Post MD Education
UNIVERSITY OF TORONTO

Best Practices in Evaluation and Assessment (BPEA)

Programs of Assessment

Donna Steele MA, MD, FRCSC
Residency Program Director
Assistant Professor, Department of Obstetrics and Gynecology, Faculty of Medicine, University of Tor

Christopher Li MD, FRCPC, DABSM
Residency Program Director, Faculty of Medicine, University of Toronto

Glenys A. Babcock PhD
Manager, Data & Analytics
Post-MD Education, Faculty of Medicine, University of Toronto

UT CBME / CBD

Best Practices in Evaluation and Assessment (BPEA)
Programs of Assessment

Donna Steele, Christopher Li, Glenys A. Babcock

1. Executive Summary

Programs of Assessment for competency-based medical education (CBME) will be most successful if they explicitly:

- Delineate individual specific competencies required, rather than listing generic or summative expectations;
- Link each assessment to a particular CanMEDS role;
- Evaluate each competency for each resident with multiple assessment tools, used by multiple assessors, over time; and
- Know and keep in mind the key real-life and workplace factors that can undermine validity and/or reliability of the assessment tools.

The academic and practitioner literature on Programs of Assessment and on CBME Assessment Tools is rich, complex, and growing. A key learning point is that each learning context is unique and the development of CBME assessment programs must be approached with that in mind.

Given the complexities in the existing literature and the fact that the field is in its adolescence, Program Directors and other key educators need to have simple distilled tools at hand that they can rely on. We have developed a preliminary Matrix of Assessment Tools that may be the foundation on which a more substantial Matrix of Assessment Tools is built.

2. Background

This paper explores three facets of Programs of Assessment:

1. Best tools for different purposes (e.g., direct observation, CanMEDS roles, types of competencies, levels of competence, assessments that support learning, assessments that support decisions on promotions and/confirmation of progress);
2. Principles and processes to select the appropriate number and variety of tools; and
3. Contextual considerations for implementing assessment tools in the context of competency-based medical education.

3. Methodology

The research methodology was carried out in two phases. In Phase 1, a literature review of approximately 50 academic articles was undertaken to explore principles, processes and contextual considerations in the utilization of Assessment Tools was undertaken. The great majority of these articles were selected and provided to the research team by Dr. Susan Glover Takahashi.

Most of the Phase 1 academic articles were secondary reviews of key considerations in utilizing assessment tools for competency-based medical education; some articles reported findings from primary research on assessment in competency-based medical education.

The team members individually produced annotated bibliographies of key articles, and then, through a collaborative, iterative process, identified salient and dominant themes.

In Phase 2 of the methodology, focused on evaluations of specific relevant assessment tools. To this end, an environmental scan and an extended literature review were conducted to focus on evaluations of specific relevant assessment tools. The environmental scan primarily covered practitioner-oriented organizations (e.g., Royal College of Physicians and Surgeons) and university medical programs (e.g., University of Ottawa). The extended literature review drew in additional academic articles and practitioner reports related to specific assessment tools, specific categories of assessment tools, and the relationship between specific assessment tools and particular CanMEDS domains.

An in-depth evaluation of specific assessment tools and assessment tool types was beyond the scope of both the literature review and environmental scan due to time constraints.

4. Results and Discussion

> *—Programs of Assessment— "Each single assessment is a biopsy and a series of biopsies will provide a more complete, more accurate picture" (van der Vleuten)*[1]

The primary output of the research is a Summary of Assessment Tools for competency-based medical education. The matrix, found in Table 1, lists various types of assessment tools, along with their applicability for assessment of each CanMEDs role, utility for summative and formative assessment, strengths and limitations, and references. In addition, where feasible, specific examples of the type of tool are given. In the reference section key references are arranged by tool (Key References by Tool) and by general discussion of assessment methods (Key General References), as well as in a more detailed listing of relevant articles.

The research also produced a long list of key areas for consideration in the development and implementation of a program of assessment for competence-based medical education.

Table 1 Summary of Assessment Tools (Green = well suited to evaluate role; yellow = might be suitable to evaluate role; red = not suitable to use for this role)

