Adv in Health Sci Educ (2016) 21:897–913
DOI 10.1007/s10459-015-9653-6

REFLECTIONS

Towards a program of assessment for health professionals: from training into practice

Kevin W. Eva[1] · Georges Bordage[2] · Craig Campbell[3] · Robert Galbraith[4] · Shiphra Ginsburg[5] · Eric Holmboe[6] · Glenn Regehr[1]

Received: 18 June 2015 / Accepted: 16 November 2015 / Published online: 21 November 2015

Abstract Despite multifaceted attempts to "protect the public," including the implementation of various assessment practices designed to identify individuals at all stages of training and practice who underperform, profound deficiencies in quality and safety continue to plague the healthcare system. The purpose of this reflections paper is to cast a critical lens on current assessment practices and to offer insights into ways in which they might be adapted to ensure alignment with modern conceptions of health professional education for the ultimate goal of improved healthcare. Three dominant themes will be addressed: (1) The need to redress unintended consequences of competency-based assessment; (2) The potential to design assessment systems that facilitate performance improvement; and (3) The importance of ensuring authentic linkage between assessment and practice. Several principles cut across each of these themes and represent the foundational goals we would put forward as signposts for decision making about the continued evolution of assessment practices in the health professions: (1) Increasing opportunities to promote learning rather than simply measuring performance; (2) Enabling integration across stages of training and practice; and (3) Reinforcing point-in-time assessments with continuous professional development in a way that enhances shared responsibility and accountability between practitioners, educational programs, and testing organizations. Many of the ideas generated represent suggestions for strategies to pilot test, for infrastructure to build, and for harmonization across groups to be enabled. These include novel

✉ Kevin W. Eva
kevin.eva@ubc.ca

[1] Centre for Health Education Scholarship, University of British Columbia, JPPN 3324, 910 West 10th Avenue, Vancouver, BC V5Z 1M9, Canada

[2] University of Illinois at Chicago, Chicago, IL, USA

[3] Royal College of Physicians and Surgeons of Canada, Ottawa, ON, Canada

[4] National Board of Medical Examiners, Philadelphia, PA, USA

[5] University of Toronto, Toronto, ON, Canada

[6] Accreditation Council for Graduate Medical Education, Chicago, IL, USA

strategies for OSCE station development, formative (diagnostic) assessment protocols tailored to shed light on the practices of individual clinicians, the use of continuous workplace-based assessment, and broadening the focus of high-stakes decision making beyond determining who passes and who fails. We conclude with reflections on systemic (i.e., cultural) barriers that may need to be overcome to move towards a more integrated, efficient, and effective system of assessment.

Keywords Assessment · Health professional education · Competency-based education · Continuing professional development

Background

Society's implied social contract with any health profession includes the obligation of that profession to self-regulate in a manner that ensures the protection of patients (Cruess and Cruess 2014). To address this obligation, regulatory authorities in the US and Canada have long included formal national examinations as part of their efforts to construct a fair and equitable quality assurance process. The primary mandate of these examinations has been to ensure that doctors have the knowledge and skill required to provide safe and effective health care (Swanson and Roberts 2016) and there is good evidence that scores on such examinations are associated with performance during candidates' subsequent careers (Cadieux et al. 2007; Wenghofer et al. 2009; Tamblyn et al. 2007).

There is also evidence, however, that increases in the amount of high-stakes assessment have not led to improvements in quality and safety; rather, the opposite has been observed with the most recent data (James 2013) suggesting that the actual number of deaths due to medical error and poor quality healthcare in the US may be three times the number reported by the seminal Institute of Medicine Report that was published nearly two decades ago (Kohn et al. 1999). It is unlikely that any one intervention could overcome such pernicious findings and it is inappropriate to blame any one aspect of the system. At a minimum, however, these findings mandate a critical look at our current approaches to assessment.

Such an analysis does not deny the importance of the gatekeeping function of high quality testing buttressed by high psychometric standards, but it does argue for supplementation by ongoing assessment that emphasizes quality improvement efforts during the practice years. In other words, effective continuing professional development is vital for all healthcare providers, regardless of where they sit on the pass-fail continuum according to high-stakes assessment practices (Eva et al. 2013). This argument is not a fundamental shift away from concerns about patient care towards concerns about physician growth, but is a recognition that continuing education offers a means towards improving patient care given that many more patients will be impacted upon by physicians who pass our current exams than by those who fail.

The purpose of this reflections article is, therefore, to cast a critical lens on current assessment practices to offer insights into ways in which they might be adapted to ensure alignment with modern conceptions of health professional education for the ultimate goal of improved healthcare.

Methods and conceptual framing

To ensure that any assessment program has a positive influence on patient care by promoting lifelong professional development, it is important to consider both the implicit messages sent to candidates and stakeholders as well as any unintended consequences of adopting a particular assessment strategy. This is Messick's (1989) notion of consequential validity and requires a broader, scoping review not only of current best practices in assessment, but also of the alignment between current assessment systems and modern understanding of health professional practice (Cizek 2012). In considering these issues, van der Vleuten's (1996) model of utility (reliability, validity, feasibility, acceptability, and educational impact) continues to provide a useful model from which to judge the adequacy of any assessment system. As van der Vleuten notes, compromise is necessary across these factors. In an ideal world this compromise should result from deliberate decision making rather than uninspected imbalance (Razack et al. 2015). At the same time, the broad conception of validity that Kane (1992), Messick (1989), and others (Cook et al. 2015; Cook 2014; Downing 2003) have put forward demands that a wide variety of additional factors be taken into account when deciding whether or not a system of assessment is fit for purpose. For example, the creation of a coherent and integrated system of assessment that promotes ongoing learning across the continuum of training and practice requires a process that (a) is made efficient for candidates, ensuring appropriate and comprehensive coverage of many aspects of performance while eliminating unnecessary redundancy; (b) emphasizes the primacy of learning by harnessing the power of feedback (Boud and Molloy 2013; Galbraith et al. 2011); and (c) creates a shared accountability between the learner, educational programs, and regulatory authorities for engaging in continuous performance improvement (Mylopoulos and Scardamalia 2008; Bordage et al. 2013).

