What we measure . . . and what we should measure in medical education

John R Boulet[1] & Steven J Durning[2]

CONTEXT As the practice of medicine evolves, the knowledge, skills and attitudes required to provide patient care will continue to change. These competency-based changes will necessitate the restructuring of assessment systems. High-quality assessment programmes are needed to fulfil health professions education's contract with society.

OBJECTIVES We discuss several issues that are important to consider when developing assessments in health professions education. We organise the discussion along the continuum of medical education, outlining the tension between what has been deemed important to measure and what should be measured. We also attempt to alleviate some of the apprehension associated with measuring evolving competencies by discussing how emerging technologies, including simulation and artificial intelligence, can play a role.

METHODS We focus our thoughts on the assessment of competencies that, at least historically, have been difficult to measure. We highlight several assessment challenges, discuss some of the important issues concerning the validity of assessment scores, and argue that medical educators must do a better job of justifying their use of specific assessment strategies.

DISCUSSION As in most professions, there are clear tensions in medicine in relation to what should be assessed, who should be responsible for administering assessment content, and how much evidence should be gathered to support the evaluation process. Although there have been advances in assessment practices, there is still room for improvement. From the student's, resident's and practising physician's perspectives, assessments need to be relevant. Knowledge is certainly required, but there are other qualities and attributes that are important, and perhaps far more important. Research efforts spent now on delineating what makes a good physician, and on aligning new and upcoming assessment tools with the relevant competencies, will ensure that assessment practices, whether aimed at establishing competence or at fostering learning, are effective with respect to their primary goal: to produce qualified physicians.

Medical Education 2019: 53: 86–94
doi: 10.1111/medu.13652

[1]Foundation for Advancement of International Medical Education and Research (FAIMER), Philadelphia, Pennsylvania, USA
[2]Department of Medicine, Uniformed Services University of the Health Sciences, Bethesda, Maryland, USA

Correspondence: John R Boulet, Foundation for Advancement of International Medical Education and Research, 3624 Market Street, Philadelphia, Pennsylvania 19104, USA.
Tel: 00 1 215 823 2227; E-mail: jboulet@ecfmg.org

INTRODUCTION

The practice of medicine has and will continue to change. Advances in technology and changes in patient care models, combined with a diffusion of scopes of practice, will require that physicians and other health care practitioners learn new skills and procedures. The breadth and depth of medical knowledge in the field will continue to expand at a rapid pace. Likewise, previously learned patient management strategies will be replaced with more effective ones, further changing the health care provider landscape.[1]

All of these changes, either directly or indirectly, have led to changes in individual assessments, assessment programmes and the ways in which stakeholders use assessment data to improve quality.[2,3] Unfortunately, many assessment processes remain outdated, are based on what is easy to measure, or have limited evidence to support their use.[4–7] As a result, and based on a number of frameworks, their validity is questionable.[8] On a more positive note, many current changes focus on what should be measured, however difficult, as opposed to what is inexpensive or easy to measure. More importantly, there has been a movement towards integrated longitudinal assessment programmes that allow for a more continuous evaluation of knowledge, skills and attitudes. These programmes, which better integrate education and assessment, may be more likely to produce competent lifelong learners.[9] From a practical perspective, technological advances now allow for more efficient collection, storage and processing of assessment data. For integrated longitudinal assessment programmes, or any model that relies on linking the results from several assessments together, these advances can offer both labour- and time-saving benefits.

In this article, we outline some of the issues we believe are important to consider when developing assessments. We focus on synthesising our current understanding of the tensions associated with assessment as opposed to suggesting specific interventions to address them. Although our arguments pertain to all health professions, we organise the discussion along the continuum of medical education, from selection into medical school, through undergraduate and postgraduate training, to independent practice. We purposely focus on the assessment of competencies that, at least historically, have been difficult to measure. We outline specific assessment challenges, many of which are currently being addressed in medical schools, postgraduate programmes and as part of post-licensure activities. We also examine some of the important issues concerning the validity of assessment scores and argue that we must do a better job of justifying our use of specific assessment strategies and associated assessment instruments. Finally, we contend that embracing technology as part of the assessment process can help ease some of the administrative burden associated with measuring relevant student, resident and practising physician competencies. This, in itself, could alleviate some of the assessment-related tension in medical education.

