Assessment Pearls for Competency-Based Medical Education

Susan Humphrey-Murto, MD, MEd, FRCPC
Timothy J. Wood, PhD
Shelly Ross, MA, PhD
Walter Tavares, PhD
Brent Kvern, MD, CCFP, FCFP
Ravi Sidhu, MD, MEd, PhD
Joan Sargeant, BN, MEd, PhD
Claire Touchie, MD, MHPE, FRCPC

Competency-based medical education (CBME) is changing the way we teach and assess residents, yet many educators feel ill-prepared to function in this new environment. The daunting task of developing entrustable professional activities (EPAs) and milestones is arguably surpassed only by our trepidation about the need for new assessment strategies.[1] As a group of MD and PhD medical educators who develop local- and national-level examinations and teach a national-level assessment course, we have identified the need for basic principles to guide clinicians who have been tasked with developing assessment strategies in CBME. The 10 practical "assessment pearls" we offer in this perspective reflect an evidence-based foundation that should serve as a user-friendly guide to these clinician educators. A glossary of terms is provided in the BOX for readers new to the field.

1. All Assessments Are Samples

It is not possible to assess everything residents are expected to demonstrate, so we must deliberately sample representative knowledge and skills using a carefully constructed blueprint.[2] A blueprint defines *what* is being assessed; it also provides evidence that the assessment strategy is valid by ensuring sufficient and appropriate sampling. For example, 1 EPA for internal medicine is "admit and manage a medical inpatient with a new acute problem on a medical floor."[3,4] Several competencies are required for a resident to complete this EPA: knowledge of basic science, clinical features and management strategies, as well as communication skills and the ability to perform an appropriately focused physical examination. For each competency, a sampling strategy is required. It might seem evident that one could not assess a resident on every type of patient who requires a focused physical examination, but one could sample deliberately—for example, based on the primary system involved, such as a patient presenting with a respiratory, cardiovascular, or musculoskeletal problem.

DOI: http://dx.doi.org/10.4300/JGME-D-17-00365.1

2. The Higher the Stakes, the More Samples Are Needed

All measurements have error, and the higher the stakes, the more assessment points (or samples) are required.[5] For example, if the purpose of the assessment is to provide feedback to residents on the management of patients in ambulatory clinic, a single assessment would be appropriate. On the other hand, if the Clinical Competency Committee (CCC) wanted to make decisions on learner promotion (ie, the assessment is for higher-stakes pass/fail decisions), then using only 1 faculty assessment would not be defensible. The small sample of observed behavior is insufficient to make a meaningful interpretation of the resident's performance score for such an important decision. Defining the purpose (high/low stakes) of the assessment can help determine how many samples are needed.

3. Assess What Is Important, Not Just What Is Easy

Educators often default to assessing what is easy, rather than assessing what is important. We know that using a written examination to assess all aspects of clinical competence is not adequate. The intrinsic (nonmedical expert) roles in frameworks such as CanMEDS 2015[6] and the Accreditation Council for Graduate Medical Education (ACGME) competencies[7] can be difficult to assess (eg, professionalism), yet they are important elements of physician competence. Newer tools have been developed to help meet these challenges, such as multi-source feedback for team skills and communication,[8] narrative feedback for patient-centered care, communication, and professionalism,[9] and the Ottawa Surgical Competency Operating Room Evaluation for operative competency.[10] Clearly, medical educators have made progress on developing tools to assess different areas of competence, not just the easier competencies to assess such as medical knowledge (ACGME) or the CanMEDS medical expert role.