EVALUATION / ASSESSMENT TOOLS	EXAMPLE OF TOOL	CanMEDS DOMAIN							TYPE OF USE		STRENGTHS	LIMITATIONS
		MEx	Com	Coll	Lea	Adv	Sch	Prf	Form	Sum		
Written exercises												
Multiple-choice questions											Can assess many content areas in relatively little time, high reliability, can be graded by computer	Difficult to write, especially in certain content areas; can result in cueing; can seem artificial and removed from real situations
Key-feature and script-concordance questions											Assess clinical problem-solving ability, avoid cueing, can be graded by computer	Not yet proven to transfer to real-life situations that require clinical reasoning
Short-answer questions											Avoid cueing, assess interpretation and problem-solving ability	Reliability dependent on training of graders
Structured essays											Avoid cueing, use higher-order cognitive processes	Time-consuming to grade, must work to establish interrater reliability, long testing time required to encompass multiple domains
Clinical Setting Assessments												
Global ratings with comments at end of rotation	ITER, FITER										Can be constructed and completed quickly and easily; use of multiple independent raters can overcome some variability due to subjectivity	Often based on second-hand reports and case presentations rather than on direct observation, subjective, rater may bias scores to extremes or may avoid using extremes, less reproducible for non-medical expert roles
Structured direct observation with ratings checklists	Mini-CEX, P-Mex, O-Score, Daily Encounter Forms										Feedback provided by credible experts	Selective rather than habitual behaviours observed, relatively time-consuming

EVALUATION / ASSESSMENT TOOLS	EXAMPLE OF TOOL	CanMEDS DOMAIN							TYPE OF USE		STRENGTHS	LIMITATIONS
		MEx	Com	Coll	Lea	Adv	Sch	Prf	Form	Sum		
Standardized oral examinations											Assesses clinical decision-making & application of medical knowledge, feedback by credible experts	Subjective, sex and race bias has been reported, time consuming, require training of examiners, summative assessments need two or more examiners
Case logs											Useful for determining scope of patient care experience, regular review can help the resident track what cases or procedures must be sought out to meet learning objectives	Numbers reported do not necessarily indicate competence
Medical record review, consult letter review											Can provide evidence about clinical decision-making, follow-through in patient management, advocacy, appropriate use of resources	Retrospective and feedback may not be timely, requires agreed-upon standards of care and rater training, outcomes may reflect health care team rather than resident decisions
Clinical simulations												
Standardized patients / OSCE	OSCE, OSATS										Tailored to educational goals; reliable, consistent case presentation and ratings; can be observed by faculty or standardized patients; realistic; useful for measuring specific clinical skills and abilities including physical exam, history taking, communication, generating differential diagnosis, clinical decision making	Timing and setting may seem artificial, require suspension of disbelief, checklists may penalize examinees who use shortcuts, expensive
Incognito standardized patients											Very realistic, most accurate way of assessing clinician's behaviour	Requires prior consent, logistically challenging, expensive

EVALUATION / ASSESSMENT TOOLS	EXAMPLE OF TOOL	CanMEDS DOMAIN							TYPE OF USE		STRENGTHS	LIMITATIONS
		MEx	Com	Coll	Lea	Adv	Sch	Prf	Form	Sum		
High-technology simulations	Mannequins, Virtual Reality Simulators										Tailored to educational goals, can be observed by faculty, often realistic and credible	Timing & setting may seem artificial, require suspension of disbelief, checklists may penalize examinees using shortcuts, expensive
Multisource ("360 degree") assessments	SPRAT, PAR											
Peer assessments											Ratings encompass habitual behaviours, credible source, correlates with future academic & clinical performance	Confidentiality, anonymity, and trainee buy-in essential
Allied Health Assessments											Ratings encompass habitual behaviours, credible source	Provide global impressions rather than analysis of specific behaviours, ratings generally high with little variability
Patient assessments	ABIM patient satisfaction questionnaire										Can be a credible source of assessment	Provide global impressions rather than analysis of specific behaviours, ratings generally high with little variability, patient literacy may be inadequate, may be difficult to collect sufficient responses for reliable data, may be difficult to separate resident performance from that of health care system
Self-assessments											Foster reflection and development of learning plans	Do not accurately describe actual behaviour unless training and feedback provided
Portfolios	MAINCERT										Foster reflection and development of learning plans, accommodate evidence of learning from a range of different contexts, based on real experience of the learner	Learner selects best case material, time-consuming to prepare and review

* Table 1 was prepared with the guidance of Dr. Susan Glover Takahashi and Dr. Marla Nayer, using additional resources [2-4]

4.1. Key Considerations when utilizing assessment tools for Competency-Based Medical Education

4.1.1. Context

Competency is a complex, integrated set of behaviours built on knowledge, skills, and attitudes.

Assessment in Competency-Based Medical Education (CBME) is mainly formative rather than summative, is criterion-referenced rather than norm-referenced, is authentic as it is workplace-based, relies on direct observation in the clinical setting, and measures specific well-defined tasks of the profession.

Competence is not something one can simply check off. The fact that every candidate who passes a minimal competence exam is effectively labelled competent overlooks the reality that:

- there is always considerable variability of performance within the passing range,
- even the top performers have room for improvement, and
- knowledge and skill are subject to drift and deterioration (decay) over time.