ASSESSMENT OF PHYSICIANS (AND THOSE WHO WANT TO BE)

Most practising physicians, having themselves experienced certification and licensure examinations, are well aware of the role of assessment in medical education. Various forms of assessment are used in the selection of medical students, the evaluation of progress in undergraduate medical education, residency selection, specialty certification and maintenance of licensure or certification (revalidation). Although the purposes of these assessments can be quite different, they all rely on the measurement of the knowledge, skills and attitudes needed by individuals at specific time-points in their careers. Notwithstanding that these assessments have changed over time, perhaps for the better, we argue that more work is needed to better align these assessments, in terms of what they measure, with the evolving skill sets required for practice. Moreover, we assert that measuring these changing skill sets can be made more efficient and effective through the adoption of various technologies and new assessment methods.

In the following sections, we specifically outline how assessments used for medical student selection, progression in the curriculum and licensure/certification have changed over the years, often by incorporating new measurement domains and methods. Here, we aim to provide the reader with a general overview of the challenges faced by the assessment community as the knowledge, skills and attitudes required in the profession evolve. Following this, we provide a brief overview of new assessment methods, highlighting the role that technology has played in expanding the measurement domain and making data collection,

storage and dissemination more efficient. Although these technological advances will not eliminate all the tension associated with assessing evolving medical competencies, they should make the process more effective, efficient and meaningful. Finally, we present some fundamental validity issues. If we are not measuring the right constructs, as part of assessments either 'for' or 'of' learning, the consequential impact of the assessment will be muted and tensions may rise.

Drawing from the literature on test-enhanced learning, it is arguable that repeated testing produces superior transfer of learning relative to repeated studying.[10] Thus, to the extent that we can efficiently administer more assessments, the learner will ultimately benefit.[11] However, if assessments are not properly targeted at the right proficiencies, or the content is not weighted appropriately, they may drive learning in the wrong direction.[12,13] Thus, the measurement tension that we see in medical education revolves, at least to some extent, around the trade-offs associated with administering more (valid) assessments and the resources required to do so.

ASSESSMENTS USED FOR SELECTION

Assessments used for the selection of medical students have evolved over time.[14,15] In line with this, numerous research articles about how to best select medical students and residents have been published.[16] Some of these investigations have even reported data to support the fact that assessment results are predictive of future performance (i.e. establish predictive validity).[17–19] Putting aside the validity of these assessments or, more appropriately, the inferences we may wish to make based on assessment scores, there has been a general recognition that, in addition to knowledge, other attributes and skills are required to be successful as a medical student or as a resident (postgraduate trainee). Historically, primarily knowledge-based assessments (e.g. the Medical College Admission Test [MCAT]) were used as initial screens for entry into medical school. Many of these written assessments were later modified to include measures of problem solving and clinical reasoning. The use of multiple mini-interviews (MMIs) is now part of many medical school selection programmes. These interviews, often structured like objective structured clinical examinations (OSCEs), allow programmes to assess verbal and non-verbal skills that are difficult, or impossible, to measure using standardised written examinations or through the evaluation of transcripts. Schools using the MMI format believe it produces a more reliable assessment of a candidate.[20] Because students interact with multiple interviewers over the course of the MMI, the opinion of a single interviewer is not over-emphasised. Unlike the traditional one-on-one interview, the MMI allows applicants multiple opportunities to showcase their skills, including their abilities to communicate and reason.

Many factors could impact the future success of a medical student and some of these cannot be controlled. It seems reasonable, however, that attributes other than, or in addition to, intelligence or application of knowledge should be emphasised. Patient care relies on leadership, teamwork, communication, conscientiousness, adaptiveness and a host of other individual and group attributes. Although some of these can be taught, others are more trait-like (e.g. personality), and not easily modified. Their measurement, at least in terms of validity considerations, deserves attention.

ASSESSMENTS USED IN EDUCATION PROGRAMMES

Once a student enters a health professions programme, at either undergraduate or graduate level, he or she will encounter numerous assessments, of which many are aimed at improving knowledge and skills, and others are designed to support decisions regarding competency. Knowledge, and its application, is at the core of most assessment programmes, at least for more 'junior' years. Other competencies become more important as training progresses. Arguably, it is these competencies, including leadership, teamwork, interpersonal skills, communication, adaptability, use of information technology, and awareness of and ability to navigate the health care system, amongst others, that define the professional. In many countries, this has been tacitly acknowledged through the development of a variety of performance-based assessments, including, amongst others, the OSCE, mini-clinical evaluation exercise (mini-CEX) and chart-stimulated recall (CSR).[21,22] These performance-based assessments, whether conducted in a standardised environment or in the workplace, aim to measure more than just knowledge. Although they are administratively burdensome, and typically costly, they allow for the assessment of competencies that are important to medical practice.