4. All Assessment Involves Judgment

Perfect objectivity and standardization is neither possible nor desirable. Considering that in order for a test score to be generated, a performance must be observed and then converted into a score, which then must be interpreted—*every* step involves judgment. At an individual level, judgment occurs when the preceptor (or rater) assesses a resident's performance in a particular context, and determines what feedback to provide. Rater cognition has received increased attention as a strength, where multiple raters provide different perspectives and have the potential to provide richer data about a resident's performance.[11] On the other hand, excessive variability in ratings by different raters has raised concerns about reliability and validity.[12] Training raters may improve assessment quality, but results are mixed.[13] Judgment also occurs collectively when CCC members review a resident's entire portfolio. Judgment in that context will determine what elements of the portfolio should be provided to the committee for review, and how to weigh individual elements. For example, the committee may judge a professionalism item to be so egregious that it would override excellent knowledge assessments and delay promotion to the next stage. Determining the threshold or standard for promotion clearly requires collective judgments and the consensus of experts.[11,14]

5. Quantitative and Qualitative Methods Complement One Another

Quantitative data have traditionally been considered more desirable for their supposed objectivity, yet a limitation of numbers is that they do not provide specific data about how to improve. Narratives have been shown to capture elements of performance that an accumulation of numbers may mask[15]; further, in unstandardized situations—such as most workplace assessment situations—narrative provides much better data for feedback and learning.[11]

6. No Single Assessment Tool Can Capture All Aspects of Clinical Competence

Clinical competence is a complex construct necessitating a diverse set of assessment tools and strategies. To illustrate, competence involves knowledge, which may be best assessed with written examinations; clinical skills require direct observation, such as an objective structured clinical examination; managing patients on an inpatient unit involves use of workplace-based assessment tools, such as 360-degree feedback; and finally, assessing diagnostic reasoning skills may require multiple tools for assessment in the clinical setting using retrospective case analysis, and for rare events using computer-based case simulations.[19]

BOX Glossary of Terms

- **Clinical Competency Committee (CCC):** In the context of competency-based medical education, the CCC is a committee that includes members of the faculty who will use a combination of assessment data gathered from multiple sources to evaluate learners' progress and make high-stakes decisions.
- **Competence:** An array of abilities that enables the trainee or physician to do all tasks of practice effectively and consistently. It is considered a complex construct.
- **Competency:** An observable ability of a trainee or physician. Example: perform a complete and accurate physical examination.
- **Competency-Based Medical Education:** An approach to educating physicians that is oriented to outcome abilities and organized around competencies. It de-emphasizes time-based training and promises more flexibility and learner-centeredness.[16]
- **Construct:** An intangible collection of abstract concepts that are inferred from behavior.[5] For example "clinical competence" and "professionalism" are constructs that may be of interest to assess, but are inferred from the trainee's behavior in the workplace.
- **Entrustable Professional Activity (EPA):** Units of professional practice, defined as real-life tasks essential for a particular specialty and can be entrusted to a trainee once competence has been attained.[3] Example of an EPA: "Manage care of patients with acute common diseases across multiple care settings."[17]
- **Low-Stakes and High-Stakes Assessments:** Low-stakes assessments have limited consequences for the trainee in terms of promotion, selection, or certification, whereas high-stakes assessments can have far-reaching consequences such as failure to become certified.[11]
- **Milestones:** The expected ability of a trainee at a stage of expertise, as he or she moves from novice to expert. Example of neurology physical examination: level 1—performs complete neurological examination—to level 4—efficiently performs a relevant neurological examination accurately incorporating all additional appropriate maneuvers.[18]
- **Programmatic Approach to Assessment:** The use of several assessment methods arranged longitudinally and constructed deliberately to optimize learning and assessment. The program would include several low-stakes assessment data points that are aggregated for higher-stakes pass/fail decisions.[11]
- **Rater Cognition:** The mental processes that occur during scoring, at either a conscious or unconscious level.

7. Feedback Is an Essential Element of Assessment

Assessment should inform learners how they are progressing toward becoming experts, and formative feedback is an essential part of that. Van der Vleuten

and colleagues,[11,20] in describing programmatic assessment, indicated that best assessment practice is not only about doing well enough to pass a unit of instruction, but also providing an opportunity for formative feedback that contributes to improved performance.

8. Assessment Drives Learning

Learners will "study to the test" whenever possible, focusing their study strategies on concepts they know will be examined. Test-enhanced learning involves learning as trainees prepare for the test, complete it, and then receive feedback. Individuals tested on material have improved recall, compared to those who simply studied the material.[21] Similarly, assessment drives learning in the workplace: when learners see that all aspects of being a physician are being assessed, the importance of mastering those nonmedical expert competencies becomes apparent. Assessment strategies should therefore be designed with this in mind.