To explore these issues, the authorship team began by reviewing four recent overviews of the current state of health professional assessment: (1) The Future of Medical Education in Canada (AFMC 2010); (2) Assessment in Postgraduate Medical Education (Regehr et al. 2011); (3) Assessment Strategies within the Revised Maintenance of Certification Program (RCPSC 2011); and (4) The Ottawa Conference consensus statement (Norcini et al. 2011). From that foundation we launched iterative discussions, both in person and asynchronously, over the course of 3 years about the key issues facing medical training, regulatory, licensing, and certification authorities in the near future and sought out literature that would inform these discussions with the intent to offer additional perspectives and issues that are meant to supplement these four documents. Our goal was not to debate the merits of any particular form of assessment. Rather, it was to re-formulate general principles that could be relevant to anyone involved with assessment, be they individual course directors, national testing agencies, or anything in between.

Results

The three broad themes discussed below are not intended to be either exclusive or exhaustive. Where possible, we have tried to offer paths for further exploration of the issues raised. In an effort to balance the discussion of these themes between broad perspectives on assessment and the operational needs of the health professions we have focused on three levels of consideration: (1) Conceptual—issues about how, why, and when different assessment practices impact upon the culture of the profession; (2)

Logistical—specific avenues of exploration through which the conceptual issues might be redressed within practical realities; and, (3) Systemic—cultural issues inherent in current practice and education systems that create barriers that need to be overcome. The first two levels will be discussed independently for each theme after which the third, integrating level, will be examined more generally.

Theme 1: Overcoming unintended consequences of competency-based assessment

Conceptual issues

Modern systems of medical education are increasingly centered on competency-based frameworks that outline the core knowledge, skills, and attitudes that individual physicians are expected to maintain. (Frank et al. 2010; Morcke et al. 2013) This movement has had definite positive impact by offering an explicit and broadly focused model from which educators, regulatory authorities, and licensing bodies can guide educational and assessment practices for the sake of improved patient care (Holmboe et al. 2010). While there are well-recognized psychometric challenges for the assessment of many competencies, in general these challenges can be overcome through deliberate and diverse sampling across different situations (Norcini et al. 2003; Eva and Hodges 2012) and the direct observations required offer increased opportunities for formative feedback.

Implicit in common models of competency-based assessment, however, are a variety of assumptions that may have unintended and undesirable consequences (Ginsburg et al. 2010). Most central is the notion that competence is something one can check off. Claiming that a student can "perform a complete and appropriate assessment of a patient," for example, ignores the robust literature indicating that contextual factors play an important role in our ability to perform any task (Eva 2003; Colliver 2002) and risks sending an implicit message that once a task can be achieved there is no further work to be done (Neve and Hanks 2016; Norman et al. 2014; Newell et al. 2001). The fact that every candidate who passes a minimal competence exam is effectively labeled competent overlooks the realities that (a) there is always considerable variability of performance within the passing range, (b) even the top performers have room for improvement, and (c) knowledge and skill are subject to drift and deterioration (decay) over time (Choudhry et al. 2005; Norman et al. 2014). A considerable amount of information is collected on students during their time in medical school, postgraduate training, and while undertaking licensing and certification examinations that is simply ignored when the individual is, at a particular point in time, deemed 'good enough' (Klass 2007). This creates a number of problems.

First, focusing on a determination of 'competent' contributes to assessment protocols being seen as hurdles that one simply needs to get over rather than as diagnostic opportunities that can be put to use for further pedagogic benefit. Second, focusing exclusively on the pass-fail cut-point removes any incentive, and creates considerable disincentive, for disclosing difficulties and continuing to pursue improvement (Eva et al. 2012). Passing the examination may then indicate that the weaknesses one experiences are unimportant (Butler 1987). Third, such competence-based models may reduce the degree of support educators feel compelled to provide given that there is little need to offer guidance to trainees who have been deemed competent. Finally, using the label 'competent' overlooks

the well-established view that knowledge and skills must be continuously used for them to be maintained (Ericsson 2004; Eva 2002). Having successfully crammed to pass an exam should not be viewed as an indication that one will remember the material after the exam is completed (Custers 2010).

Moreover, the "state of independence" that underlies the label of competent runs counter to modern perspectives on expertise, which differentiate between the routine expert who achieves a certain degree of performance and simply reproduces that performance repeatedly (Regehr 1994) and the adaptive expert who continuously reinvests her energies into better understanding and innovating within the domain of practice for the sake of continuous performance improvement (Mylopoulos and Regehr 2011). All of this sums together to create a state in which focusing assessment efforts on the achievement of competence by exclusively striving to identify a minority of individuals who do not meet an established threshold eliminates opportunities to provide formative guidance directing future learning for the majority. While some individuals definitely need to be excluded from practice until such time as their ability can be improved, formally supporting the others in their performance improvement efforts could have a bigger impact on the quality of healthcare that patients receive.

Logistical considerations

Addressing such unintended consequences would require greater emphasis on longitudinal and direct observation, with accompanying feedback, across medical school, postgraduate training, and continuing professional development. This could, as a result, lead to a radical increase in assessment. That must be considered in light of the observations that physicians in practice are already overworked and the organizations responsible for implementation of assessment do not have endless resources. Further, given the value of progressive independence (Kennedy et al. 2009) and of desirable difficulties (i.e., being challenged in a manner that drives learning) for performance improvement (Guadagnoli et al. 2012; Eva 2009; Bjork 1994), there are dangers associated with trainees or practitioners being observed constantly. Thus, as we move forward in addressing the unintended consequences of competency-based assessment it will be important to optimize the time and resources that are available rather than assuming that more is necessarily better.