In both undergraduate and graduate medical education, there has been a movement towards the

measuring of growth. To measure growth, point-in-time assessments need to be more frequent. There must also be systems in place to store and analyse performance data. In undergraduate medical education, the use of assessments with overlapping or equivalent content, often referred to as 'progress testing', allows for more systematic measurement of growth in student abilities.[23] This growth, although relatively easy to measure for knowledge, can also be measured for clinical skills by employing linked OSCEs.[24] There has also been a movement towards programmatic assessment, in which individual assessment methods are chosen specifically to align with curriculum outcomes. Here, where efforts have been made to develop proper assessment blueprints, meaningful point-in-time assessment data can be aggregated longitudinally to support decisions regarding specific competencies.[25,26] Graduate medical education programmes now focus on competency-based medical education, using specific 'milestones' to both facilitate improvements to the curricula and to document the progression of trainees.[27–29] The goals of these milestones are to define the essential competencies in each specialty. Being more explicit about the competencies, and focusing on the individual learner, allows for the development of meaningful assessments that will help the individual trainee progress to become the best possible provider. Developing information systems to store longitudinal performance data will facilitate the achievement of assessment goals, including those related to learning and competency attainment.

The assessments used in education programmes have clearly extended beyond the measurement of knowledge. There has also been a movement towards more formative assessment practices that focus on the provision of meaningful feedback. From an educational perspective, both of these trends are positive. They should lead to the development of more capable practitioners if the assessments are well constructed, are supported by a suitable information technology infrastructure, measure the important domains or competencies, are administered relatively frequently, and yield reliable and valid ability estimates.

ASSESSMENTS USED FOR LICENSURE, CERTIFICATION OR RECERTIFICATION

Licensure examinations are used in many countries to decide who can, or cannot, practise medicine.[30] Historically, these assessments were primarily concerned with measuring knowledge or, at best, the application of knowledge (e.g. in pharmacology). They typically employ selected-response formats. These assessments are costly to construct because they require large item banks. Nevertheless, they are efficient and yield fairly precise estimates of ability, at least on the construct (e.g. interpretation of laboratory tests), or constructs, being measured. However, although knowledge and application of knowledge are important, the provision of adequate patient care requires other abilities. The assessment of these abilities can be accomplished in many ways, including through OSCEs and various *in situ* workplace-based formats such as the mini-CEX and CSR.

In 1992, the Medical Council of Canada (MCC) introduced the MCC Qualifying Examination Part II (MCCQE II) for the Licentiate of the Medical Council of Canada (LMCC). Canadian medical regulatory authorities normally require that prospective physicians have the LMCC to apply for a medical licence within their province or territory. The MCCQE Part II utilises standardised patients (SPs), who are lay people trained to simulate the conditions of real patients. Examinees interact with several SPs, allowing for the assessment of their history taking, physical examination and communication abilities. Similar types of assessment were introduced by the Educational Commission for Foreign Medical Graduates (ECFMG) in 1998, by the National Board of Osteopathic Medical Examiners (NBOME) in 2004, and by the United States Medical Licensing Examination (USMLE) in 2004.[31] The introduction of these performance-based examinations as part of the licensure process highlights the general recognition that clinical skills are essential for the provision of patient care. Like other licensure examination formats, they are expensive to administer and their validity remains subject to debate.[32] Their introduction did, however, push medical schools to emphasise clinical skills in their curricula. From a consequential validity perspective, the assessments have certainly driven learning.[33] More importantly, and discussed in more length later in this article, new assessment methods, including simulation, made possible through technological innovation, afford the opportunity to more efficiently measure important constructs that were previously ignored. We would argue that these constructs (e.g. teamwork, communication), combined with skills to promote self-regulated learning, are likely to make the most difference in patient care.

