Once the purpose is determined, the **design** of the assessment can unfold by asking "*How* might the constructs best be measured?" One way would be to directly observe the candidate's behaviors (performance-based assessment). Another might be to probe the candidate's mind (cognitive assessment) using multiple-choice questions or case-based computer simulations. Depending upon this decision, we can next turn to the format of the assessment. Issues such as the length of the test (fixed or variable), location (test center, educational institution or home), how faithfully the test should be aligned with the medical practice environment (i.e., degree of immersion) and what software tools will be accessible during the exam (e.g., reference books, calculators, web).

Finally, the **policy space** around the assessment should be considered. What should be the consequences based on the score? Should remediation be assigned to low scorers? Should the certificate length be shorter for low scorers? Related to consequences, what will be the stakes of the exam? Should certification depend upon passing the exam, as determined by the current "bright line" passing score (i.e., high stakes)? Or should there instead be some leniency afforded to those scoring in the "gray area" between the cut score and one standard error of measurement below it? And, once the assessment has been established, what should the time period between mandated re-takes of the assessments be?

Idea 1. Diffusing the Bright Line

Summary

The "bright line" idea reduces the chance for false negative errors by creating an indeterminacy region for candidates within one standard error of measurement below the passing score. After obtaining certification, diplomates only fail subsequent MOC exams if their exam scores are statistically significant below the cut score.

The "bright line" idea explicitly takes into account the concept of false negative pass/fail decisions and creates an "indeterminacy region" based on statistical confidence. Specifically, MOC candidates who are within one standard error below the cut score are not reported as failing the exam, even though their exam scores are below the passing score. The logic is that—after passing the certification exam—diplomates have already demonstrated the required knowledge, skills, and attitudes to receive ABIM certification; they are "in the club," so to speak. On subsequent MOC exams, diplomates are only reported as failing if we are statistically confident about the fail decision; we want to minimize errors in making false negative decisions, i.e., failing candidates who really should have passed the exam.

The diplomates would not initially lose their certification but would be required to pass the exam by the end of the following year thereby getting two more attempts at passing the exam before losing certification.

GOAL: To reduce the possibility for false positives and false negatives (i.e. kicking the diplomate "out of the club" when they really should be in it).

PROS

- Reduces the possibility for false negatives and makes us very confident that a diplomate who fails should not maintain their certification based on exam results.
- Very easy to implement and can be integrated into the MOC program right away.

CONS

- May increase false positives rate if diplomates were indeterminate for the first assessment and then subsequently passed the second assessment.
- Initial certification candidates may argue that their exam should allow for the same "safety net"

Idea 2. Using Prior Test Information to Determine Exam Cycles

Summary

Rather than standardizing the MOC requirements across all examinees, this idea permits those who performed well on the MOC examination to have lighter requirements in the immediate future. Those who performed poorly would be assigned the full—or an even heavier load—to demonstrate competence.

Variable Length Exam cycles

Historically examinees are asked to pass the ABIM secure exam once each decade. This policy implies that all physicians who pass the exam will maintain their skills at an acceptable level for at least ten years. This one-size-fits-all approach requires that high performing doctors retest more frequently than is probably necessary and allows lower performers to remain certified for longer than is reasonable.

An alternative approach would be to have the time period that an examinee must pass the exam dependent on examinees scores. The basic idea is that as exam scores increase our concern that the physician will fail the test in the future decreases. Therefore, a doctor performing well above the cut score could reasonably return less frequently then a doctor who has just barely passed the exam.

By estimating the probability that an examinee's score would have fallen below the cut score over time we can produce reasonable lifespans for the exam cycle. For example if the cut score is 300 the certificate lifespans could be as follows.

Fail	2 Years	5 Years	8 Years	11 Years	15 Years	
200	**300**	**400**	**500**	**600**	**700**	**800**

This approach would reward high-performers while ensuring that lower performers are acceptably maintaining their skills.

GOAL: To reinforce performance on the secure assessment through incentives for higher performance like reduction in exam frequency.

PROS

- Physicians would only be re-tested at a frequency that is related to their ability level, resulting in positive reinforcement and incentive for high performance on the exam.
- It encourages physicians to maintain their skills at an acceptable level for the secure exam by rewarding high performance with lower assessment frequency in the future.