9. Validity Is the Most Important Characteristic of Assessment Data

Simply put, validity is the overall judgment of the degree to which theory and evidence support the interpretation of assessment scores for a specific purpose.[5,22] If a resident scores perfectly on a multiple-choice examination of knowledge, can we conclude that the resident is ready to take first call for all consultations coming to internal medicine? Although 1 interpretation of this learner's multiple-choice question score is *Yes, call ready*, many argue that demonstrating knowledge is not enough. We also need to know the resident's abilities in history taking, physical examination, management, and procedural skills. Thus, the interpretation of the multiple-choice test score as proof of readiness to take first call is not valid: the evidence (knowledge testing) does not support the purpose (overall competence). This highlights an important concept: that there is no such thing as a valid or invalid test. Validity always refers to the appropriateness of inferences or judgments based on test scores for a specific purpose. It is beyond the scope of this article to elaborate further, but validity is a unitary concept and requires multiple sources of evidence to support or refute meaningful score interpretation.[5]

10. Perfect Assessment Is an Illusion

There are many criteria that are relevant to any assessment: validity, reproducibility, equivalence, feasibility, educational effect, catalytic effect, and acceptability.[23,24] Ultimately, assessment always involves some degree of compromise. Consider a *low-stakes assessment* where the purpose is to provide residents with progress data and feedback. In this instance, feasibility (ease of administration), acceptability (for residents and faculty), education effect (facilitates feedback), and catalytic effect (provides results that enhance education) would all be considered important. If this was a *high-stakes assessment* with significant consequences, then reproducibility (statistically reliable test) and equivalence (every resident is tested in the same way) would be paramount to lead to defensible results.

We have outlined 10 assessment pearls that can help guide clinician educators and program directors tasked with transforming their programs of assessment to meet the new requirements of CBME. Important concepts in CBME will have direct implications for assessment: EPAs will demand more rigorous workplace assessments; milestones will necessitate ongoing direct observation and feedback as well as a continuous program of assessment; and CCCs will be challenged with determining how to integrate data for decision-making purposes. We hope these basic principles can serve as guideposts to ease transition during this exciting time in medical education.

References

1. Holmboe ES, Sherbino J, Long DM, et al. The role of assessment in competency-based medical education. *Med Teach*. 2010;32(8):676–682.
2. Hamdy H. Blueprinting for the assessment of healthcare professionals. *Clin Teach*. 2006;3(3):175–179.
3. Hauer KE, Kohlwes J, Cornett P, et al. Identifying entrustable professional activities in internal medicine training. *J Grad Med Educ*. 2013;5(1):54–59.
4. ten Cate O, Chen HC, Hoff RG, et al. Curriculum development for the workplace using Entrustable Professional Activities (EPAs): AMEE Guide No. 99. *Med Teach*. 2015;37(11):983–1002.
5. Downing SM. Validity: on the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830–837.
6. Frank JR, Snell L, eds. The Draft CanMEDS 2015 Physician Competency Framework. Royal College of Physicians and Surgeons of Canada. http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/framework_series_1_e.pdf. Accessed October 4, 2017.
7. Educational Commission for Foreign Medical Graduates. ACGME core competencies. http://www.ecfmg.org/echo/acgme-core-competencies.html. Accessed October 4, 2017.