There is a general recognition that recertification (or revalidation) is necessary for physicians to maintain their knowledge and skills over time, and that practitioners can improve their performance and patient care by actively participating in education and assessment activities.[34,35] Many countries, including the USA and the UK, have well-established certification and recertification programmes. Nevertheless, various stakeholders have complained that the assessments are not appropriate or are irrelevant to their everyday practice. The fundamental issue is that assessment content is not adequately aligned with the knowledge and skills required for effective practice, at least for the physician at that point in his or her career. From an assessment perspective, there has also been some debate as to whether recertification (or revalidation) examinations should be high-stakes assessments 'of learning' or more formative assessments 'for learning', or some combination thereof.[36] However, regardless of the primary purpose of the assessment, the content and constructs being measured must be aligned with the education needs of the practitioner. Ideally, to minimise the tension between the certification agencies and the diplomates they serve, these assessments should be relevant to practice, spur learning and lead to more effective patient care. Most importantly, and as will be discussed later in this article, evidence should be gathered to support these positive assessment attributes.

In the USA, specialty boards govern the certification and recertification processes. The requirements for initial specialty certification can vary, but typically include the completion of an education programme (residency, postgraduate training) and some form of standardised assessment. Passing these standardised assessments, which are often based largely on multiple-choice questions (MCQs), provides some evidence of a physician's expertise in a particular specialty or subspecialty of medical practice. For most specialties, the certification examination is taken at the end of postgraduate training and is administered in a selected-response format. More recently, certification examinations have been spread out (taken in parts, earlier and later in specialty training) and, in some cases, expanded to include some form of performance-based assessment.[37] The spacing of initial certification examinations makes sense in that it recognises that point-in-time assessments may not optimise learning, potentially increase examinee stress, and may not indoctrinate future practitioners into a 'longitudinal' assessment paradigm. As in medical licensure examinations, the introduction of performance-based assessments into specialty certification recognises the fact that abilities other than knowledge (e.g. communication, teamwork and procedures) are important parts of a physician's expertise.

Maintenance of certification (MoC), or revalidation, is a system of ongoing professional development and practice assessment and improvement. Until recently, one of the major requirements for MoC, at least in the USA, was the resitting of some form of the initial board certification examination, usually at 7- to 10-year intervals. Although this strategy ensures that 'recertification' means the same thing as 'initial certification' and lets various stakeholders (e.g. patients) know that the criteria, at least in terms of assessment, are effectively equivalent, it is well known that specialists' practice domains tend to change over time, often narrowing to specific types of patient with specific conditions. As such, it has been argued that simply repeating the certification examination, at whatever interval, is not, in itself, very meaningful, at least in terms of motivating physicians to enhance their abilities. To address this issue, some specialty boards have introduced more continuous assessment models by providing diplomates with MCQ-based assessments (approximately 30 per quarter) to be taken over the course of a year.[38] In theory, these items, which can be delivered on a smartphone, can be adapted to the diplomate's practice domain and ability, and can be quickly modified to meet a current health care concern (e.g. opioid epidemic). More importantly, the continuous assessment model helps individuals retain knowledge.[39] Much of the argument against this model, often exacerbating the tension between test developers/psychometricians and individuals concerned primarily with education outcomes, centres on security. How can we know that the individual answering the items is who he says he is? From a summative assessment perspective, this is certainly a problem. However, from a formative assessment perspective, and assuming that board-certified physicians are professionals and actually want to provide better, more informed patient care (a reasonable hypothesis), the potential security holes, perhaps exploited by a few individuals, may be outweighed by the active engagement of the majority of learners.

In postgraduate training, certification and maintenance of certification, the so-called non-technical abilities are often ignored, at least from a

standardised assessment perspective. It should be noted, however, that many of these abilities (e.g. procedural) or attitudes (e.g. professionalism) are assessed through workplace-based assessments. Although observer-based assessments (e.g. the mini-CEX) can yield error-prone estimates of ability, especially if relatively few performance samples are obtained, they can be effective formative assessment tools that provide meaningful evaluations of constructs that would be difficult or impossible to measure in non-workplace-based environments (e.g. ethical behaviour). New technology, including assessment data portals and convenient evaluation interfaces, is making it easier and less costly to gather the data needed to make inferences regarding these 'non-technical' abilities.

