Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

Core principles of assessment in competency-based medical education

Jocelyn Lockyer, Carol Carraccio, Ming-Ka Chan, Danielle Hart, Sydney Smee, Claire Touchie, Eric S. Holmboe, Jason R. Frank & on behalf of the ICBME Collaborators

To cite this article: Jocelyn Lockyer, Carol Carraccio, Ming-Ka Chan, Danielle Hart, Sydney Smee, Claire Touchie, Eric S. Holmboe, Jason R. Frank & on behalf of the ICBME Collaborators (2017) Core principles of assessment in competency-based medical education, Medical Teacher, 39:6, 609-616, DOI: 10.1080/0142159X.2017.1315082

To link to this article: https://doi.org/10.1080/0142159X.2017.1315082

Published online: 09 Jun 2017.

MEDICAL TEACHER, 2017
VOL. 39, NO. 6, 609–616
http://dx.doi.org/10.1080/0142159X.2017.1315082

Core principles of assessment in competency-based medical education

Jocelyn Lockyer[a], Carol Carraccio[b], Ming-Ka Chan[c], Danielle Hart[d,e], Sydney Smee[f], Claire Touchie[f,g], Eric S. Holmboe[h] and Jason R. Frank[i,j]; on behalf of the ICBME Collaborators

[a]Cumming School of Medicine, University of Calgary, Calgary, Canada; [b]American Board of Pediatrics, Chapel Hill, NC, USA; [c]Max Rady College of Medicine, Rady Faculty of Health Sciences, University of Manitoba, Winnipeg, Canada; [d]Hennepin County Medical Center, Minneapolis, MN, USA; [e]University of Minnesota Medical School, Minneapolis, MN, USA; [f]Medical Council of Canada, Ottawa, Canada; [g]Faculty of Medicine, University of Ottawa, Ottawa, Canada; [h]Accreditation Council for Graduate Medical Education, Chicago, IL, USA; [i]Royal College of Physicians and Surgeons of Canada, Ottawa, Canada; [j]Department of Emergency Medicine, University of Ottawa, Ottawa, Canada

ABSTRACT

The meaningful assessment of competence is critical for the implementation of effective competency-based medical education (CBME). Timely ongoing assessments are needed along with comprehensive periodic reviews to ensure that trainees continue to progress. New approaches are needed to optimize the use of multiple assessors and assessments; to synthesize the data collected from multiple assessors and multiple types of assessments; to develop faculty competence in assessment; and to ensure that relationships between the givers and receivers of feedback are appropriate. This paper describes the core principles of *assessment for learning* and *assessment of learning*. It addresses several ways to ensure the effectiveness of assessment programs, including using the right combination of assessment methods and conducting careful assessor selection and training. It provides a reconceptualization of the role of psychometrics and articulates the importance of a group process in determining trainees' progress. In addition, it notes that, to reach its potential as a driver in trainee development, quality care, and patient safety, CBME requires effective information management and documentation as well as ongoing consideration of ways to improve the assessment system.

Introduction

A major challenge in implementing competency-based medical education (CBME) is the meaningful assessment of competence. The shift to CBME has raised awareness of the limitations of existing assessment methods (see Harris et al. 2017, in this issue) and the need to develop strategies to assess the competencies expected of today's physicians in an era characterized by increasing interdependence among health care professionals, the recognition that patient safety is everyone's responsibility, and an expectation of transparency and accountability.

In designing assessment programs, it is critical to articulate its purpose. Two fundamental and yet essentially different rationales are *assessment of learning* and *assessment for learning*. Before the introduction of CBME, the former was emphasized; however, as CBME becomes established, the focus is shifting to *assessment for learning*. Van der Vleuten et al. suggest that "whenever assessment becomes a goal in itself, it is trivialized and will ultimately be abandoned. Assessment has utility insofar as it succeeds in driving learning, is integrated in a routine and ultimately comes to be regarded as indispensable to the learning practice." (2010, p. 712). Thus, if the primary purpose in assessment in CBME is to drive learning, and our secondary purpose is to make judgments about readiness to progress, we need to design assessment programs accordingly (van der Vleuten et al. 2012). *Assessment for learning* aligns with other foundational principles of CBME, including active trainee involvement in learning and assessment, the creation of an authentic environment for learning and assessment, the use of direct observation, and an emphasis on formative feedback. *Assessment of learning* aligns with the continuing need to gauge progress against targeted outcomes and criterion-referenced standards (Carraccio et al. 2002).

In a plea for new psychometric models, Schuwirth and van der Vleuten (2006) proposed that, rather than asking only whether a learner has achieved a predetermined outcome, we pose a more critical question: How big is the risk of the student performing below the standard in a future case given his or her history and the current observation?

Practice Points

- Competency-based medical education (CBME) relies on a program of assessment that includes multiple methods and multiple assessors and is embedded within an effective educational system.
- Assessment for learning plays a prominent role in CBME, since formative feedback is an essential element of developing competence.
- Faculty development to create a shared mental model of required learner behavior and expected levels of performance is foundational to CBME.
- Variance in assessor rating is not all attributable to error; some variance reflects a different lens through which an assessor sees a learner.
- The assessment instrument is primarily the individuals who conduct the assessment, rather than the tools and forms they use. As such, individuals using assessment tools and forms need training.