To achieve such optimization, it will be necessary to create a more continuous model of assessment that is integrated across the various stages of training and practice, with information being carried forward by the individual, thereby enabling the profession to hone in on the particulars of performance that would be most impactful for a particular learner at a particular time and would promote the notion that learners (trainees and practitioners) need to be accountable for their learning (van der Vleuten and Schuwirth 2005). An assessment system that recognizes the continuous nature of performance, as opposed to dichotomizing into pass-fail, would further normalize this process such that all learners would be expected to maintain ownership over a learning plan that could efficiently guide their activities while minimizing threats to the candidates' self-concept, an important determinant of performance (Eva et al. 2012; Kluger and van Dijk 2010). Thinking this way about competent practice would not remove the need to develop strategies for external observation (by supervisors or peers), but it might create a situation in which such opportunities are more rewarding as they can be deliberately directed at the key aspects of performance that would yield palpable benefit while helping learners maintain a sense of ownership over their own learning agenda (Mutabdzic et al. 2015).

To some extent these ideas are translations of the key features model of exam question development (Page and Bordage 1995; Farmer and Page 2005) with "key" now defined by what the learner most requires rather than by the critical next step for the patient in a clinical case. This analogy leads to a proposal to use the information available in any form of assessment to define the key actions that are most appropriate at a particular moment in one's development. For example, with regard to issues of knowledge acquisition, the high quality of summative assessment processes currently mounted by big testing organizations are likely to be more trustworthy compared to assessment protocols generated at local institutions. By contrast, for issues of practice performance and related competencies, the greater number of opportunities for situated and longitudinal observations in workplace and learning institutions will likely make the data from those settings more trustworthy compared to those collected by big testing organizations. As a result, we believe maximal strength will come from integration of efforts across the stakeholder organizations that are responsible for each level of training and practice. Within current assessment practices, it is conceivable that OSCE stations or key features questions could be built such that they require candidates to follow up on an error that was made during a preceding assessment moment (Bordage et al. 2013); to ask for help, guidance, or supervision (Cote and Bordage 2012); to demonstrate use of a clinical decision support; or to summarize a case for an attending, making it clear not only what is understood (Goldszmidt et al. 2013), but what the learner would take away from the situation to direct further formative development that would facilitate better care for patients. This promotes the translation of the assessment experience into a personal inquiry based learning strategy and integrates the idea of using data to make sense of one's experience and frame a plan for improvement. In an ideal world such planning would take place with a coach or peer support (Marsh and Roche 1997).

Using the resulting information to tailor components of subsequent assessment processes that require the candidate to demonstrate how they have utilized previous experiences to their patients' benefit would create considerable incentive for candidates to "nurture" their learning as a lifelong effort requiring continuous reinvestment rather than simply trusting that they know enough because their exams have been passed (Schön 1983). Doing so would require considerable harmonization of assessment practices across many groups. Working towards such a coherent and integrated assessment system, however, would create the potential to overcome negative reactions to assessment practices by establishing a cultural norm and expectation that shifts accountability toward a shared responsibility between learner and system (Galbraith et al. 2008) and harnesses feedback in ways that emphasize the primacy of learning (Eva et al. 2012), which leads us to Theme 2.

Theme 2: Striving to implement quality assurance efforts while promoting performance improvement

Conceptual issues

For assessment of any kind to provide a meaningful mechanism through which individuals can be expected to grow it needs to promote and empower self-regulating aspects of practice (Eva et al. 2010), provide high quality and credible feedback (Sargeant et al. 2011), and deliver support for the candidate (Eva et al. 2012; Marsh and Roche 1997). Perhaps because of this complexity it is commonly believed that an assessment cannot

fulfill the dual purposes of offering summative measurement and formative guidance. The distinction is an important one in that it helps to make sense of the compromises that are appropriate when determining the utility of an assessment protocol in different contexts. However, when treated as an absolute rule, this distinction can be detrimental. It risks absolving training organizations that are responsible for nurturing and supporting learners from serving as effective gatekeepers. It also risks removing responsibility from high stakes testing organizations to attend to assessment for learning. More fundamentally, the assumption that duality of purpose cannot be achieved simply mistakes the reality of the learner's experience. We presume that even the staunchest advocates of the distinction would recognize the adage that assessment drives learning and would, therefore, concede that studying or sitting a summative assessment has a formative influence (Newble and Jaeger 1983; Larsen et al. 2008; Norman et al. 2010). Further, any time one performs a task in which identity is invested there is an aspect of summative judgment even if the assessment is intended to be "purely formative." That judgment may not take the form of a high-stakes, 'pass-fail' decision, but it can certainly have a direct impact on whether an observer deems the performer worth the effort of providing feedback or whether the performer himself deems further improvement to be important (Eva et al. 2010).

Thus the question is not whether an assessment is summative or formative, but is the extent to which summative or formative purposes are foregrounded in the mind of the learner. In this regard, a more relevant continuum is the level of the stakes involved in the judgment (again, in the perception of the person being assessed). Part of the challenge with current assessment practices, especially in high-stakes contexts, is that they are routinely singular events with no effort to facilitate use of the information gained from their administration beyond a summative decision. Further, there is generally no effort directed at follow-up to determine if the information that could be gleaned from the assessment is utilized by the candidate for learning purposes.

Exacerbating the problems associated with the unsophisticated dichotomization of summative and formative testing purposes is the romanticized construction of the "self-regulating professional" as one who will rationally and neutrally accept data and strive to use it to change their own behaviour (Eva and Regehr 2013; Watling et al. 2014; Harrison et al. 2015). Many of the roles (e.g., Scholar, Professional) inherent in modern conceptions of clinical practice incorporate competencies that require individuals to continually reinvest in developing their performance over the course of their careers (Mann et al. 2009; Ericsson 2004). This demands that good data be sought by individual practitioners regarding their current level of practice. Yet, data that conflict with one's self-identity are threatening to the individual recipient (Kluger and van Dijk 2010) and create an experience of cognitive dissonance that can make it easier to discount the data than to determine how to best use them for professional growth (Eva et al. 2012). This is especially true given the confidence that follows increasing experience (Eva 2009). For recipients to be influenced by feedback they must be receptive to it (Shute 2008). For recipients to be receptive to feedback they must deem it credible, not just with respect to its validity, but with respect to believing that it is delivered with the sincere goal of helping the recipient practice better (Sargeant et al. 2011; Galbraith et al. 2011). Achieving such credibility requires more than simply convincing the recipient that the data are psychometrically sound. At the level of the individual, we must offer not just data but also guidance regarding *how* to use external evidence to improve (Marsh and Roche 1997). Culturally, we must normalize the improvement process across the range of performance, because focusing attention only on those at the bottom of the distribution reduces the need for the majority of candidates to pay attention to the data available (Kluger and van Dijk 2010; Butler 1987). Functionally,

we must strive for an integrated and continuous system with shared accountability by focusing beyond point-in-time assessment moments that will inevitably be treated simply as hurdles to be overcome before returning to one's normal stride.