NEW(ER) ASSESSMENT METHODOLOGIES

Advances in technology and assessment methods, including simulation, have expanded the scope of what can be measured, at least in a standardised way.[40] Technology, in many respects, has and will continue to expand the domain of what can be assessed. Furthermore, it can make assessment more efficient, removing the economic barriers that often curtail assessment programmes or processes. In this section, we argue that the use of technology (e.g. simulation-based assessment) can make the assessment process more effective and efficient. To the extent that we can measure the competencies required for patient care, and do so in cost-effective ways, the various tensions associated with assessment may be reduced. However, given that the administration of more assessments is viable, evidence to support their validity must still be gathered. Although test developers may see the benefits of new assessments, test takers, who often have to pay for assessments and for research to support their validity, may have a different perspective. As such, tensions could rise.

The use of OSCEs is quite prevalent and dates back nearly 50 years. Over the same period, the modelling of typical provider–patient interactions has evolved. The use of moulage, confederate family members, programmed examination tools (e.g. stethoscopes with heart sounds), hybrid stations that involve follow-up visits, adaptive simulations, etc., has led to both greater simulation fidelity and expansions in the measurement domain (e.g. measurement of teamwork).[41,42] There has also been a growing use of electromechanical manikins in all health care disciplines. Physiologically, these manikins can be programmed to react to interventions (e.g. intubation, drug administrations) just as real humans would. They too have expanded the measurement domain by allowing for the simulation of physical findings (e.g. dysrhythmia) that cannot be simulated in SPs. They have proved to be quite effective in training practitioners to deal with rare events that are encountered infrequently in real patient care settings.[43]

Although OSCEs and other performance-based assessments can be quite expensive to administer, technology can, and will continue to, mitigate the costs. There are now wearable devices that can allow for the authentic indirect observation of practitioners as they interact with patients or other health care workers. Computer-based training systems can be used to help SPs portray their cases.[44] Automated collection and analysis of performance data, including data obtained from manikins and part-task trainers equipped with sensors, are routine at many educational institutions.[45] With online assessment systems, through which automated scoring can be implemented, students can receive instant feedback and tutors/preceptors can monitor progress.[46] The application of artificial intelligence (AI), either for scoring written exercises or for automating the assessment of procedural, communication (via facial recognition or linguistic analysis) or history taking (speech to text) skills, has the potential to eliminate the need for human ratings, something that would decrease assessment costs.[47] At the very least, these technologies, including the use of cameras in examination rooms, could be used to support quality assurance initiatives.

THE VALIDITY OF ASSESSMENT SCORES

Technological advances can both expand the measurement domain in medical education and allow for the use of novel scoring tools, including various applications of AI. They will not, however, alleviate the need to gather data to support the psychometric adequacy of assessment scores or any competency decisions we make based on these scores. Instead, with the emergence of new ways to assess students, residents and practising physicians, there is even more impetus to conduct research studies to support validity arguments. These investigations, if carried out with appropriate rigour, can help support new assessment formats, thereby easing the tension for those who believe that we are not adequately assessing the important constructs in

medicine. As noted previously, however, gathering validity evidence for new assessments can be costly and this cost is often borne by the test taker. As such, assessment-related tensions, at least for some groups, may actually increase.

There have been articles written about validity frameworks and how evidence to support the validity of the scores obtained in various assessments employed in medical education can be collected.[5,19] In gathering this evidence, we must ask: Validity for what? For assessments used to make selection decisions, we need some evidence that those who are selected are up to the task (i.e. they are successful in the programme). For certification and licensure examinations, the ultimate goal of which is the protection of the public, we must ask: What evidence indicates that practitioners are qualified? For maintenance of certification, or revalidation, we must ask: How will we know that those who take the assessments are better practitioners? These are not easy questions to answer. In most instances, we cannot conduct controlled experiments (e.g. by waiving assessment requirements for a random sample of those seeking practice licences and investigating whether they have worse patient outcomes). Furthermore, for any predictive validity considerations (i.e. how assessment results are related to future performance), numerous confounding variables make it difficult, or impossible, to attribute specific outcomes (e.g. mortality) to specific providers. Nevertheless, the stakeholders are demanding, and have a right to, information that supports, or refutes, the use of specific assessments.[48]