CONTACT Jocelyn Lockyer lockyer@ucalgary.ca Department of Community Health Sciences, TRW Building, Floor 3, 3280 Hospital Dr NW, Calgary, AB T2N 4Z6, Canada

Table 1. Overview of assessment methods aligned with Miller's pyramid[a].

Stage	Corresponding methods to assess performance
Does	Chart/electronic medical record review (e.g. medical decision-making, cost-effective care, documentation)
	"Collective perspective"/clinical competency committee/residency education committee decisions
	Direct observation in clinical environments
	Efficiency data
	End-of-rotation evaluations
	Multi-source feedback/360-degree
	Patient outcomes data, including patient-reported outcome measures
	Portfolio[b]
	Procedure or case log with reflection and/or assessment
	Product review (e.g. splint, laceration repair)
	Project review (e.g. evidence-based medicine project, quality improvement project)
	Video review from clinical environments
Shows how	Objective structured clinical examination/standardized patient encounter
	Oral case presentation
	Simulated case
	Skills station
	Virtual reality/computerized patient management problem
Knows how	Chart-stimulated recall
	Development of individualized learning plan
	Mock oral boards examination/progressive case disclosure
	Oral questioning targeting patient management
	Written assignment/essay test
Knows	Multiple-choice questionnaire/short-answer test/audience response system
	Oral questioning targeting fact recall

[a]Adapted from Nyquist (2014); Hawkins & Holmboe (2017).
[b]Portfolios may fall under various stages of Miller's pyramid, depending on what is included.

Educational systems need to maximize the probability that a physician graduating from residency training can provide safe, effective, patient-centered care (Holmboe et al. 2004; Norcini et al. 2011; Kogan et al. 2014). To meet this purpose, the elements of an assessment program include actions (collecting, synthesizing, interpreting, and assigning weight to information); support (faculty development and psychometric analysis); documentation (rules, blueprints, and information management); improvement (with regard to research, development, the learning environment, and change management); and accountability (with regard to scientific research, external review, cost effectiveness, and political and legal requirements) (Dijkstra et al. 2010).

Moreover, assessment in CBME should not end with residency training. Eva et al. suggest we "prioritize continuous professional development in a manner that enhances a shared model of responsibility/accountability between practitioners and educational programs/testing organizations" (2013, p. 3). Competence is not something one can attain once and for all: there will always be another context or occasion that necessitates reassessment (Eva et al. 2013).

Our vision for meaningful competency-based assessment should include (1) timely ongoing assessments, with comprehensive periodic reviews to ensure continued progress (Archer 2010); (2) the best use of multiple assessors and assessments to enable the right assessment to be made at the right time for the right purpose, while avoiding assessor fatigue (Norcini et al. 2011; Hodges 2013); (3) a synthesis of data collected through group processes to reach judgments about competence; (4) faculty development for all assessors, who, as observers of trainees in the workplace, are the true measurement instrument; and (5) optimized relationships between the givers and receivers of formative feedback to enhance the incorporation of feedback into practice (Watling et al. 2012).

If these emerging issues are not addressed, we risk creating assessment systems that are burdensome and uninformative. This paper examines current recommendations from the literature relating to assessment and assessors. First, we describe the core principles of assessment *for* and *of* learning. Then, we describe how assessment can be optimized through multiple methods and multiple assessments, assessor selection and training, a reconceptualization of the role of psychometrics, and a recognition of the importance of group processes. Finally, we discuss information management and documentation and ways to improve assessment programs.

Core assessment principles of CBME

The first step in planning CBME assessments is to determine what information is necessary to ascertain whether goals are being met. Given that the two goals – *assessment for learning* and *assessment of learning* – are different, so too are the information management strategies for each.

Assessment for learning

Miller (1990) identified four levels of learning, conceptualized as a pyramid. Beginning at the base, the learner "knows," and then proceeds through "knows how" and "shows how" before reaching the apex, "does" (Table 1). The assessment strategies tied to each level inform and contribute to learning as well as assessment, provided that formative feedback is given. At the "does" level, assessment becomes part of the authentic context in which one works and learns; learning provides deeper meaning for the trainee and builds a substrate for the cognitive processes of clinical decision-making (Eva 2005).

The active engagement of learners in their own learning has long been understood as crucial to developing skills in lifelong learning (Dewey 1974; Knowles 1975). Assessment should be performed *by* and *with* the learner. Two strategies that embody this principle are informed self-assessment, whereby the learner is encouraged to draw on data from credible external as well as internal sources to guide learning (Sargeant et al. 2010), and the use of portfolios, which encourage learners to document and reflect

on their learning (van Tartwijk & Driessen 2009). Both strategies can have a significant impact on a trainee's ability to improve performance.

It is difficult to accurately assess oneself (Eva & Regehr 2007). However, when self-assessment involves reflection, particularly "reflection-in-action," it allows the learner to know when to stop and ask for feedback or help (Eva & Regehr 2005). This behavior is termed "self-directed assessment seeking" (Eva & Regehr 2008). Pelgrim et al. (2013) demonstrated the connection between the specificity of the feedback given by faculty and the subsequent specificity of reflections by learners, and showed that this alignment promotes the formulation of action plans, which the authors used as a proxy for the incorporation of feedback into practice. Sargeant et al. (2011) also found that informed self-assessment, especially when combined with feedback, can be a powerful catalyst for professional growth.