It is rare for trainees or physicians in practice to have better data on which to guide their continued development than they do upon engaging in formal assessment activities organized by institutions that have invested heavily in providing the best possible data. To not consider the use of assessment for performance improvement even in high stakes contexts is, therefore, a considerable missed opportunity. There will inevitably be a tension between quality control and quality improvement goals whenever the threat of high stakes assessment looms over candidates (Kluger and van Dijk 2010). We believe benefit can occur despite such tension, however, with further integration across the continuum of training and with deliberate attention paid to quality improvement as a goal such that one of the things that is "learned" through high-stakes assessments is that the material covered matters and must be considered beyond demonstrating adequate performance at a particular moment. Thus, we offer some initial thoughts regarding how such concepts might be built into current high stakes assessment activities.

Logistical considerations

A burgeoning area of research in recent years is demonstrating the conditions under which testing can have pedagogical value (Larsen et al. 2008; Kromann et al. 2010; Rohrer and Pashler 2010). For example, more frequent testing tends to yield a greater learning effect, especially when the testing format requires constructed responses (e.g., short answers) rather than recognition (e.g., MCQs; Karpicke and Roediger 2008; Kornell and Son 2009). This phenomenon creates a perspective in which shorter, more frequent, lower stakes quizzes become increasingly valuable. While such approaches might seem more feasible for learners still in training, a number of groups, including the American Board of Anesthesiology (ABA), have provided proof of concept for its use with clinicians in practice (2014). The ABA provides questions longitudinally to the diplomate who is then given a period of time to answer, providing an alternate pathway to continuous quality improvement and a new approach to keeping up with medical knowledge. The ABA's efforts provide an excellent example of how to implement a summative assessment that can help to confirm aspects of knowledge that are up-to-date while enabling learning and improvement.

One of the reasons offered against using summative assessments for formative purposes is the cost inherent in generating a high quality assessment exercise. If the gatekeeping function is to be maintained in high stakes assessments, test security is an issue and providing feedback on items may mean radically increasing the pool of quality items available for use. Similarly, if assessment is to be offered more continuously through smaller scale but more frequent testing, this too would likely require an increase in the pool of questions available. However, if the profession truly believes that assessment illuminates a road to improved healthcare, this is an investment worth making. Further, this might become more feasible with the rapid developments of automatic item generation (AIG) processes (Gierl and Lai 2013; Gierl et al. 2012), mitigating test security issues by allowing new tests to be built relatively efficiently. If adequate databases of questions could be built, through AIG and other approaches, several possibilities for the use of testing to support performance improvement could be considered.

The learning associated with this sort of assessment program could be further enhanced with the creation of a test-tailoring platform in which the domains of focus and item

blueprints could be specified for individual trainees and physicians in near real time. Customized tests would both support learning and create habits of engaging in deliberate practice improvement activities. In an ideal world, item databases would be created that would allow practitioners to sync their current scope of practice (based on electronic health records, prescription habits, etc.) to a rubric that defines the item database such that formative tests could be maximized to yield optimal guidance regarding mechanisms of improvement. This is closer to reality now than it was 20 years ago as the amount and quality of data that physicians have available about their practice is increasing (Ellaway et al. 2014). The success of such targeted assessment and feedback could then be tracked through the individual's practice patterns by considering the very patient outcomes that led to the generation of the formative test.

Establishing such an iterative program of assessment, in combination with harmonization across levels of training and across stakeholder organizations, could be still further enhanced through innovative examination protocols for candidates to demonstrate how feedback on areas requiring improvement led to an action plan that formed the basis for additional learning. This would extend the scope of examinations away from the single moment in time in which the candidate is physically present for the exam. An OSCE station, for example, could involve review of data collected from each candidate's actual patient encounters and require them to demonstrate how they have understood their experiences with workplace-based assessments and evaluation of patient outcomes. Similarly, the generation of a "Diagnostic OSCE" late in undergraduate MD training or early in postgraduate training could be used deliberately to identify aspects of performance that would benefit from further development and could form the basis for tailoring subsequent assessment efforts. Ideally this process would be repeated at the end of residency and, in both instances, would allow further exploration of the candidates' conceptualization of practice (Bogo et al. 2011) while enabling the stakes of any given assessment to be lowered because motivation would come from the need to demonstrate follow-up rather than the need to pass the exam. How to implement such processes in a manner that will be deemed authentic, and therefore used, is the focus of Theme 3.

Theme 3: Authentically linking assessment and practice

Conceptual issues

To be maximally effective as an educational tool, any system of assessment should model the realities of practice as closely as possible. Such alignment increases acceptability and makes claims of validity much more credible (Bernabeo et al. 2013). Authenticity does not mean using high fidelity simulation to mimic practice (Norman et al. 2012). Rather, authenticity is achieved when assessment protocols accurately reflect the domain of practice such that "studying to the test" or learning to "game the system" equates with learning to practice well. Too often we hear statements from clinical preceptors to their trainees along the lines of "in reality I would do X, but for your exam you should do Y." Such disconnects threaten to undermine the entire system and create a culture in which assessments are viewed merely as hurdles to be overcome to prove oneself competent. They encourage trainees to practice in a manner that is misaligned with reality and encourage educational programs to teach to the exam rather than to learners for the benefit of patients.