The introduction of new assessment modalities and new measured constructs magnifies validity concerns. Often, it is very difficult to define the construct of interest, let alone measure it. This does not absolve assessment developers of their responsibility to gather evidence to support the validity of their assessment scores, or any competency decisions based on the scores. Instead, this process should be looked upon as a research challenge that will demand resources and cooperation amongst various stakeholders from undergraduate, graduate and continuing medical education programmes. As much of the validity argument rests with providing evidence that performance at a certain point in time is related to future performance, the development of longitudinal databases, in which assessment results can be linked to an individual, is essential. Fortunately, many medical schools and residency programmes have developed information systems that allow for the storage and retrieval of longitudinal data concerning progress through the curriculum and the application of learning analytics.[49] Likewise, with electronic medical records, unique provider identifiers and access to patient records, it has become easier to conduct studies that specifically link assessment data and quality of care.[50]

CONCLUSIONS

As in any profession, there are clearly tensions in medicine as to what should be assessed, who should be responsible for administering assessment content, and how evidence should be gathered to support the evaluation process. Although there have been great developments in assessment practices, including a broadening of the measurable content domain (expansion of the competencies, work-based assessments), a movement towards more longitudinal, programmatic delivery models, implementation of technological improvements in simulation and scoring, and more rigorous studies to support the validity of scores (or decisions based on scores), there is still room for improvement.

From the physician's perspective, assessments, regardless of purpose, need to measure domains that are important to the practice of medicine. To us, this is a key step in lessening the tension between test developers and test takers. Knowledge is certainly required, but there are without doubt other qualities that are important, and perhaps far more important. Thinking about the future of assessment begs the question: What are we not measuring that we should be measuring? A secondary query relates to how this could be accomplished. For the first question, it is clear that the evolving practice of medicine will necessitate the measurement of certain constructs (or domains) that are currently not emphasised. Teamwork, which is not measured in many of the current standardised assessments, is a fundamental part of patient care. Other domains related to system-based practice (e.g. understanding the costs and benefits of interventions) can be measured indirectly through OSCEs, but are often not measured at all. We should also be thinking about what the practice of medicine will look like in the future. What will the physical examination of the future entail? Is it necessary to memorise and regurgitate so much information, or is the ability to seek out information and synthesise it using point-of-care resources more relevant to practice? Answers to these questions will

inform the development of meaningful assessments. Once the measurement domain is better defined, the 'how' questions will be easier to answer.

Technology will clearly be important, both for developing more relevant assessments (e.g. using electromechanical manikins to measure teamwork in the management of critical care events) and for making assessments more efficient (e.g. automated scoring). Likewise, the ability to construct and maintain longitudinal datasets, including links to patient data, will allow for the conduct of meaningful validity studies that connect assessment results to future short- and long-term outcomes. Effort spent now on delineating what makes a good physician, and on aligning new and upcoming assessment tools with the relevant competencies, will ensure that assessment practices, whether aimed at establishing competence or at fostering learning, are effective with respect to their primary goal: to produce qualified physicians.

Contributors: both authors contributed to the conception and drafting of this article. Both authors contributed to the critical revision of the paper and have approved the final manuscript for publication.
Acknowledgements: none.
Funding: none.
Conflicts of interest: none.
Ethical approval: not applicable.