CONS

- The frequency of testing in a secure setting would vary significantly by ability estimates, resulting in greater complexity in managing when diplomates must re-take the exam at the operational level.
- May increase anxiety for diplomates who consistently pass the exam, but at a lower ability level.

Idea 3. Shorter Check-up Exams using Test Information

Summary

Rather than ask MOC candidates to sit for the secure exam every ten years, we instead ask them to complete short check-up exams every two or three years. Most physicians would never need to take a longer exam but those we are less confident in their ability will.

.

Another approach to using test information to lower the burden on examinees would be to periodically check-up on examinees rather than requiring they pass the exam every ten years. These “Check-Up Exams” could be much shorter and lower stakes than the current MOC exam. The details of this are completely up for debate and would require some supporting analyses but one possible approach is presented below.

After physicians are certified they could take a 50-item Check-Up exam every two years. Based on this exam we would calculate the probability that the doctor’s true ability is below the cut score. If we are 85% confident that the examinee would pass the exam she/he would continue to be certified. If we are less than 85% confident she/he would remain certified but would be required to pass the complete longer exam the following year and the process would start over.

GOAL: To reduce anxiety for diplomates through lower-stakes assessments.

PROS

- Many physicians would not need to take a full-length, high-stakes assessment.
- The chances of an examinee with true ability below the cut score consistently passing a short exam would be very small.
- Check-up exams are lower, but not low, stakes. While no certificates would be revoked based on this exam, we would know sooner if a doctor’s skills degrade below acceptable levels.
- It encourages physicians to maintain their skills over time, rather than cramming for a test.

CONS

- The frequency of testing in a secure setting (either test center or remote proctoring from home) would increase.
- Some physicians would still experience the higher levels of anxiety when they must complete the longer exam following poor performance on shorter check-up exams.

Idea 4. Frequent Low-Stakes Assessments Resulting in a High-Stakes Decision

Summary

Rather than ask MOC candidates to sit for the secure exam every ten years, we instead ask them to complete (at home) a low-stakes, formative assessment every six months. They must achieve a certain level of mastery to retain their certification.

Replace the high-stakes secure exam with more frequent low stakes tests that would cumulatively fold into a high stakes score. This approach would require physicians to demonstrate competency while allowing them to track and realize when they need to elevate their steady state competency.

Implementation could, for instance, take the form of a low stakes assessment administered every six months, scheduled like jury duty or maybe windows of time, and not intended to be studied for but to assess steady state abilities. Imagine if there were 1000 facts that were deemed essential and that physicians should be expected to know cold – facts that were prioritized by the exam committees – and that were closed book – and some small fraction were presented on a regular basis – but you were handed them – and you knew you were expected to know them – and as they changed you would get updates – and then the rest, like real practice, was open-book. And the frequent tests could be taken in your home – with remote proctoring – in chunks so you could get breaks in between. Physicians would get in a rhythm of every 6 months being assessed – getting feedback – and over 5 years the physician would need to maintain a certain score in order to retain their certification – it is cumulative. And the physician would know all the time if they were on track or not so they could improve/study on their own. This would alleviate anxiety.

GOAL: To reduce anxiety for diplomates through lower-stakes assessments.

PROS

- Lower-stakes assessment administered more frequently, but in a more relaxed administration setting, would reduce stress and anxiety.
- Potential for portions of assessment to be open-book would better reflect how physicians perform their role in actual practice.
- Diagnostic feedback would promote physicians' awareness of their own strengths and weaknesses for both their practice and the assessment.
- It encourages physicians to maintain their skills over time, rather than cramming for a test.

CONS

- Take-home exams create numerous security challenges that may not be easily met with remote proctoring.
- Open-book exams will require appropriate resources that are agreed upon prior to administration.
- Scoring approach is an innovative one that could present significant psychometric challenges.
- The frequency of testing in a secure setting (either test center or remote proctoring from home) would increase.