Although this section is the shortest in this reflections piece, this aspect of our deliberations may be the most fundamental to enabling the cultural shift encouraged by efforts to broaden the definition of good medical practice. It is important not only that assessment processes accurately reflect the aspects of practice that stakeholders desire to promote, but assessment candidates should be able to express an understanding of why their behavior might differ in "typical" practice or why their behaviour might be variable within their practice. In other words, it might be appropriate for assessment-driven behavior to deviate from one's normal practice because context matters. Practicing in rural and remote areas, for example, will not be the same as practicing in large urban academic tertiary care centres and assessment practices should provide some sense of whether or not candidates demonstrate appropriate (i.e., safe) awareness of variation in their practice. Having candidates state their broader conceptualizations of practice by indicating why their practice might change from one context to the next might give assessors better information regarding the driving motivation for behaviours observed. It is only by marrying abstract standards of practice with meaningful understanding of local variability that assessment can be truly authentic in the eyes of the individual being assessed.

Workplace-based assessment practices such as in-training evaluation reports, mini-CEXs, patient reported outcomes, and direct observation of procedural skills (Pugh et al. 2014; Kogan and Holmboe 2013; Kogan et al. 2009; Norcini and Burch 2007; Norcini 2005; Norcini et al. 2003; Galbraith et al. 2011), in many ways represent the next frontier of assessment technology. They are not currently part of the most high stakes assessment activities despite their potential for assessing a greater variety of dimensions of practice and their capacity to better reflect what individuals actually do in their day-to-day activity. Of course, there remain concerns about standardization and reliability with most workplace-based assessment practices, but collection of data over many evaluators, rotations, and cases does tend to yield sufficient reliability (Ginsburg et al. 2013) and uniformity of opinion may not be the ultimate goal in all contexts (Gingerich et al. 2014).

Logistical considerations

At its root, making assessment authentic requires having candidates engage with clinical scenarios that are not clear and obvious cut-outs from a blueprint. In in vivo, work-based, situations such as using practice data, peer review, or portfolios, generating "authentic" assessment would seem straightforward as the data are by definition based on the individual's practice. When the stakes are high and momentary, however, even one's personal "reflections" can become fictional when the system encourages them to be written for external review (Hays and Gay 2011). We see value, therefore, in leaving control of learner portfolios (Galbraith et al. 2008) in the hands of the learner to engender a sense of accountability and responsibility (van Tartwijk and Driessen 2009) while also enabling deliberate exploration of practice patterns, successes, and concerns without fear of the repercussions that can arise from placing great weight on any one assessment moment. This view is reinforced by recent literature examining physicians' engagement with Practice Improvement Modules, which has suggested that physicians are more likely to believe the data available because they collected it (Bernabeo et al. 2013). Placing the emphasis on learning from workplace-based assessment in a manner that accumulates towards a higher stakes decision, as advocated by van der Vleuten and Schuwirth (2005), is important to engage communities of practitioners to discuss and learn from each other the reasons why their practice might deviate from that of others and when or if such differences are safe and appropriate.

In ex vivo assessment situations such as OSCEs, the cases must allow uncertainty and avoid prompting statements such as "here comes the breaking bad news station." Doing so might involve allowing stations with multiple pathways even at the cost of absolute standardization (Hodges 2003). Models are being developed, as many groups have been experimenting with sequential OSCE processes in which later stations revisit previously encountered patients at an ostensibly later point in time, offering test results that were not available previously or focusing on some other form of follow-up visit (MacRae et al. 1997; Hatala et al. 2011). Within station, it is also conceivable that standardized patients could be trained to offer information midway through a case that contradicts the most apparent diagnosis from the early portion of the encounter. Doing so would further provide some indication of candidates' capacity to overcome their first impressions and avoid falling prey to premature closure (Eva and Cunnington 2006).

At the same time, there is a tendency to infer the cause of behaviours, trusting that the right things, when done, were done for the right reasons (Ginsburg et al. 2004). Given that context influences performance there might be value in establishing opportunities for examiners to explore the reasoning underlying candidates' behaviour (Bogo et al. 2011; Kogan et al. 2011; Williams et al. 2014). This could be done through post-encounter probes that are akin to debriefing sessions post simulation encounters in that both require the candidate to explain why certain things were done (Williams et al. 2014), why alternative actions were ruled out, and if or how decision-making might have changed if the context had differed in specified ways. Gaining a better understanding of candidates' conceptualizations of practice might help to account for some of the apparent inconsistency both in behaviour and in rater perception of that behaviour while also enabling a strategy for assessing other aspects of competence embedded within the Scholar and Professional roles (Bogo et al. 2011). Whether or not any of these innovations prove effective in adequately measuring candidate performance, movement beyond simple "examine the knee" types of cases is thought necessary to enable assessment of holistic aspects of practice that avoid the atomization of medical practice that the specification of a series of competencies can create (Eva and Hodges 2012).

Systemic considerations

In offering these reflections we fully recognize that good assessment is time and resource intensive. Given the generic nature of the issues raised here, without specific focus on any one setting or level of training, it is impossible to specify with any precision the cost of the concepts outlined. We do know, however, that a considerable sum of money is already spent on assessment that might well be diverted in ways that would better align assessment processes with the ideas we are exploring, especially the quality, safety, and improvement of patient care. We are also aware that many of the issues we raise are cultural in their roots and that changes in the direction we are suggesting might be threatening to many stakeholders. However, any assessment process will be threatening and can create a source of frustration or even outright rebellion amongst practitioners (as recently seen in relation to Maintenance of Certification practices in the US). We suggest that this is more likely to be the case if the assessment community continues to emphasize summative processes based on a 'standard practice' that does not exist, thereby creating an unnatural, high stakes test of competence. It seems antithetical to the very reasoning behind assessment (the protection of patients) to suggest that we should not think about how to improve current

assessment practices, not only in terms of their role in gatekeeping but also in terms of their opportunities for shaping further learning. Leadership is called for now, just as it was when substantial funds were devoted to the development of Multiple Choice Question technology from the 1950s onward. Such leadership will only be achieved through effective collaboration across educational and testing organizations and with providers of continuing education services to enable practices and expectations of healthcare practitioners to be established early in training.