4.1.2. Framing competencies

Each specialty in the medical profession uniquely defines a “competent” practitioner.

This necessitates a strategic planning phase of identifying and defining the competencies needed for professional practice.

The critical knowledge, skills, and attitudes underpinning each competency need to be clearly written and measurable in order to produce anchors that reflect the achievement of that competency.

Core competencies should be: both specific and comprehensive, durable, trainable, measurable, and related to professional activities, as well as being connected to other competencies. They should include three facets of competence: knowledge, attitudes, and skills.

4.1.3. Mastery learning

Mastery learning is a hybrid approach to competency-based education that emphasizes learners’ achievement of consistently high levels of performance within competency-based education programs. As such, it addresses one criticism of CBME—that CBME typically evaluates a basic (minimal) level of skill.

4.1.4. Resident development of competence

Development of competence requires residents to have adequate exposure to the tasks of the specialty and graduated supervision with guided development of skills, as well as assessment of competence.

Experience. A prerequisite to developing competence in the tasks of a profession is exposure to these tasks. An experience matrix codifies the different procedures/conditions with which a resident is expected to be familiar in a rotation and logs their exposure, as well as to the degree of supervision needed in the encounter.

Portfolios play an important role in both training and evaluation. They provide a practical approach to measuring competence and documenting professional development. Portfolios promote active engagement of the learner and enable residents to be proactive in driving their learning and professional development.

Meta-Competencies. Faculty need to evaluate meta-competency — that is, to recognise the complex mix of individual knowledge, skills, and attitudes, as well as cultural and social context - required for safe and effective practice in actual healthcare environments. Evaluation of meta-competency addresses the ability to competently perform in a universe of similar situations and allows observed performance to be extrapolated to performance in practice situations that are not directly evaluated. Such evaluation requires criterion-referenced, not norm-referenced, assessment standards. This requires careful delineation of the methods, tools, and processes used to generate information about the learner's readiness to progress.

Features of an Effective Clinical Performance Assessment:

- Emphasizes the primacy of learning as an integral part of assessment.
- Provides timely and frequent feedback.
- Assesses performance over longitudinal experiences rather than short blocks/rotations.
- Enables integration across stages of training and practice.
- Authentically links assessment and practice.
- Includes self-assessments.
- Includes assessments from a range of assessors who are in a position to give a relevant judgment of one or more aspects of the resident's performance, including peer assessors.
- Emphasizes healthcare processes and outcomes, including strengthening the ability of the assessments to predict who will perform well against those outcomes and who will further develop in their ability after training.
- Shifts accountability in a model of shared responsibility between the individual and the educational system.
- Embeds continuous education integrated across various stages of training and practice.
- For formative evaluations, does not focus exclusively on the pass-fail cut-point, which removes any disincentive for disclosing difficulties.

4.1.5. Validity

As always, careful consideration and rigorous testing is required with new CBME programs and new instruments to ensure that the recorded assessment actually measures what we expect it

to be measuring, and to the correct degree. At the same time, consideration must be given to newer discourses on validity. One emphasizes validity as an argument-based evidentiary-chain, where evidence is presented to support or refute the interpretation of assessment results".[5] In this case a validation process is used to verify that there is sufficient evidence and that use of the tool was appropriate for making the interpretations desired about the learners' performances. In another current discourse validity is considered as a social imperative, and this requires the consideration for the consequences of assessment.[6] This perspective is characterized by a "bird's eye view" of assessment that does not simply consider the tools being used but also the broader issues of the individuals and society.

4.1.6. Reliability

In the context of CBME, two very different types of "reliability" are important.

One type of reliability refers to the assessment tools and processes. This is the standard research methodology definition of reliability, which is usually theoretical or hypothetical in nature: If the evaluation had been conducted by a different person or in a different setting, would the same result have been recorded?

The second type of reliability, discussed in some depth in the academic literature, refers to the reliability of a candidate's competency or performance across scenarios and over time.

In any given CBME assessment scenario, one, both, or neither type of reliability may be present.

4.1.7. Sample of assessments

Competence is not something one can check off. Residents must be tested in multiple settings, using a variety of tools, by multiple assessors, over a period of time. In essence, a 'sample' of the resident's capability is required in order to confidently generalize as to a resident's "competency."

Increasing the number of observations and observers improves the reliability of rater-based assessments. Moreover, increasing the number of observers improves reliability to a greater extent than increasing the number of assessments from any one observer.

Because assessing competencies is such a complex endeavour, it is necessary to use a variety of instruments.

The number of observations (sample size) needed for adequate reliability is highly context-dependent and tool-dependent. When patient assessments are used, more assessments may be needed.