Likewise, the fact that CBME and portfolio assessment share certain principles creates synergies when portfolios are used to assess competencies. Here we use a broad definition of a portfolio as a framework and process for collecting, analyzing, and documenting the successful acquisition of competence and performance (Holmboe et al. 2008). Fundamental to both is the active engagement of the learner in the process, leading to assessment as the "teachable moment." The use of portfolios is best conceived in an active sense: "to portfolio" (Friedman Ben David et al. 2001). Moreover, portfolios, particularly electronic versions, contribute to both the effectiveness and the efficiency of information management by stimulating reflection and informed self-assessment, providing a longitudinal view of learner development and organizing the myriad of assessments from multiple assessors using multiple tools.

Assessment of learning

Training outcomes must now go beyond the traditional domains and encompass a broad range of abilities as captured in competency frameworks such as the CanMEDS Roles (Frank & Danoff 2007), the Accreditation Council for Graduate Medical Education (ACGME) Core Competencies (Swing 2007), or Good Medical Practice (General Medical Council 2013). This presents new challenges for assessment. Further, Kogan and Holmboe (2013) and Hodges (2013) recommend expanding assessment beyond single patient–provider encounters to embrace competencies such as population care and teamwork.

Traditionally, assessment has focused on educational outcomes such as the acquisition of knowledge or the demonstration of certain competencies in controlled settings. With CBME comes a shift to work-based assessment, and our thinking must shift to assessments that take into account the impact of trainees' competence on the quality of care provided to the patient (Kogan & Holmboe 2013). Medical education will need to embrace a continuous quality-improvement process to ensure that innovation in education leads not only to improved learner outcomes but also to better patient care, the latter being the ultimate goal.

Optimizing assessment

Optimizing an assessment program in the era of CBME will require (1) multiple methods; (2) multiple assessors; (3) the selection and training of assessors; (4) a reconceptualization of the role of psychometrics; and (5) a recognition of the importance of group process in reaching critical decisions about competence.

Multiple methods

Various assessment modalities are possible in CBME. The information sought, the level of performance, the learner's stage within Miller's pyramid (Miller 1990), and the institution's capabilities can all influence the choice of assessment technique.

An assessment program should collect information purposefully, using both structured and unstructured measures; it should value quantitative and qualitative data and ensure that the richness and rigor of the data used align with the stakes of the decision being made (Schuwirth & Ash 2013). A comprehensive program must include non-standardized methods if it hopes to gather information that supports inferences about future real-world practice (van der Vleuten et al. 2012).

Because all assessment methods have their limitations, multiple methods are needed to compensate for the shortcomings of any one technique (van der Vleuten 1996). Similarly, using quantitative and qualitative data in combination can bring greater meaning to learner assessment. Traditionally, the focus has been on quantitative data, which were equated with objectivity and reliability (sometimes at the expense of real-world validity). However, qualitative methods of assessment are rigorous, provided they incorporate strategies to establish the trustworthiness of the data (van der Vleuten et al. 2010); thus, work-based assessments, which rely on qualitative data, can be both defensible and desirable. To realize the promise of CBME, medical educators and training programs will need to embrace the "messiness" of work-based assessment and its reliance on qualitative data.

Multiple assessors

Just as we need multiple methods of assessment to compensate for the shortcomings of any one method, so do we need multiple assessors to compensate for rater shortcomings such as biases, halo effects, and leniency. In the past, such shortcomings have shifted assessment strategies away from expert global judgments and toward more "reliable" checklists, such as those used in observed structured clinical examinations. However, subsequent comparisons of expert judgments with checklists yielded the surprising finding that the former were more reliable (Regehr et al. 1998).

As our understanding of the value of expert opinion has advanced, a growing body of literature is focusing on the unexpected variance in rater judgment, previously attributed to "noise," that occurs when two raters witness the behavior of one individual in the same encounter. Rater training has been shown to be helpful in calibrating raters and in addressing some – but not all – of this variance. Gingerich et al. (2011) postulate that raters spontaneously categorize new individuals on the basis of preformed schemas of interactions with previous individuals, in much the same way as pattern recognition influences clinical

decision-making. These schemas or narratives might not be readily translated into the numerical judgments typically required of most rating scales, thus accounting for some of the unexpected variance. Therefore, a rater-based assessment program that incorporates qualitative assessments may be more effective. Factors such as the time allowed to observe the learner and to complete the rating, as well as the expertise of the rater relative to the content of the assessment, are also important to the outcome (Govaerts et al. 2011; Yeates et al. 2012).

Assessor selection and training

More often than may be acknowledged, assessor selection depends on who is available, who volunteers, who has formal assessment responsibilities, and who can be convinced to perform the assessment for a specific task or event. Those recruited are assumed to have the knowledge of the competencies being assessed by virtue of their medical training and area of practice. Secondary consideration, if any, is given to the assessors' skills as an observer and assessor.