REFERENCES

1 Laiteerapong N, Huang ES. The pace of change in medical practice and health policy: collision or coexistence? *J Gen Intern Med* 2015;**30** (6):848–52.
2 Vanderbilt AA, Perkins SQ, Muscaro MK, Papadimos TJ, Baugh RF. Creating physicians of the 21st century: assessment of the clinical years. *Adv Med Educ Pract* 2017;**8**:395–8.
3 Haist SA, Butler AP, Paniagua MA. Testing and evaluation: the present and future of the assessment of medical professionals. *Adv Physiol Educ* 2017;**41** (1):149–53.
4 Epstein RM. Assessment in medical education. *N Engl J Med* 2007;**356** (4):387–96.
5 Lineberry M, Soo PY, Cook DA, Yudkowsky R. Making the case for mastery learning assessments: key issues in validation and justification. *Acad Med* 2015;**90** (11):1445–50.
6 Whitehouse A, Higginbotham L, Nathavitharana K, Singh B, Hassell A. Team assessment of behaviour: a high stakes assessment with potential for poor implementation and impaired validity. *Clin Med (Lond)* 2015;**15** (1):7–9.
7 Lockyer J. Multisource feedback: can it meet criteria for good assessment? *J Contin Educ Health Prof* 2013;**33** (2):89–98.
8 Messick S. Meaning and values in test validation: the science and ethics of assessment. *Educ Res* 1989;**18** (2):5–11.
9 van der Vleuten CP, Schuwirth LW, Driessen EW, Dijkstra J, Tigelaar D, Baartman LK, van Tartwijk J. A model for programmatic assessment fit for purpose. *Med Teach* 2012;**34** (3):205–14.
10 Larsen DP, Butler AC, Lawson AL, Roediger HL III. The importance of seeing the patient: test-enhanced learning with standardized patients and written tests improves clinical application of knowledge. *Adv Health Sci Educ Theory Pract* 2013;**18** (3):409–25.
11 Butler AC. Repeated testing produces superior transfer of learning relative to repeated studying. *J Exp Psychol Learn Mem Cogn* 2010;**35** (5):1118–33.
12 Wormald BW, Schoeman S, Somasunderam A, Penn M. Assessment drives learning: an unavoidable truth? *Anat Sci Educ* 2009;**2** (5):199–204.
13 Cilliers FJ. Is assessment good for learning or learning good for assessment? A. Both? B. Neither? C. It depends?*Perspect Med Educ* 2015;**4** (6):280–1.
14 Eva KW, Rosenfeld J, Reiter HI, Norman GR. An admissions OSCE: the multiple mini-interview. *Med Educ* 2004;**38** (3):314–26.
15 Patterson F, Cleland J, Cousans F. Selection methods in healthcare professions: where are we now and where next? *Adv Health Sci Educ Theory Pract* 2017;**22** (2):229–42.
16 Simpson PL, Scicluna HA, Jones PD, Cole AM, O'Sullivan AJ, Harris PG, Velan G, McNeil HP. Predictive validity of a new integrated selection process for medical school admission. *BMC Med Educ* 2014;**14**:86.
17 Patterson F, Knight A, Dowell J, Nicholson S, Cousans F, Cleland J. How effective are selection methods in medical education? A systematic review. *Med Educ* 2016;**50** (1):36–60.
18 Patterson F, Cousans F, Edwards H, Rosselli A, Nicholson S, Wright B. The predictive validity of a text-based situational judgment test in undergraduate medical and dental school admissions. *Acad Med* 2017;**92** (9):1250–3.
19 Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-clinical evaluation exercise: a review of the research. *Acad Med* 2010;**85** (9):1453–61.
20 Rees EL, Hawarden AW, Dent G, Hays R, Bates J, Hassell AB. Evidence regarding the utility of multiple mini-interview (MMI) for selection to undergraduate health programs: a BEME systematic review: BEME Guide No. 37. *Med Teach* 2016;**38** (5):443–55.
21 Norcini JJ, McKinley DW. Assessment methods in medical education. *Teach Teach Educ* 2007;**23** (3): 239–50.
22 Carraccio C, Englander R. The objective structured clinical examination: a step in the direction of