When mini–clinical evaluation exercises (Mini CEX) are used to assess a variety of different clinical encounters over a period of time, with a number of different assessors, the Mini CEXs can provide a nuanced, comprehensive, valid and reliable measure of a resident's performance.

By using a combination of observational assessment methods in a portfolio, a valid and reliable summative decision can be made with a feasible number of assessments — for example, seven mini-CEXs, eight direct observations of practical skills (DOPS), and one multi-source feedback (MSF).

4.1.8. Consequences of the assessment process

It is possible for the *process* of assessment to influence learning and learning outcomes. As such, assessment can be viewed as an "intervention" with potential costs, benefits, and harms. (In the context of CBME, "consequences" are not a synonym for impact or outcome.) Consequences may arise not only from the assessment itself, but also from decisions and actions based on the assessment.

Potential consequences of the assessment process may be beneficial or harmful, and intended or unintended. For example, the presence of a senior colleague (assessor) may provide a resident with confidence such that the resident performs much better in the assessment than in their usual day-to-day performance. In another simple example, if an assessment is structured in such a way as to produce a high level of anxiety in the resident, the resident may perform much more poorly in the assessment than in their usual day-to-day performance. Consequences of a CBME assessment "intervention" can affect both the validity and the reliability of an assessment.

4.1.9. Limited cognitive capacity for assessing

Human cognitive capacity is limited, and may affect ratings when the rate demand is high (e.g. volume of ratings; long or challenging assessment task). As their workload gets heavier, it has been shown that raters often find ways to reduce the task, rather than ways to improve performance. Mental functions required for rating complex performances are limited by capacity, which leads to judgment error on the part of the raters, which can also artificially inflate the inter-item consistency scores when raters give the same or similar score for each item.

5. Summary

Top Four Take-Away Messages

- The rich academic literature on assessment tools for competency-based medical education strongly suggests the need to rigorously test the reliability and validity of assessment tools, within specific contexts, *in situ,* with actual residents and teachers.
- Even so, there appears to be a high level of confidence in many types of assessment tools, and in specific assessment tools, such that the unknown validity and reliability of actual tools in a particular context need not be a barrier to implementation.
- Training residents and teachers in the use of specific assessment tools will help improve the tools' reliabilities.

- Assessment programs must consider assessment tools that are best suited for the intended purpose and provide consistent and valid information about resident performance.

6. References

6.1. References in Paper

1. van der Vleuten CPM, Schulwirth L, Scheele F, Driessen EW, Hodges B. The Assessment of Professional Competence: building blocks for theory development. *Best Practice and Research clinical Obstetrics and Gynecology.* 2010;24(6):703-719.
2. Accreditation Council for Graduate Medical Education (ACGME), American Board of Medical Specialties (ABMS). *Toolbox of Assessment Methods; A product of the joint initiative of the ACGME Outcome Project Ver 1.1.* Ver 1.1 ed. USA: ACGME and ABMS; 2000.
3. Epstein RM. Assessment in Medical Education. *New England Journal of Medicine.* 2007;356(4):387.
4. Glover Takahashi S, Abbott C, Oswald A, Frank JR, eds. *CanMEDS Teaching and Assessment Tools Guide.* Ottawa, ON: Royal College of Physicians and Surgeons of Canada; 2015.
5. Downing SM. Validity: on the meaningful interpretation of assessment data. *Medical Education.* 2003;37:830-837.
6. St-Onge C, Young M, Eva KW, Hodges B. Validity: one word with a plurality of meanings. *Advances in Health Science Education.* 2016;early online.
7. Pangaro L, Ten Cate O. Frameworks for learner assessment in medicine: AMEE Guide No. 78. *Medical Teacher.* 2013;35(6):e1197-e1210.

6.2. Key References by Tool

6.2.1. Written Exercises

Case, S. M. and D. B. Swanson. (1998). "Writing Written Test Questions for the Basic and Clinical Sciences." 3rd Edition. Retrieved May 17, 2017, from http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf.

6.2.2. Clinical Setting Assessments

Moonen-van Loon, J. M. W., K. Overeem, H. H. L. M. Donkers, C. Van Der Vleuten and E. W. Diessen (2013). "Composite reliability of a workplace-based assessment toolbox for postgraduate medical education." Advances in Health Science Education **18**(5): 1087-1102.

Tugwell, P. and C. Dok (1985). Medical record review. Assessing Clinical Competence. V. Neufeld and G. Norman. NY, NY, Springer Publishing Company. **7:** 142-182.

6.2.3. Clinical Simulations

Turner J, Dankoski M. Objective Structured Clinical Exams: A Critical Review. *Family Medicine.* 2008;40(8):574-578.