One of the primary reasons to train assessors is to ensure that patients cared for by learners receive safe, effective, patient-centered care (Kogan et al. 2014). When we think about assessment as a domain of expertise (Govaerts et al. 2011), we need to think about how someone becomes a competent assessor. The knowledge required is at least twofold: knowledge of the competencies being assessed (Ponnamperuma 2013), and an understanding of the observational and recording tasks intrinsic to the assessor role (Kogan & Holmboe 2013). A supervising clinician who contributes to the summative assessment of senior learners will need considerable skill in the competencies being assessed and, arguably, should already have experience as an assessor. Moving to CBME will challenge institutions to create educational communities in which assessment is integrated into learning and, moreover, the acquisition of assessment skills is integrated into teaching.

Generally speaking, training provides assessors with the opportunity to become familiar with the goals of assessment and with assessment instruments. Although the need for such training seems clear, how to go about it is not always so. Various approaches to assessor training have been developed (e.g. Woehr & Huffcutt 1994), although these have been researched predominantly in the context of personnel appraisal (Woehr & Huffcutt 1994; Smithers 1998) and assessor training research has been conducted only recently in the context of medical education.

The techniques that show some promise in the field of assessor training are behavioral observation training (BOT), performance dimension training (PDT), and frame of reference training (FoRT). BOT familiarizes assessors with observation processes and improves observation skills by means of three strategies: increased frequency of the observation (the "practice makes perfect" principle); proper preparation for observations; and provision of simple tools to record observations (Holmboe et al. 2008). PDT is an interactive group process that assists assessors in learning and applying behavioral criteria and standards for competencies and is an important precursor of FoRT (Holmboe et al. 2004). FoRT is also an interactive process that seeks to align assessor judgments with a common criterion-based frame of reference to enable assessors to make accurate distinctions between levels of performance (Lievens 2001; Kogan et al. 2014). Both PDT and FoRT rely on the use of case material (video tape review, objective structured teaching examination, etc.) for learning and deliberate practice.

Assessor training must be both feasible and meaningful, and it must be integrated into ongoing faculty development. Training in assessment cannot overcome all the limitations inherent in rater cognition (Gingerich et al. 2011; Yeates et al. 2012; Govaerts & van der Vleuten 2013), and much research into effective rater training is needed. Assessment ability is acquired, not innate; it requires deliberate practice and refinement over time (Berendonk et al. 2013; Govaerts et al. 2013; Kogan & Holmboe 2013). Therefore, one-time training interventions, no matter how appropriate, are insufficient.

Reconceptualizing the role of psychometrics

Two decades ago, the merits of limiting assessment decisions to traditional psychometric approaches came into question. Van der Vleuten (1996) expanded thinking around assessment, defining assessment utility as the product of reliability, validity, cost, practicality, and educational impact. More recently, Norcini et al. (2011) concluded that a "good assessment" should be characterized by validity or coherence; reproducibility or consistency (reliability); equivalence with other assessment approaches; feasibility; acceptability; and a consideration of the educational effect and/or the catalytic effect on learning.

At the same time, traditional thinking about reliability and validity was questioned. Hodges (2013) pointed out that the notion of *subjectivity* had taken on the connotation of *bias,* and that standardization was touted as the ticket to reliability, even though adequate sampling mitigates bias (Eva & Hodges 2012) and is the main determinant of reliability (van der Vleuten & Schuwirth 2005). One can have objective measures (such as standardized checklists) that yield unreliable scores, and subjective measures (such as expert judgments using global rating scales) that provide reliable scores (van der Vleuten 1996). Thinking about validity has also evolved. Validity is no longer seen as an inherent property of a tool. Instead, validity evidence is something that we accumulate, on the basis of the methods and tools that we use, to support the decisions we make. Building evidence of validity is a process that begins with articulating the inference or judgment we want to make. From there, we identify the best evidence needed to support that judgment, collect the evidence using appropriate sampling with multiple methods, and develop faculty members who can effectively use the tools to assess learners and provide them with feedback on their performance (Kane 2013).

Since the introduction of CBME, a common practice has been to reduce competencies to small units of behavior for the purposes of assessment. This "atomization" can lead to trivialization and may actually threaten validity (van der Vleuten & Schuwirth 2005). Hodges (2013) also highlights the threat to validity posed by the "opposite" practice of aggregating sub-scores from instruments with different purposes to "reconstitute competence." The introduction of

entrustable professional activities (EPAs) (ten Cate & Scheele 2007; ten Cate 2013), in which progressive levels of supervision and delegation lead to independent practice, and the Milestone Project (Nasca et al. 2012), which uses narrative descriptions of behaviors for the levels of performance related to competencies, respond to these concerns.

Hodges (2013) sums up much of the new direction in psychometric discourse in the title of his article "Assessment in the post-psychometric era: Learning to love the subjective and collective." He suggests that a "collective perspective" can frame both data-gathering and the aggregation and interpretation of independent judgments. The use of an assessment framework that integrates EPAs and milestones supports this "both/and" approach. Holistic assessment based on EPAs, which are professional activities that require integrated competencies, in combination with milestones, which provide a more granular description of individual competencies and the substrate of formative feedback, will advance competency-based assessment.