Idea 5. Diagnosing Mastery of the Breadth of the Discipline

Summary

Rather than ask MOC candidates to sit for the secure exam "cold" after 6-9 years since their last high-stakes experience, a powerful diagnostic assessment would first be taken in Part II to reveal loss of breadth and depth of blueprinted content. The assessment would provide immediate feedback, based on the web-based adaptive assessment and prepare physicians for the high stakes exam taken in a p proctored setting.

Part II SEP products have never been considered "test preparation" materials. Instead, they have been designed to point out recent advances in medicine and provide hyperlinks to reading material in the recent advances topics. The idea proposed here is to create one new Part II product for each discipline, to be designed as *preparation* for the high-stakes Part III exams that follow.

To create the diagnostic assessment, new non-secure question pools would need to be written with different characteristics than the secure pool. Whereas the bulk of the questions in the secure question pool are written to a single (cut score) level of difficulty, there would need to be a large number of questions at *all* levels of difficulty in the new pools. The diagnostic pool would also need to feature probing questions that "drill down" to more detailed topics than secure exam questions do.

With this new question pool in place, the new assessment could be administered online at home with immediate feedback on strengths and weaknesses. The assessment could choose next questions based on responses made so far (i.e., be "adaptive" or "tailored"). The feedback could be rich in diagnostic information and could even take the additional step of prescribing study plans to brush up on topics of non-mastery. This would prepare physicians for taking the high-stakes exam and reduce anxiety.

GOAL: To reduce anxiety for diplomates through the delivery of a Part II product that would provide rich feedback on strengths and weaknesses as preparation for the secure assessment.

PROS

- Being given the ability to use a Part II product to adequately prepare for the examination would reduce the stress and anxiety for diplomates for the Part III exam.
- Diagnostic feedback would promote physicians' awareness of their own strengths and weaknesses they could use for self-reflection and to better prepare for the secure exam.
- The assessment would use examinee time efficiently where content already mastered would be skipped.

CONS

- Building a larger, non-secure item pool that drills down to deep levels of the blueprint would require a significant amount of exam development time.

Idea 6. Practice Evaluation Using Population Health Measures

Summary

Physicians could opt for a performance assessment in lieu of passing the secure exam by demonstrating the capacity to maintaining control of physician-sensitive measures of their patient population.

Physicians would need to demonstrate the capacity to maintain control of physician-sensitive measures of his or her patient population. The process involves continuous collection for at least 10 years of 5 clinician-level, HEDIS metrics (NCQA or NQF endorsed) appropriate for the physician's practice and relevant to the patient population that the physician serves. These measures would come from a verified EHR or NIH grade registry with patients attributed to the physician's care. The registries for the patient population metrics must meet meaningful use criteria. Data reported would include at least 40 time points (i.e., quarterly over 10 years or more) demonstrating statistical quality control of the 5 measures. The physician provides attestation that the measures submitted are attributed to her care and provided to patients she serves. The physician's performance measures may be shown graphically demonstrating that the clinician can provide quality care within statistical limits of control.

These quality measures and the methodology proposed are identical to the process used by large health systems and government or private healthcare payers to compare clinician performance. The system proposed provides objective measures for evaluating a physician's practice behaviors that are widely recognized and used throughout the nation's healthcare system in a statistically valid manner. Physicians may be allowed to pick their personal measures or if the Board so wished, could devise a list of clinician-sensitive measures from which each physician may pick to demonstrate their high quality of patient care. If the physician shows that they can provide quality care at a certain threshold they would be able to opt out of taking the secure exam.

GOAL: To provide an objective assessment of a physician's quality of care for patients relevant to the physician's practice career and allow MOC to measure the highest quality standards for care.

PROS

- Enables physicians to demonstrate their quality of care in medical areas that they routinely practice with objective, statistically valid, and nationally recognized indicators of quality.
- Data can be collected and organized in a meaningful way for physicians to make compliance with MOC requirements more convenient.

CONS

- Using practice analysis alone is missing the component of a direct assessment of medical knowledge.
- Physicians in smaller practices may not have the resources needed to collect and report objective quality measures and other physicians may not want to share quality measures with the ABIM.
- Collecting and organizing data in a meaningful way to provide results on a physician's care quality takes significantly more time than an assessment.