Boulet, J. R. (2008). "Summative Assessment in Medicine: The Promise of Simulation for High-stakes Evaluation." Academic Emergency Medicine **15**: 1017-1024.

6.2.4. Multisource (360 degree) assessments

Archer, J., M. McGraw and H. Davies (2010). "Assuring validity of multisource feedback in a national programme." Postgrad Med J 86: 526-531.

6.2.5. Peer assessments

Thomas, P., K. Gebo and D. Hellmann (1999). " A pilot study of peer review in residency training." Journal of General Internal Medicine **14**: 551-554.

6.2.6. Patient assessments

Challis, M. (1999). "AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education." Medical Teacher **21**(4): 370-386.

6.3. General References

1. Archer, J., M. McGraw, and H. Davies, *Assuring validity of multisource feedback in a national programme.* Postgrad Med J, 2010. **86**: p. 526-531.

2. Baartman, L.K.J., *The wheel of competency assessment: Presenting quality criteria for competency assessment programs.* Studies in Educational Evaluation, 2006. **32**(2): p. 153.

3. Baker, K., *Determining Resident Clinical Performance: Getting Beyond the Noise.* Anaesthesiology, 2011. **115**(4): p. 862-878.

4. Bogo, M., et al., *Adapting objective structured clinical examinations to assess social work students' performance and reflections.* Journal of Social Work Education, 2011. **47**(1): p. 5-18.

5. Boulet, J.R., *Summative Assessment in Medicine: The Promise of Simulation for High-stakes Evaluation.* Academic Emergency Medicine, 2008. **15**: p. 1017-1024.

6. Brannick, M., H. Erol-Korkmaz and M. Prewett (2011). "A systematic review of the reliability of objective structured clinical examination scores." Medical Education **45**: 1181-1189.

7. Carraccio, C. and R. Englander (2013). "From Flexner to competencies: reflections on a decade and the journey ahead." Academic Medicine **88**(8): 1067-1073.

8. Case, S. M. and D. B. Swanson. (1998). "Writing Written Test Questions for the Basic and Clinical Sciences." 3rd Edition. Retrieved May 17, 2017, from http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf.

9. Challis, M. (1999). "AMEE Medical Education Guide No. 11 (revised): Portfolio-based learning and assessment in medical education." Medical Teacher **21**(4): 370-386.

10. Cook, D. A. and M. Lineberry (2016). "Consequences Validity Evidence: Evaluating the Impact of Educational Assessments." Academic Medicine **91**(6): 785-795.

11. Cook, D. A., R. Brydges, B. Zendejas and S. J. Hamstra (2013). "Technology-enhanced simulation to assess health professionals: A systematic review of validity evidence, research methods, and reporting quality." Academic Medicine **88**: 872-883.

12. Cruess, R., J. H. McIlroy, S. Cruess, S. Ginsburg and Y. Steinert (2006). "The Professionalism Mini-Evaluation Exercise: A Preliminary Investigation." Academic Medicine **81**(10): S74-S78.

13. Department of Family Medicine and University of Ottawa. "Learning Strategies Mapped to Competency Area." Retrieved January 1, 2017, from http://www.academicsupportplan.com/Documents/Grid.pdf .

14. Dreyfus, H. L. and S. E. Dreyfus (1986). Mind over machine: The power of human intuition and expertise in the era of the computer. New York, The Free Press.

15. Driessen, E. and F. Scheele (2013). "What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research." Medical Teacher **35**: 569-574.

16. Driessen, E., C. Van Der Vleuten, L. Schuwirth, J. VanTarwijk and J. Vermut (2005). "The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study." Medical Education **39**(2): 214-220.

17. Duffy, F., G. Gordon, G. Whelan, Cole-Kelly, R. Frankel, N. Buffone, S. Lofton, M. Wallace, L. Goode, L. Langdon and Participants in the American Academy on Physician and Patient's Conference on Education and Evaluation of Competence in Communication and Interpersonal Skills (2004). "Assessing competence in communication and interpersonal skills: the Kalamazoo II report." Academic Medicine **79**(6): 495-507.

18. Eva, K. W., G. Bordage, C. Campbell, R. Galbraith, S. Ginsburg, E. Holmboe and G. Regehr (2016). "Towards a program of assessment for health professionals: from training into practice." Advances in Health Science Education **21**(4): 897-913.

19. Hatala, R., D. A. Cook, R. Brydges and R. Hawkins (2015). "Constructing a validity argument for the Objective Structured Assessment of Technical Skills (OSATS): A systematic review of validity evidence." Advances in Health Science Education **20**(5): 1149-1175.