Importance of group process

Although assessment processes and research have typically focused on the assessment of one individual by another, or of an individual method such as a knowledge exam, interest is growing in the use of group processes to improve judgments of overall competence. Invoking the "wisdom of crowds," Surowiecki (2005) describes how good group process can be employed in decision-making. Despite the need for multiple assessors, no combination of assessors and assessment methods can measure "all things." A synthesis process is still essential, and group judgment may provide the best means of maximizing the reliability of entrustment decisions. There is some evidence in medicine to support the "wisdom of crowds" principle. Hemmer et al. (2000) found that many deficiencies in professionalism were detected and discussed only in a group evaluation session. Schwind et al. (2004) found that 18% of competency deficiencies in a surgery residency were detected only during discussion at a clinical competency committee. A group process conducted by clinical competency committees is now a required component of the Next Accreditation System in the United States for graduate medical education (Nasca et al. 2012).

Information management and documentation

Shifting to CBME exposes the unmet challenges of learner assessment that had defined the status quo. Attempts to address these challenges have provoked many faculty and program directors to voice concern about the potentially prohibitive assessment burden of CBME. This concern is valid, as faculties will need to assess broad and diverse competencies, to engage in faculty development to produce better assessors, and to widely sample learner progress so that the reliability of expert judgments is enhanced. For CBME to be successful, strategies to mitigate the impact of the change must be employed; the use of technology can be helpful in this regard.

The Internet, handheld devices, innovative software, and other technologies have the potential to facilitate CBME (Table 2). Connected platforms for the dissemination and collection of assessments, along with a relational database that facilitates the aggregation and analysis of data will be critical. Currently, data synthesis and aggregation are done manually at great expense of time and effort. With these developments, technology can (1) prompt reflection on performance at the individual and program level; (2) inform progression decisions and other important judgments about learners; (3) provide individual learner feedback regarding skill progression along a developmental continuum; (4) create dashboards (graphic representations of an individual's or program's performance against a reference group) to stimulate trainees' reflection on learning needs and performance gaps; and (5) create dashboards for program directors to compare individual and aggregate learner progress with local and national peer groups in other programs as well as to provide comparisons against national or international standards (Schumacher et al. 2014).

Such technologies can facilitate assessment from undergraduate training through to continuing professional development. Technologies could play a role in formative and summative assessments, in both low-stakes and high-stakes contexts, including for certification and revalidation or maintenance of certification. However, maintaining the security of patient and learner data will be critically important.

There are three documentation issues to highlight: assessment program auditing, transparency, and accountability. The audit trail, which is part of the validity evidence for qualitative assessment, is likewise an important component of the validity evidence for any high-stakes decision. Auditing establishes trustworthiness by addressing the dependability or conformability of a judgment (van der Vleuten et al. 2010). Auditing involves documenting the process, supporting the outcome (e.g. if learners question a decision), and providing evidence of quality for external reviewers. The clinical competency committee required by the ACGME for assessing learner milestones sets the stage for auditing documentation going forward (Nasca et al. 2012). Transparency goes hand in hand with creating an audit trail, and documenting the rules, evidence, thought processes, and reasons for decision-making are essential to both. Taking assessment beyond a judgment based on a set of scores to a judgment that includes an interpretation of those scores should be more trustworthy in the end (Govaerts & van der Vleuten 2013). A robust and transparent program of assessment should ideally make the "summative" decisions about learner performance for the benefit of the public more, rather than less, straightforward. Ultimately, there has to be accountability. Those who synthesize and reach decisions are responsible to the trainee and to the larger health care system that trusts their judgments about a physician's ability to progress further.

Improving the assessment system

Dijkstra et al. (2010) and van der Vleuten et al. (2012) make a strong case for creating an assessment program that allows for both ongoing learner assessment and program evaluation. Programs that can forward information from one phase of learning to the next will enable learners to

Table 2. Information management options and uses for assessment[a].