8. Donnon T, al Ansari A, Al Alawi S, et al. The reliability, validity, and feasibility of multisource feedback physician assessment: a systematic review. *Acad Med*. 2014;89(3):511–516.
9. Ross S, Poth CN, Donoff M, et al. The competency-based achievement system: using formative feedback to teach and assess competencies with family medicine residents. *Can Fam Physician*. 2011;57(9):e323–e330.
10. Gofton WT, Dudek NL, Wood TJ, et al. The Ottawa Surgical Competency Operating Room Evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. 2012;87(10):1401–1407.
11. van der Vleuten CP, Schuwirth LW, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–214.
12. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Acad Med*. 2011;86(suppl 10):1–7.
13. Lee V, Brain K, Martin J. Factors influencing mini-CEX rater judgments and their practical implications: a systematic literature review. *Acad Med*. 2017;92(6):880–887.
14. Hauer KE, ten Cate O, Boscardin CK, et al. Ensuring resident competence: a narrative review of the literature on group decision making to inform the work of clinical competency committees. *J Grad Med Educ*. 2016;8(2):156–164.
15. Cook DA, Kuper A, Hatala R, et al. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med*. 2017;91(10):1359–1369.
16. Frank JR, Mungroo R, Ahmad Y, et al. Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Med Teach*. 2010;32(8):631–637.
17. Caverzagie KJ, Cooney TG, Hemmer PA, et al. The development of entrustable professional activities for internal medicine residency training: a report from the Education Redesign Committee of the Alliance for Academic Internal Medicine. *Acad Med*. 2015;90(4):479–484.
18. Leep Hunderfund AN, Reed DA, Starr SR, et al. Ways to write a milestone: approaches to operationalizing the development of competence in graduate medical education. *Acad Med*. 2017;92(9):1328–1334.
19. Ilgen JS, Humbert A, Kuhn G, et al. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. *Acad Emerg Med*. 2012;19(12):1454–1461.
20. van der Vleuten CP, Schuwirth LW, Driessen EW, et al. 12 tips for programmatic assessment [published online ahead of print November 20, 2014]. *Med Teach*. 2014:1–6. doi:10.3109/0142159X.2014.973388.
21. Larsen DP, Butler AC, Roediger HL III. Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Med Educ*. 2013;47(7):674–682.
22. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006;119(2):166.e7–166.e16.
23. Norcini J, Anderson B, Bollela V, et al. Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206–214.
24. van der Vleuten CP. The assessment of professional competence: developments, research, and practical implications. *Adv Health Sci Educ*. 1996;1(1):41–67.

Susan Humphrey-Murto, MD, MEd, FRCPC, is Associate Professor, Department of Medicine, Faculty of Medicine, University of Ottawa, Ontario, Canada; **Timothy J. Wood, PhD,** is Professor, Department of Innovation in Medical Education, Faculty of Medicine, University of Ottawa; **Shelly Ross, MA, PhD,** is Associate Professor and Director, CBAS Program Research and Innovation, Department of Family Medicine, University of Alberta, Edmonton, Alberta, Canada; **Walter Tavares, PhD,** is Assistant Professor and Scientist, The Wilson Centre, Department of Medicine, University of Toronto, Ontario, Canada, and affiliated with Post Graduate Medical Education/Continued Professional Development, University of Toronto, and Paramedic and Senior Services, Community and Health Services Department, Regional Municipality of York, Newmarket, Ontario, Canada; **Brent Kvern, MD, CCFP, FCFP,** is Associate Professor, Department of Family Medicine, Max Rady College of Medicine, University of Manitoba, Winnipeg, Manitoba, Canada, Site Director, Family Medicine, St. Boniface General Hospital, Winnipeg, Manitoba, Canada, and Chair, Central Examination Committee, Medical Council of Canada, Ottawa, Ontario, Canada; **Ravi Sidhu, MD, MEd, PhD,** is Associate Professor of Surgery and Associate Dean, Postgraduate Medical Education, University of British Columbia, Vancouver, British Columbia, Canada; **Joan Sargeant, BN, MEd, PhD,** is Professor, Division of Medical Education, Faculty of Medicine, Dalhousie University, Halifax, Nova Scotia, Canada; and **Claire Touchie, MD, MHPE, FRCPC,** is Associate Professor, Department of Medicine, Faculty of Medicine, University of Ottawa, and Chief Medical Education Advisor, Medical Council of Canada.

The authors would like to thank the Department of Innovation in Medical Education at the University of Ottawa for ongoing research assistant support.

Corresponding author: Susan Humphrey-Murto, MD, MEd, FRCPC, University of Ottawa Faculty of Medicine, Riverside Hospital, Box 37, 1967 Riverside Hospital, Ottawa, ON K1H 7W9 Canada, 613.737.8899 ext 81850, shumphrey@toh.ca