20. Iobst, W., J. Sherbino and O. Ten Cate (2010). "Competency-based Medical Education in postgraduate medical education." Medical Teacher **32**: 651-656.

21. Issenberg, S. and W. McGaghie (2013). Looking to the future. International Best Practices for Evaluation in the Health Professions. W. McGaghie. London, UK, Radcliffe Publishing Ltd**:** 341-359.

22. Ketteler, E. R., E. D. Auyang, K. E. Beard, E. L.McBride, R. McKee, J. C. Russell, N. L.Szoka and M. T. Nelson (2014). "Competency Champions in the Clinical Competency Committee." Journal of Surgical Education **71**(1): 36-38.

23. Lineberry, M., Y. S. Park, D. A. Cook and R. Yudkowsky (2015). "Making the Case for Mastery Learning Assessments: Key Issues in Validation and Justification." Academic Medicine **90**(11): 1445-1450.

24. Lynch, D., P. Suurdyk and A. Eisner (2004). "Assessing professionalism: a review of the literature." Medical Teacher **26**(4): 366-373.

25. MacEwan, M. J., N. L. Dudek, T. J. Wood and W. T. Gofton (2016). "Continued Validation of the O-SCORE (Ottawa Surgical Competency Operating Room Evaluation): Use in the Simulated Environment." Teaching & Learning in Medicine **28**(1): 72-79.

26. Martin, J., G. Regehr and R. K. Reznick (1997). "Objective structured assessment of technical skill (OSATS) for surgical residents." British Journal of Surgery **84**(2): 273-278.

27. McGaghie, W. C. (2015). "Mastery Learning: It Is Time for Medical Education to Join the 21st Century." Academic Medicine **90**(11): 1438-1441.

28. Meade, L. B., S. H. Borden, P. McArdle, M. J. Rosenblum, M. S. Picchioni and K. T. Hinchley (2012). "From theory to actual practice: Creation and application of milestones in an internal medicine residency program." Medical Teacher **34**(9): 717-723.

29. Memon, M. A., J. G. Rowland and B. Memon (2010). " Oral assessment and postgraduate medical examinations: establishing conditions for validity, reliability and fairness." Advances in Health Science Education **15**: 277-289.

30. Moonen-van Loon, J. M. W., K. Overeem, H. H. L. M. Donkers, C. Van Der Vleuten and E. W. Diessen (2013). "Composite reliability of a workplace-based assessment toolbox for postgraduate medical education." Advances in Health Science Education **18**(5): 1087-1102.

31. Pangaro, L. and O. Ten Cate (2013). "Frameworks for learner assessment in medicine: AMEE Guide No. 78." Medical Teacher **35**(6): e1197-e1210.

32. Regehr, G., H. MacRae, R. K. Reznick and D. Szalay (1998). "Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination." Academic Medicine **73**(9): 993-997.

33. Regehr, G., K. W. Eva, S. Ginsburg, Y. Halwani and R. Sidhu (2011). Assessment in Postgraduate Medical Education: Trends and Issues in Assessment in the Workplace, Association of Faculties of Medicine of Canada (AFMC).

34. Scheele, F., P. Teunissen, S. V. Luijk, E. Heineman, L. Fluit, H. Mulder, A. Meininger, M. Wijnen-Meijer, G. Glas, H. Sluiter and T. Hummel (2008). "Introducing competency-based postgraduate medical education in the Netherlands." Medical teacher **30**(3): 248-253.

35. Schuwirth, L. W. T. and C. P. M. van der Vleuten (2011). "General overview of the theories used in assessment: AMEE Guide No. 57." Medical teacher **33**(10): 783-797.

36. Shaneyfelt, T., K. Baum, D. Bell, D. Feldstein, T. Houston, S. Kaatz, C. Whelan and M. Green (2006). "Instruments for evaluating education in evidence-based practice: a systematic review." JAMA **296**(9): 1116-1127.

37. Tamblyn, R., S. Benaroya, L. Snell, P. McLeod, B. Schnarch and M. Abrahamowicz (1994). "The feasibility and value of using patient satisfaction ratings to evaluate internal medicine residents." Journal of General Internal Medicine **9**(3): 146-152.

38. Tavares, W., S. Ginsburg and K. W. Eva (2016). "Selecting and Simplifying Rater Performance and Behavior when Considering Multiple Competencies." Teaching & Learning in Medicine **28**(1): 41-51.

39. Ten Cate, O. (2005). "Entrustability of Professional Activities and competency-based training." Medical Education **39**: 1176-1177.

40. The Royal Australian College of Physicians (2015). When learning at work is not enough: Embedding medical education in the DNA of organisational systems and physician practice. Australia, The Royal Australian College of Physicians.