A common criticism of the medical training system is the sharp transitions experienced when moving from pre-clerkship to clerkship, from clerkship to postgraduate training, and from postgraduate training to practice (Jarvis-Selinger et al. 2012; Teunissen and Westerman 2011). Some degree of transition pain is inevitable, but the challenges might be reduced by efforts to create a cohesive system of assessment. Enabling supervisors, mentors, program directors, and colleges to receive high quality information regarding each individual's relative strengths and weaknesses such that further educational opportunities might be crafted efficiently would maximize the opportunity to resolve any issues even as the next stage of training is experienced and the process of discovery begins anew.

Finally, it will be important that these changes are understood to be value added to the recipients of such assessment and feedback. Encouraging active engagement will require a reward structure that allows data and candidate responses to be recognized as evidence that continuing professional development is being undertaken and that the candidate experiences authentically reflect the practice in which physicians are engaged, thereby having clear relevance to their patients. Thus, we do not see this process of change as a top down exercise that is imposed upon trainees and practitioners but rather as a co-productive collective exercise that truly engages learners in benefiting patients.

Summary

Conceptions of best practice in health professional assessment are evolving away from simply focusing on "knows how and shows how" processes towards processes that catalyze quality improvement and patient safety. There is growing availability of more robust and timely performance measurement through local Electronic Medical Records and large clinical databases. These forms of information are calling into question the exclusive reliance on traditional assessment approaches thanks to their potential to provide a real-time authentic "window" into a physician's practice. As thinking and data sources evolve, we consider the following to be a list of common issues facing any individuals concerned with mounting high quality assessment practices:

1. Broadening the base of assessment beyond knowledge tests;
2. Rigorously focusing data collection and decision-making practices in a manner that enables the assessment body to draw relevant and meaningful inferences;
3. Adding emphasis on healthcare processes and outcomes, including strengthening of the ability of the assessments to predict who will perform well against those outcomes and who will further develop in their ability after training;
4. Building a coherent and integrated system of assessment across the continuum of training to practice;
5. Emphasizing the primacy of learning as an integral part of assessment;
6. Harnessing the power of feedback; and

7. Shifting accountability towards a model of shared responsibility between the individual and the educational system.

Continuing the evolution of assessment practices in the manner outlined here will require time, energy, and resources. However, patient safety challenges and the licensing and certification of physicians are not going to stop while these issues are resolved. As such, we would advocate for engaging a set of pilot efforts aimed at quickly determining the feasibility of multiple strategies that might facilitate movement in the desired direction rather than waiting for a comprehensive system of assessment to be designed and investing heavily in a complete infrastructure, the components of which will undoubtedly be variable in effectiveness. While there are no simple answers, there are many testing organizations working towards determining how to offer authentic, tailored, and meaningful assessment practices for professional regulation (e.g., Eva et al. 2013; Hawkins et al. under review; Krumholz et al. under review). Fundamental to all of these efforts is that we avoid confusing quality assurance with quality improvement, reliability with usefulness, precision of measurement with being actionable and that we avoid confusing the desire on the part of practitioners to practice well with the desire to be told how they are doing (Mann et al. 2011).

Acknowledgments This work was supported by the Medical Council of Canada (MCC) through the work of the authors as members of the Medical Education Assessment Advisory Committee. The focus was not constrained to MCC practices, however, and the content of the paper does not necessarily reflect MCC policy.

References

ABA: American Board of Anesthesiology. (2014). MOCA minute. http://www.theaba.org/MOCA/MOCA-Minute. Last accessed November 2, 2015.

AFMC: Association of Faculties of Medicine in Canada. (2010). The Future of Medical Education in Canada (FMEC): A collective vision for MD education. Retrieved from http://www.afmc.ca/fmec/pdf/collective_vision.pdf.

Bernabeo, E., Hood, S., Iobst, W., Holmboe, E., & Caverzagie, K. (2013). Optimizing the implementation of practice improvement modules in training: Lessons from educators. *Journal of Graduate Medical Education, 5*(1), 74–80.

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

Bogo, M., Regehr, C., Logie, C., et al. (2011). Adapting objective structured clinical examinations to assess social work students' performance and reflections. *Journal of Social Work Education, 47*, 5–18.

Bordage, G., Meguerditchian, A. N., & Tamblyn, R. (2013). Avoidable adverse events: A content analysis of a national qualifying examination. *Academic Medicine, 88*, 1493–1498.

Boud, D., & Molloy, E. (Eds.). (2013). *Feedback in higher and professional education: Understanding it and doing it well*. London: Routledge.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*, 474–482.

Cadieux, G., Tamblyn, R., Dauphinee, D., & Libman, M. (2007). Predictors of inappropriate antibiotic prescribing among primary care physicians. *CMAJ, 177*(8), 877–883.

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

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justification on test use. *Psychological Methods, 17*, 31–43.

Colliver, J. A. (2002). Educational theory and medical education practice: A cautionary note for medical school faculty. *Academic Medicine, 77*(12), 1217–1220.