Tools and materials					
Type	Examples	Uses	CanMEDS Role assessed[b]	Benefits	Challenges
Computers and tablets Mobile technology	Audience response systems (clickers) and smartphones Computers on wheels, desktop or laptop computers and tablets	Access to websites, programs, email communications, and documents Interactive feedback Part of toolkit in summative assessment methods, e.g. within OSCEs (S) Point of care and "just in time" information searches, e.g. assessment of scholarly and time-management skills (F) Pre- and post-session quizzes (F/S)	All	Can provide teachers and learners alike with a gauge to learning through pre- and post-session questions Ease of search, mobility Immediacy of feedback User comfort high	Upfront costs of mobile devices Uploading of software onto computer to use mobile technology
Web-based	CanMEDS Interactive http://canmeds.royalcollege.ca/ Curriculum management systems Electronic surveys	Accessible via mobile or desktop platforms Direct observation forms, multi-source feedback forms (F) Final in-training assessment forms (S) Houses educational standards, materials, and tools Reflection, self-assessment and documentation re activities within the Intrinsic Roles Role-specific assessment tools (F)	All	Access at point of care or teaching Can be useful for bedside teaching, direct observation, etc. (i.e. work-based assessments) Reduced effort in sending forms, data collation, and report generation; improved convenience for users of forms; improved completion and return	High cost (although some are free of charge)
Digital	Electronic medical record Health electronic record	To support chart audits and related workplace-based assessments (F/S)	Communicator (written) Medical Expert	Potential to track completeness of charting, quality of computerized order entry, documentation, etc., as well as patient outcomes such as length of stay and complication rates. When decision support systems are built in, can prompt use of clinical practice guidelines or standardized protocols or redirect a provider when ordering an inappropriate test. The flagging by such systems can be monitored and used to enhance assessment	Different software do not share data for inter-center collaboration or communication High cost (upfront startup and maintenance costs and human resources)
Audio and video	Electronic portfolio/log-book with audio, video, and/or text entries	Reflection, self-assessment, and documentation of cases, procedures, narratives (F) Sampling of reflection, self-assessment, and documentation of cases, procedures, narratives may be used as part of toolkit (S)	Communicator Professional	On-hand documentation, criteria for assessment on hand, prior learning needs available for review, ongoing updating, reflection	Human resources needed for faculty required to support assessment time, faculty development
Social networks	Group webpages, wikis, blogs, Twitter, etc.	Assessment of professionalism in terms of communication by voice, chat, instant messaging, video conferencing, blogs, and tweets in an interactive learning environment.	Communicator Professional	Ease of sharing information Increased learner-learner interaction and learner motivation as well as learner–teacher interactions Multiple modalities of data available to meet different leaner needs	Assessors and organizations need to consider issues of privacy and confidentiality regarding use of data for assessment purposes Determining quality of the data Managing a large volume of data
Virtual classroom	Communication between learners as well as with teachers via webcam, microphone, and real-time chatting Telehealth/web conferencing systems, e.g. Go-To-Meeting or Adobe Connect, to simulate classroom or meetings	To increase opportunities for learners to demonstrate role as Scholar/Teacher (e.g. facilitate teaching sessions) or Manager/Collaborator Role (e.g. attend meetings) (F/S) Polls, quizzes (F/S)	Collaborator Leader Scholar	Can connect teachers and learners across different sites. Increased connectivity for learners with program, especially if programing is shared across multiple sites Synchronous or asynchronous use	High cost; dependence on Internet connection quality
Learner management systems	Blackboard and other web-based learning management systems	Submission and tracking tools for online assessment (synchronous or asynchronous) (F/S) To track attendance, time on task, learner progress patterns	Leader Professional	Secure content management and sharing as well as supporting virtual collaboration Online assessments, learner tracking, and assignment management	Dependence on Internet connection quality High cost (startup, maintenance, and human resources)

(*continued*)

Table 2. Continued

Tools and materials					
Type	Examples	Uses	CanMEDS Role assessed[b]	Benefits	Challenges
Intelligence tools	Procedural models, games, virtual reality, dexterity analysis devices (through motion tracking), computer modeling, etc.	Part of toolkit within simulation-based summative assessment methods (S) To provide controlled contexts that support provision of feedback (F)	Medical Expert	Electronic reminders for both faculty and learners Synchronous or asynchronous use Can support learning about rare cases and/or learning for high- risk environments. More elaborate assessments including team performance can be assessed in venues such as simulated operating suites or in situ simulation whereby a clinical space is used to stage the simulation	High cost (startup, maintenance, and human resources)

F: formative; S: summative; OSCE: Objective Structured Clinical Examination.
[a]Adapted from: Felkey et al. (2005); Courts & Tucker (2012); Hicks et al. (2014); "Educational technology," Wikipedia (http://en.wikipedia.org/wiki/Educational_technology).
[b]This table uses the CanMEDS Framework as an example, but the classifications shown can be applied to other competency-based frameworks.

focus on gaps and build on strengths over time (Eva et al. 2013). Similarly, at the program level, data collected from the assessment of several trainees can be aggregated to assess curriculum effectiveness or to determine whether a cohort was able to reach desired levels of competence within a reasonable period. Data collected for a cohort should be applied in a continuous process of quality improvement and innovation in medical education. Programs should continually ask what works, for whom, in what circumstances, and why (Pawson & Tilley 1997). As Pawson has noted, program interventions are almost always "partial solutions" that must be continually refined and revised (Pawson 2013).

Conclusions

Assessment in a CBME environment requires attention to ensure that it provides feedback *for* and *of* learning. More assessments will be needed by trained assessors on an ongoing basis. This paper highlights the importance of multiple assessments with multiple assessors, assessor training, a reconceptualization of the role of psychometrics, and the need for solid group processes for decision-making. The effective use of technology can help to manage information and track progress toward competence, while facilitating audits and transparency. CBME requires a dynamic environment that is attentive to the demands of the health care system and continually strives to optimize assessments for and of learning.

Disclosure statement

Eric Holmboe is employed by the ACGME and receives royalties for a textbook on assessment from Mosby-Elsevier. Resources and secretariat support for this project was provided by the Royal College of Physicians and Surgeons of Canada.

Notes on contributors

Jocelyn Lockyer, PhD, is the Senior Associate Dean of Education, and Professor, Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Canada.

Carol Carraccio, MD, is the Vice President, Competency-Based Assessment, American Board of Pediatrics, USA.