41. Thomas, P., K. Gebo and D. Hellmann (1999). " A pilot study of peer review in residency training." Journal of General Internal Medicine **14**: 551-554.

42. Tsugawa, Y. and Y. Tokuda (2009). "Professionalism Mini-Evaluation Exercise for medical residents in Japan: a pilot study." Medical Education **43**: 968-978.

43. Tugwell, P. and C. Dok (1985). Medical record review. Assessing Clinical Competence. V. Neufeld and G. Norman. NY, NY, Springer Publishing Company. **7:** 142-182.

44. Turner, J. and M. Dankoski (2008). "Objective Structured Clinical Exams: A Critical Review." Family Medicine **40**(8): 574-578.

45. van der Vleuten, C. (2015). "Competency-based education is beneficial for professional development." Perspectives on Medical Education **4**: 323-325.

46. van der Vleuten, C. P. M., L. Schulwirth, F. Scheele, E. W. Driessen and B. Hodges (2010). "The Assessment of Professional Competence: building blocks for theory

development." Best Practice and Research clinical Obstetrics and Gynecology **24**(6): 703-719.

47. Van Tartwijk, J. and E. W. Driessen (2009). "Portfolios for assessment and learning; AMEE Guide No. 45 " Medical Teacher **31**(9): 790-801.

48. Veloski, J., S. Fields, J. Boex and L. Blank (2005). "Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002." Academic Medicine **80**(4): 336-370.

49. Waas, W., C. Van Der Vleuten, J. Shatzer and R. Jones (2001). "Assessment of clinical competence." The Lancet **357**(9260): 945-949.

7. Appendix 1: Annotated Bibliography

Regehr, G., K. W. Eva, S. Ginsburg, Y. Halwani and R. Sidhu (2011). Assessment in Postgraduate Medical Education: Trends and Issues in Assessment in the Workplace, Association of Faculties of Medicine of Canada (AFMC).

This review of the literature suggests, identifies, and discusses three high-level issues that the authors believe are of critical importance and must be addressed in the near future.

Eva, K. W., G. Bordage, C. Campbell, R. Galbraith, S. Ginsburg, E. Holmboe and G. Regehr (2016). "Towards a program of assessment for health professionals: from training into practice." Advances in Health Science Education 21(4): 897-913.

This is a reflection paper which casts a critical lens on current assessment practices, and offers insights into ways they might be adapted to ensure alignment with modern conceptions of health professional education, for the ultimate goal of improved healthcare.

Specifically, it highlights the need to overcome unintended consequences of competency-based assessment, to design assessment systems that facilitate performance improvement, and to authentically link assessment and practice.

Tavares, W., S. Ginsburg and K. W. Eva (2016). "Selecting and Simplifying Rater Performance and Behavior when Considering Multiple Competencies." Teaching & Learning in Medicine 28(1): 41-51.

An excellent article that explores the implications of rater fatigue on ratings. Based on an experimental design study of excessive rating demands and informed by a number of theories in cognitive psychology, this study explores the alignment between imposed rating demands/load and inherent human cognitive architecture. It concludes that raters instinctually try to minimize the “load”/burden of the task by focusing on the rating items they feel are most important for rating, rather than considering *individually* each item in a ratings battery.

Pangaro, L. and O. Ten Cate (2013). "Frameworks for learner assessment in medicine: AMEE Guide No. 78." Medical Teacher 35(6): e1197-e1210.

This AMEE Guide makes a distinction between analytic, synthetic, and developmental frameworks. Analytic frameworks deconstruct competence into individual pieces, to evaluate each separately. Synthetic frameworks attempt to view competence holistically, focusing evaluation on the performance in real-world activities. Developmental frameworks focus on stages or milestones in the progression toward competence. Most frameworks have one predominant perspective; some have a hybrid nature.

Post MD Education
UNIVERSITY OF TORONTO

Table 2 from AMEE Guide No. 78: Summary of frameworks for assessment of competence. definitions, examples, assumptions, advantages, and limits[7]

	Analytic	Synthetic	Developmental
Definitions	Divide competence into domains	Combine domains into tasks	Describe progress through levels
Examples	Knowledge-skills-attitudes; ACGME*; CanMEDS**	Entrustable professional activities (EPAs)***; Reporter-interpreter-manager-educator (RIME)§	Dreyfus and Dreyfus (1986) (Novice – Advanced beginner – Competent Expert – Master)
Assumptions	Together the discrete elements equal competence; they can be measured discretely	Complex social tasks require multiple domains applied by the learner simultaneously	There are stages, each one superseding the prior
Advantages	Theoretically covers all aspects; allows discrete assessment allow feedback on specific facets and domains individually	Strong connection with workplace activities; high level of authenticity	Can encompass multi-year training and allow assessment of personal progress of an individual
Limits	Tends to lead to extensive descriptions. Not easily comprehensible by clinicians. Connection with clinical activities can be weak	Holistic assessment may not identify specific reasons for failure to progress	Different domains may be at different levels of proficiency; norm-based evaluation of progress may collide with fixed standards