Cook, D. A. (2014). When I say… validity. *Medical Education, 48*(10), 948–949.
Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*, 560–575.
Cote, L., & Bordage, G. (2012). Content and conceptual frameworks of preceptor feedback in response to residents' educational needs. *Academic Medicine, 87*(9), 1274–1281.
Cruess, R., & Cruess, S. (2014). Updating the Hippocratic Oath to include medicine's social contract. *Medical Education, 48*(1), 95–100.
Custers, E. (2010). Long-term retention of basic science knowledge: A review study. *Advances in Health Sciences Education, 15*(1), 109–128.
Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*, 830–837.
Ellaway, R. H., Pusic, M. V., Galbraith, R. M., & Cameron, T. (2014). Developing the role of big data and analytics in health professional education. *Medical Teacher, 36*(3), 216–222.
Ericsson, K. A. (2004). Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Academic Medicine, 79*, S70–S81.
Eva, K. W. (2002). The aging physician: Changes in cognitive processing and their impact on medical practice. *Academic Medicine, 77*, S1–S6.
Eva, K. W. (2003). On the generality of specificity. *Medical Education, 37*, 587–588.
Eva, K. W. (2009). Diagnostic error in medical education: Where wrongs can make rights. *Advances in Health Sciences Education, 14*, 71–81.
Eva, K. W., Bordage, G., Campbell, C., Galbraith, R., Ginsburg, S., Holmboe, E., & Regehr, G. (2013). Medical Education Assessment Advisory Committee report to the Medical Council of Canada on Current Issues in Health Professional and Health Professional Trainee Assessment. Retrieved from http://mcc.ca/wp-content/uploads/Reports-MEAAC.pdf.
Eva, K. W., & Cunnington, J. P. (2006). The difficulty with experience: Does practice increase susceptibility to premature closure? *Journal of Continuing Education in the Health Professions, 26*(3), 192–198.
Eva, K. W., & Hodges, B. D. (2012). Scylla or Charbydis? Can we navigate between objectification and judgment in assessment? *Medical Education, 46*, 914–919.
Eva, K. W., Munoz, J., Hanson, M. D., Walsh, A., & Wakefield, J. (2010). Which factors, personal or external, most influence students' generation of learning goals? *Academic Medicine, 85*, S102–S105.
Eva, K. W., & Regehr, G. (2013). Effective feedback for maintenance of competence: From data delivery to trusting dialogues. *CMAJ, 185*, 463–464.
Eva, K. W., Regehr, G., & Gruppen, L. D. (2012). Blinded by 'insight': Self-assessment and its role in performance improvement. In B. D. Hodges & L. Lingard (Eds.), *The question of competence: Reconsidering medical education in the twenty-first century* (pp. 131–154). Ithaca, NY: Cornell University Press.
Farmer, E. A., & Page, G. (2005). A practical guide to assessing clinical decision-making skills using the key features approach. *Medical Education, 39*, 1188–1194.
Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R., et al. (2010). Competency-based medical education: theory to practice. *Medical Teacher, 32*(8), 638–645.
Galbraith, R. M., Clyman, S., & Melnick, D. E. (2011). Conceptual perspectives: Emerging changes in the assessment paradigm. In J. P. Hafler (Ed.), *Extraordinary learning in the workplace* (pp. 87–100). Berlin: Springer.
Galbraith, R. M., Hawkins, R. E., & Holmboe, E. S. (2008). Making self-assessment more effective. *Journal of Continuing Education in the Health Professions, 28*(1), 20–24.
Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education, 47*(7), 726–733.
Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education, 46*(8), 757–765.
Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education, 48*(11), 1055–1068.
Ginsburg, S., Eva, K., & Regehr, G. (2013). Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Academic Medicine, 88*(10), 1539–1544.
Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. *Academic Medicine, 85*(5), 780–786.
Ginsburg, S., Regehr, G., & Lingard, L. (2004). Basing the evaluation of professionalism on observable behaviours: A cautionary tale. *Academic Medicine, 79*(10, Suppl), S1–S4.
Goldszmidt, M., Minda, J. P., & Bordage, G. (2013). What physicians reason about during clinical encounters: Time to be more explicit. *Academic Medicine, 88*(3), 390–394.

Guadagnoli, M., Morin, M. P., & Dubrowski, A. (2012). The application of the challenge point framework in medical education. *Medical Education, 46*(5), 447–453.

Harrison, C. J., Könings, K. D., Schuwirth, L., Wass, V., & van der Vleuten, C. (2015). Barriers to the uptake and use of feedback in the context of summative assessment. *Advances in Health Sciences Education, 20*(1), 229–245.

Hatala, R., Marr, S., Cuncic, C., & Bacchus, C. M. (2011). Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance. *BMC Medical Education, 11*, 23.

Hawkins, et al. (under review). The ABMS MOC Part III examination: Value, concerns and alternative formats.

Hays, R., & Gay, S. (2011). Reflection or 'pre-reflection': What are we actually measuring in reflective practice? *Medical Education, 45*(2), 116–118.

Hodges, B. (2003). OSCE! variations on a theme by Harden. *Medical Education, 37*(12), 1134–1140.

Holmboe, E. S., Sherbino, J., Long, D. M., Swing, S. R., & Frank, J. R. (2010). The role of assessment in competency-based medical education. *Medical Teacher, 32*(8), 676–682.

James, J. T. (2013). A new, evidence-based estimate of patient harms associated with hospital care. *Journal of Patient Safety, 9*(3), 122–128.

Jarvis-Selinger, S., Pratt, D. D., & Regehr, G. (2012). Competency is not enough: integrating identity formation into the medical education discourse. *Academic Medicine, 87*(9), 1185–1190.

Kane, M. T. (1992). An argument-based approach to validation. *Psychological Bulletin, 112*, 527–535.

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

Kennedy, T. J., Regehr, G., Baker, G. R., & Lingard, L. A. (2009). 'It's a cultural expectation…' The pressure on medical trainees to work independently in clinical practice. *Medical Education, 43*(7), 645–653.

Klass, D. A. (2007). Performance-based conception of competence is changing the regulation of physicians' professional behavior. *Academic Medicine, 82*(6), 529–535.

Kluger, A. N., & van Dijk, D. (2010). Feedback, the various tasks of the doctor, and the feedforward alternative. *Medical Education, 44*, 1166–1174.

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. S. (2011). Opening the black box of postgraduate trainee assessment in the clinical setting via observation: A conceptual model. *Medical Education, 45*, 1048–1060.

Kogan, J. R., & Holmboe, E. (2013). Realizing the promise and importance of performance-based assessment. *Teaching and Learning in Medicine, 25*(Suppl 1), S68–S74.

Kogan, J. R., Holmboe, E. S., & Hauer, K. R. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302*, 1316–1326.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.). (1999). *To err is human: building a safer health system*. Washington, DC: National Academy Press, Institute of Medicine.

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

Kromann, C. B., Bohnstedt, C., Jensen, M. L., & Ringsted, C. (2010). The testing effect on skills learning might last 6 months. *Advances in Health Sciences Education, 15*(3), 395–401.

Krumholz, et al. (under review). Recommendations to the American Board of Internal Medicine (ABIM): A vision for certification in internal medicine in 2020.