Ming-ka Chan, MD, is an Associate Professor at Department of Pediatrics and Child Health, University of Manitoba, Canada.

Danielle Hart, MD, is the Program Director, Emergency Medicine Residency, Director of Simulation, Interdisciplinary Simulation and Education Center, Hennepin County Medical Center, and Assistant Professor, Department of Emergency Medicine, University of Minnesota Medical School, USA.

Sydney Smee, PhD, is the Assessment Advisor for the Evaluation Bureau of the Medical Council of Canada, Ottawa, Canada.

Claire Touchie, MD, is the Chief Medical Education Advisor, Medical Council of Canada, and Associate Professor of Medicine, University of Ottawa, Canada.

Eric S. Holmboe, MD, is the Senior Vice President for Milestone Development and Evaluation, Accreditation Council for Graduate Medical Education, USA.

Jason R. Frank, MD, is the Director, Specialty Education, Strategy and Standards in the Office of Specialty Education at the Royal College of Physicians and Surgeons of Canada, and the Director of Educational Research & Development in the Department of Emergency Medicine, University of Ottawa, Canada.

References

Archer JC. 2010. State of the science in health professional education: effective feedback. Med Educ. 44:101–108.

Berendonk C, Stalmeijer RE, Schuwirth LWT. 2013. Expertise in performance assessment: assessors' perspectives. Adv Health Sci Educ Theory Pract. 18:559–571.

Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. 2002. Shifting paradigms: from Flexner to competencies. Acad Med. 77:361–367.

Courts B, Tucker J. 2012. Using technology to create a dynamic classroom experience. TLC. 9:121–128.

Dewey J. 1974. John Dewey on education: selected writings. Chicago: University of Chicago Press.

Dijkstra J, van der Vleuten CPM, Schuwirth LWT. 2010. A new framework for designing programmes of assessment. Adv Health Sci Educ Theory Pract. 15:379–393.

Eva KW. 2005. What every teacher needs to know about clinical reasoning. Med Educ. 39:98–106.

Eva K, Bordage G, Campbell C, Galbraith R, Ginsberg S, Holmboe E, Regehr G. 2013. Report to the Medical Council of Canada on current issues in health professional and health professional trainee assessment [Internet]. Ottawa: Medical Council of Canada; [cited 2014 Sep 4]. Available from: http://mcc.ca/wp-content/uploads/Reports-MEAAC.pdf.

Eva KW, Hodges BD. 2012. Scylla or charbydis? Can we navigate between objectification and judgement in assessment? Med Educ. 46:414–419.

Eva KW, Regehr G. 2005. Self-assessment in the health professions: a reformulation and research agenda. Acad Med. 80:S46–S54.

Eva KW, Regehr G. 2007. Knowing when to look it up: a new conception of self-assessment ability. Acad Med. 82(10 Suppl):S81–S84.

Eva KW, Regehr G. 2008. "I'll never play professional football"; and other fallacies of self-assessment. J Contin Educ Health Prof. 28:14–19.

Felkey BG, Fox BI, Thrower MR. 2005. Health care informatics: a skills-based resource. Washington: American Pharmacists Association.

Frank JR, Danoff D. 2007. The CanMEDs initiative: implementing an outcomes-based framework of physician competencies. Med Teach. 29:642–647.

Friedman Ben David M, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ. 2001. AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. Med Teach. 23:535–551.

General Medical Council. 2013. Good Medical Practice: Working with doctors, working with patients [Internet]. General Medical Council: Manchester; [cited 2014 Oct 28]. Available from: www.gmc-uk.org/guidance.

Gingerich A, Regehr G, Eva KW. 2011. Rater-based assessments as social judgments: rethinking the etiology of rater errors. Acad Med. 86:S1–S7.

Govaerts MJB, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. 2011. Workplace-based assessment: effects of rater expertise. Adv Health Sci Educ. 16:151–165.

Govaerts M, van der Vleuten CPM. 2013. Validity in work-based assessment: expanding our horizons. Med Educ. 47:1164–1174.

Govaerts MJB, van de Wiel MWJ, Schuwirth LWT, van der Vleuten CPM, Muijtjens AMM. 2013. Workplace-based assessment: raters' performance theories and constructs. Adv Health Sci Educ. 18:375–396.

Harris P, Bhanji F, Topps M, Ross S, Lieberman S, Frank JR, Snell L, Sherbino J; ICBME Collaborators. 2017. Evolving concepts of assessment in a competency-based world. Med Teach. 39:603–608.

Hawkins RE, Durning SJ. 2017. Program evaluation. In: Holmboe ES, Durning SJ, Hawkins RE, editors. Practical guide to the evaluation of clinical competence. 2nd ed. Philadelphia: Elsevier; p. 303–331.

Hemmer PA, Hawkins R, Jackson JL, Pangaro LN. 2000. Assessing how well three evaluation methods detect deficiencies in medical students' professionalism in two settings of an internal medicine clerkship. Acad Med. 75:167–173.

Hicks PJ, Schwartz A, Clyman SG, Nichols DG. 2014. The Pediatrics Milestones: pursuit of a national system of workplace-based assessment through key stakeholder collaboration. Acad Pediatr. 14:S10–S12.