8. Appendix 2: Examples of Assessment Tool Matrices

Table 1 from Epstein (2007)[3]

Method	Domain	Type of Use	Limitations	Strengths
Written Exercises				
Multiple-choice questions in either single-best-answer or extended matching format	Knowledge, ability to solve problems	Summative assessments within courses or clerkships; nation- al in-service, licensing, and certification examinations	Difficult to write, especially in certain content areas; can result in cueing; can seem artificial and removed from real situations	Can assess many content areas in relatively little time, have high reliability, can be graded by computer
Key-feature and script-concordance questions	Clinical reasoning, problem- solving ability, ability to apply knowledge	National licensing and certification examinations	Not yet proven to transfer to real- life situations that require clinical reasoning	Assess clinical problem-solving ability, avoid cueing, can be graded by computer
Short-answer questions	Ability to interpret diagnostic tests, problem-solving ability, clinical reasoning skills	Summative and formative assessments in courses and clerkships	Reliability dependent on training of graders	Avoid cueing, assess interpretation and problem-solving ability
Structured essays	Synthesis of information, interpretation of medical literature	Preclinical courses, limited use in clerkships	Time-consuming to grade, must work to establish interrater reliability, long testing time required to encompass a variety of domains	Avoid cueing, use higher-order cognitive processes
Assessments by Supervising Clinicians				
Global ratings with comments at end of rotation	Clinical skills, communication, teamwork, presentation skills, organization, work habits	Global summative and some- times formative assessments in clinical rotations	Often based on second-hand re- ports and case presentations rather than on direct observation, subjective	Use of multiple independent raters can overcome some variability due to subjectivity
Structured direct observation with checklists for ratings (e.g., mini–clinical-evaluation exercise or video review)	Communication skills, clinical skills	Limited use in clerkships and residencies, a few board- certification examinations	Selective rather than habitual behaviours observed, relatively time-consuming	Feedback provided by credible experts
Oral examinations	Knowledge, clinical reasoning	Limited use in clerkships and comprehensive medical school assessments, some board- certification examinations	Subjective, sex and race bias has been reported, time-consuming, require training of examiners, summative assessments need two or more examiners	Feedback provided by credible experts

Method	Domain	Type of Use	Limitations	Strengths
Clinical Simulations				
Standardized patients and OSCEs	Some clinical skills, interpersonal behaviour, communication skills	Formative and summative assessments in courses, clerk- ships, medical schools, national licensure examinations, board certification in Canada	Timing and setting may seem artificial, require suspension of disbelief, checklists may penalize examinees who use shortcuts, expensive	Tailored to educational goals; reliable, consistent case presentation and ratings; can be observed by faculty or standardized patients; realistic
Incognito Standardized patients	Actual practice habits	Primarily used in research; some courses, clerkships, and residencies use for formative feedback	Requires prior consent, logistically challenging, expensive	Very realistic, most accurate way of assessing clinician's behavior
High-technology simulations	Procedural skills, teamwork, simulated clinical dilemmas	Formative and some summative assessment	Timing and setting may seem artificial, require suspension of disbelief, checklists may penalize examinees who use shortcuts, expensive	Tailored to educational goals, can be observed by faculty, often realistic and credible
Multisource (360-degree) assessments				
Peer Assessments	Professional demeanor, work habits, interpersonal behavior, teamwork	Formative feedback in courses and comprehensive medical school assessments, formative assessment for board recertification	Confidentiality, anonymity, and trainee buy-in essential	Ratings encompass habitual behaviors, credible source, correlates with future academic and clinical performance
Patient Assessments	Ability to gain patients' trust; patient satisfaction, communication skills	Formative and summative, board recertification, use by insurers to determine	Provide global impressions rather than analysis of specific behaviors, ratings generally high with little variability	Credible source of assessment
Self Assessments	Knowledge, skills, attitudes, beliefs, behaviors	Formative	Do not accurately describe actual behavior unless training and feedback provided	Foster reflection and development of learning plans
Portfolios	All aspects of competence, especially appropriate for practice-based learning and improvement and systems-based practice	Formative and summative uses across curriculum and with- in clerkships and residency programs, used by some U.K. medical schools and specialty boards	Learner selects best case material, time-consuming to prepare and review	Display projects for review, foster reflection and development of learning plans