Larsen, D. P., Butler, A. C., & Roediger, H. L, 3rd. (2008). Test-enhanced learning in medical education. *Medical Education, 42*(10), 959–966.

MacRae, H. M., Cohen, R., Regehr, G., Reznick, R., & Burnstein, M. (1997). A new assessment tool: the patient assessment and management examination. *Surgery, 122*(2), 335–343.

Mann, K., Gordon, J., & MacLeod, A. (2009). Reflection and reflective practice in health professions education: A systematic review. *Advances in Health Sciences Education, 14*(4), 595–621.

Mann, K. V., van der Vleuten, C., Eva, K., Armson, H., Chesluk, B., Dornan, T., et al. (2011). Tensions in informed self-assessment: How the desire for feedback and reticence to collect and use it conflict. *Academic Medicine, 86*, 1120–1127.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187–1197.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: American Council on Education and Macmillan.

Morcke, A. M., Dornan, T., & Elka, B. (2013). Outcome (competency) based education: an exploration of its origins, theoretical basis and empirical evidence. *Advances in Health Sciences Education, 18*, 851–863.

Mutabdzic, D., Mylopoulos, M., Murnaghan, M. L., Patel, P., Zilbert, N., Seemann, N., et al. (2015). Coaching surgeons: Is culture limiting our ability to improve? *Annals of Surgery, 262*(2), 213–216.

Mylopoulos, M., & Regehr, G. (2011). Putting the expert together again. *Medical Education, 45*(9), 920–926.
Mylopoulos, M., & Scardamalia, M. (2008). Doctors' perspectives on their innovations in daily practice: implications for knowledge building in health care. *Medical Education, 42*(10), 975–981.
Neve, H., & Hanks, S. (2016). When I say … capability. *Medical Education, 50* (in press).
Newble, D. I., & Jaeger, K. (1983). The effect of assessments and examinations on the learning of medical students. *Medical Education, 17*(3), 165–171.
Newell, K. M., Liu, Y., & Mayer-Kress, G. (2001). Time scales in motor learning and development. *Psychological Review, 108*, 57–82.
Norcini, J. J. (2005). Current perspectives in assessment: The assessment of performance at work. *Medical Education, 39*(9), 880–889.
Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., et al. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 conference. *Medical Teacher, 33*(3), 206–214.
Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: A method for assessing clinical skills. *Annals of Internal Medicine, 138*(6), 476–481.
Norcini, J., & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher, 29*(9), 855–871.
Norman, G., Dore, K., & Grierson, L. (2012). The minimal relationship between simulation fidelity and transfer of learning. *Medical Education, 46*(7), 636–647.
Norman, G., Neville, A., Blake, J. M., & Mueller, B. (2010). Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. *Medical Teacher, 32*(6), 496–499.
Norman, G. R., Norcini, J., & Bordage, G. (2014). Competency-based education: Milestones or millstones. *Journal of Graduate Medical Education, 6*(1), 1–6.
Page, G., & Bordage, G. (1995). The medical council of Canada's key feature project: A more valid written exam. of clinical decision-making skills. *Academic Medicine, 70*, 104–110.
Pugh, D., Hamstra, S. J., Wood, T. J., Humphrey-Murto, S., Touchie, C., Yudkowsky, R., Bordage, G. (2014). A procedural skills OSCE: Assessing technical and non-technical skills of internal medicine residents. Advances in health sciences education. Retrieved from http://link.springer.com/article/10.1007/s10459-014-9512-x?sa_campaign=email/event/articleAuthor/onlineFirst.
Razack, S., Hodges, B., Steinert, Y., & Maguire, M. (2015). Seeking inclusion in an exclusive process: Discourses of medical school student selection. *Medical Education, 49*, 36–47.
RCPSC: Royal College of Physicians and Surgeons of Canada. (2011). Assessment strategies within the revised maintenance of certification program, draft recommendations.
Regehr, G. (1994). Chickens and children do not an expert make. *Academic Medicine, 69*, 970–971.
Regehr, G., Eva, K., Ginsburg, S., Halwani, Y., & Sidhu, R. (2011). Future of medical education in Canada postgraduate project environmental scan. Paper 13. Assessment in postgraduate medical education: Trends and issues in assessment in the workplace. Retrieved from http://www.afmc.ca/pdf/fmec/13_Regehr_Assessment.pdf.
Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Research, 38*, 406–412.
Sargeant, J., Eva, K. W., Armson, H., Chesluk, B., Dornan, T., Holmboe, E., et al. (2011). Features of assessment learners use to make informed self-assessments of clinical performance. *Medical Education, 45*, 636–647.
Schön, D. (1983). *The reflective practitioner: How professionals think in action*. London: Temple Smith.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*, 153–189.
Swanson, D., & Roberts, T. (2016). Trends in national licensing examinations. *Medical Education, 50*(1) (in press).
Tamblyn, R., Abrahamowicz, M., Dauphinee, D., et al. (2007). Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA, 298*(9), 993–1001.
Teunissen, P. W., & Westerman, M. (2011). Opportunity or threat: The ambiguity of the consequences of transitions in medical education. *Medical Education, 45*(1), 51–59.
van der Vleuten, C. (1996). The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education, 1*, 41–67.
van der Vleuten, C. P., & Schuwirth, L. W. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309–317.
van Tartwijk, J., & Driessen, E. W. (2009). Portfolios for assessment and learning: AMEE Guide no. 45. *Medical Teacher, 31*(9), 790–801.
Watling, C., Driessen, E., van der Vleuten, C. P., & Lingard, L. (2014). Learning culture and feedback: An international study of medical athletes and musicians. *Medical Education, 48*(7), 713–723.

Wenghofer, E., Klass, D., Abrahamowicz, M., et al. (2009). Doctor scores on national qualifying examinations predict quality of care in future practice. *Medical Education, 43*(12), 1166–1173.

Williams, R. G., Klamen, D. L., Markwell, S. J., Cianciolo, A. T., Colliver, J. A., & Verhulst, S. J. (2014). Variations in senior medical student diagnostic justification ability. *Academic Medicine, 89*(5), 790–798.