Hodges B. 2013. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach. 35:564–568.

Holmboe ES, Davis MH, Carraccio C. 2008. Portfolios. In: Holmboe ES, RE Hawkins RE, editors. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby.

Holmboe ES, Hawkins RE, Huot SJ. 2004. Effects of training in direct observation of medical residents' clinical competence: a randomized trial. Ann Intern Med. 140:874–881.

Kane MT. 2013. Validating the interpretations and uses of test scores. J Educ Meas. 50:1–73.

Knowles M. 1975. Self-directed learning: a guide for learners and teachers. New York: Pearson Learning/Cambridge Adult Education.

Kogan JR, Conforti LN, Iobst WF, Holmboe ES. 2014. Reconceptualizing variable rater assessments as both an educational and clinical care problem. Acad Med. 89:721–727.

Kogan JR, Holmboe E. 2013. Realizing the promise and importance of performance-based assessment. Teach Learn Med. 25:S68–S74.

Lievens F. 2001. Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. J Appl Psychol. 86:255–264.

Miller GM. 1990. The assessment of clinical skills/competence/performance. Acad Med. 65:S63–S67.

Nasca TJ, Philibert I, Brigham T, Flynn TC. 2012. The next GME accreditation system-rationale and benefits. N Engl J Med. 366:1051–1056.

Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, et al. 2011. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 33:206–214.

Nyquist JG. 2014. Techniques for assessment of learner performance in teaching and assessing the competencies, 10th ed. Los Angeles: University of Southern California.

Pawson R. 2013. The science of evaluation: a realist manifesto. London: Sage Publications.

Pawson R, Tilley N. 1997. Realistic evaluation. San Francisco: Sage Publications.

Pelgrim EAM, Kramer AWM, Mokkink HGA, van der Vleuten CPM. 2013. Reflection as a component of formative assessment appears to be instrumental in promoting the use of feedback: an observational study. Med Teach. 35:772–778.

Ponnamperuma G. 2013. Workplace based assessment. In: Walsh E, editor. Oxford textbook of medical education. Oxford: Oxford University Press.

Regehr G, MacRae H, Reznick RKS, D. 1998. Comparing psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Acad Med. 73:993–997.

Sargeant J, Armson H, Chesluk B, Dornan T, Eva K, Holmboe E, Lockyer J, Loney E, Mann K, van der Vleuten CP. 2010. The processes and dimensions of informed self-assessment: a conceptual model. Acad Med. 85:1212–1220.

Sargeant J, Eva KW, Armson H, Chesluk B, Dornan T, Holmboe E, Lockyer JM, Loney E, Mann KV, van der Vleuten CP. 2011. Features of assessment learners use to make informed self-assessments of clinical performance. Med Educ. 45:636–647.

Schumacher DJ, Spector ND, Calaman S, West DC, Cruz M, Frohna JG, Gonzalez Del Ray Gustafson KK, Poynter SE, Rosenbluth G, et al. 2014. Putting the pediatrics milestones into practice: a consensus roadmap and resource analysis. Pediatrics. 133:898–906.

Schuwirth LWT, Ash J. 2013. Assessing tomorrow's learners: in competency-based education only a radically different holistic method of assessment will work. Six things we could forget. Med Teach. 35:555–559.

Schuwirth LWT, van der Vleuten CPM. 2006. A plea for new psychometric models in educational assessment. Med Educ. 40:296–300.

Schwind CJ, Williams RG, Boehler ML, Dunnington GL. 2004. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? Acad Med. 79:453–457.

Smithers JW. 1998. Performance appraisal: state of the art in practice. San Francisco: Jossey-Bass.

Surowiecki J. 2005. The wisdom of crowds. New York: Anchor Books.

Swing SR. 2007. The ACGME outcome project: retrospective and prospective. Med Teach. 29:648–654.

ten Cate O, Scheele F. 2007. Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? Acad Med. 82:542–547.

ten Cate O. 2013. Nuts and bolts of entrustable professional activities. J Grad Med Educ. 5:157–158.

van der Vleuten CPM. 1996. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ. 1:41–67.

van der Vleuten CPM, Schuwirth LWT. 2005. Assessing professional competence: from methods to programmes. Med Educ. 39:309–317.

van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, van Tartwijk J. 2012. A model for programmatic assessment fit for purpose. Med Teach. 34:205–214.

van der Vleuten CPM, Schuwirth LWT, Scheele F, Driessen EW, Hodges B. 2010. The assessment of professional competence: building blocks for theory development. Best Pract Res Clin Obstet Gynaecol. 24:703–719.

van Tartwijk J, Driessen EW. 2009. Portfolios for assessment and learning: AMEE Guide no. 45. Med Teach. 31:790–801.

Watling C, Driessen E, van der Vleuten CPM, Vanstone M, Lingard L. 2012. Understanding responses to feedback: the potential and limitations of regulatory focus theory. Med Educ. 46:593–603.

Woehr DJ, Huffcutt AI. 1994. Rater training for performance appraisal: a quantitative review. J Occup Org Psych. 67:189–205.

Yeates P, O'Neill P, Mann K, Eva KW. 2012. Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. JAMA. 308:2226–2232.