competency-based evaluation. *Arch Pediatr Adolesc Med* 2000;**154** (7):736–41.
23 Schuwirth LW, van der Vleuten CP. The use of progress testing. *Perspect Med Educ* 2012;**1** (1):24–30.
24 Pugh D, Touchie C, Wood TJ, Humphrey-Murto S. Progress testing: is there a role for the OSCE? *Med Educ* 2014;**48** (6):623–31.
25 Schuwirth LW, van der Vleuten CP. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 2011;**33** (6):478–85.
26 van der Vleuten CP, Schuwirth LW, Driessen EW, Govaerts MJ, Heeneman S. 12 Tips for programmatic assessment. *Med Teach* 2015;**37**(7):641–6.
27 Frank JR, Snell LS, ten Cate O et al. Competency-based medical education: theory to practice. *Med Teach* 2010;**32** (8):638–45.
28 Holmboe ES, Call S, Ficalora RD. Milestones and competency-based medical education in internal medicine. *JAMA Intern Med* 2016;**176** (11):1601–2.
29 Teherani A, Chen HC. The next steps in competency-based medical education: milestones, entrustable professional activities and observable practice activities. *J Gen Intern Med* 2014;**29** (8):1090–2.
30 Archer J, Lynn N, Coombes L, Roberts M, Gale T, Price T, Regan de Bere S. The impact of large scale licensing examinations in highly developed countries: a systematic review. *BMC Med Educ* 2016;**16** (1):212.
31 Boulet JR, Smee SM, Dillon GF, Gimpel JR. The use of standardized patient assessments for certification and licensure decisions. *Simul Healthc* 2009;**4** (1):35–42.
32 Burdick WP, Boulet JR, LeBlanc KE. Can we increase the value and decrease the cost of clinical skills assessment? *Acad Med* 2017;**93**:690–2.
33 Wood T. Assessment not only drives learning, it may also help learning. *Med Educ* 2009;**43** (1):5–6.
34 Gray BM, Vandergrift JL, Lipner RS. Association between the American Board of Internal Medicine's general internist's Maintenance of Certification requirement and mammography screening for Medicare beneficiaries. *Womens Health Issues* 2018;**28** (1):35–41.
35 Vandergrift JL, Gray BM, Weng W. Do state continuing medical education requirements for physicians improve clinical knowledge? *Health Serv Res* 2017;**53** (3):1682–701.
36 Lockyer J, Bursey F, Richardson D, Frank JR, Snell L, Campbell C. Competency-based medical education and continuing professional development: a conceptualization for change. *Med Teach* 2017;**39** (6):617–22.
37 Zhou Y, Sun H, Lien CA, Keegan MT, Wang T, Harman AE, Warner DO. Effect of the BASIC examination on knowledge acquisition during anesthesiology residency. *Anesthesiology* 2018;**128** (4):813–20.
38 Sun H, Zhou Y, Culley DJ, Lien CA, Harman AE, Warner DO. Association between participation in an intensive longitudinal assessment program and performance on a cognitive examination in the Maintenance of Certification in Anesthesiology Program®. *Anesthesiology* 2016;**125** (5):1046–55.
39 Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci* 2006;**17** (3):249–55.
40 Ryall T, Judd BK, Gordon CJ. Simulation-based assessments in health professional education: a systematic review. *J Multidiscip Healthc* 2016;**9**:69–82.
41 Boulet JR, Natt N, Hawkins R. Direct observation: standardized patients. In: Holmboe ES, Hawkins RE, Durning SJ, eds. *Practical Guide to the Evaluation of Clinical Competence*, 2nd edn. Philadelphia, PA: Mosby/Elsevier 2017;102–18.
42 Bruen C, Kreiter C, Wade V, Pawlikowska T. Investigating a self-scoring interview simulation for learning and assessment in the medical consultation. *Adv Med Educ Pract* 2017;**8**:353–8.
43 Griswold S, Fralliccardi A, Boulet J, Moadel T, Franzen D, Auerbach M, Hart D, Goswami V, Hui J, Gordon JA. Simulation-based education to ensure provider competency within the health care system. *Acad Emerg Med* 2018;**25** (2):168–76.
44 Errichetti A, Boulet JR. Comparing traditional and computer-based training methods for standardized patients. *Acad Med* 2006;**81** (10 Suppl):91–4.
45 Azari DP, Frasier LL, Quamme SRP, Greenberg CC, Pugh CM, Greenberg JA, Radwin RG. Modeling surgical technical skill using expert assessment for automated computer rating. *Ann Surg* 2017; doi: 10.1097/SLA.0000000000002478 [Epub ahead of print.]
46 Walsh K. Point of view: online assessment in medical education – current trends and future directions. *Malawi Med J* 2015;**27** (2):71–2.
47 Williamson DM, Xi X, Breyer FJ. A framework for evaluation and use of automated scoring. *Educ Meas Issues Pract* 2012;**31**:2–13.
48 Byrne BJ, Frintner MP, Abraham HN, Starmer AJ. Attitudes and experiences of early and midcareer pediatricians with the maintenance of certification process. *Acad Pediatr* 2017;**17** (5):487–96.
49 Chan T, Sebok-Syer S, Thoma B, Wise A, Sherbino J, Pusic M. Learning analytics in medical education assessment: the past, the present and the future. *Acad Emerg Med Educ Train* 2018;**2** (2):178–87.
50 Norcini JJ, Lipner RS, Kimball HR. Certifying examination performance and patient outcomes following acute myocardial infarction. *Med Educ* 2002;**36** (9):853–9.

Received 4 February 2018; editorial comments to authors 6 March 2018; accepted for publication 31 May 2018