THE FUTURE OF MEDICAL EDUCATION IN CANADA
POSTGRADUATE PROJECT

L'AVENIR DE L'ÉDUCATION MÉDICALE AU CANADA
PROJET POSTDOCTORAL

13 Assessment in Postgraduate Medical Education: Trends and Issues in Assessment in the Workplace

Lead
Glenn Regehr

Authors
Glenn Regehr
Kevin Eva
Shiphra Ginsburg
Yasmin Halwani
Ravi Sidhu

A Paper Commissioned as part of the Environmental Scan for the Future of Medical Education in Canada Postgraduate Project

This Environmental Scan was commissioned by the Future of Medical Education in Canada Postgraduate (FMEC PG) Project, funded by Health Canada and supported and managed by a consortium comprised of the Association of Faculties of Medicine of Canada (AFMC), the College of Family Physicians of Canada (CFPC), le Collège des médecins du Québec (CMQ) and the Royal College of Physicians and Surgeons of Canada (RCPSC), with the AFMC acting as the project secretariat.

Acknowledgements
The authors wish to acknowledge the support of the University of British Columbia, the University of Toronto and McGill University in this study.

How to cite this paper: Regehr G, Eva K, Ginsburg S, Halwani Y, Sidhu R. Assessment in Postgraduate Medical Education: Trends and Issues in Assessment in the Workplace. Members of the FMEC PG consortium; 2011.

Published by: members of the FMEC PG consortium.

Executive Summary

Despite the impressively large set of clinical assessment tools described in the medical education literature, even the most optimistic reviewers of this literature acknowledge that we are not yet well equipped to effectively evaluate clinical competence. However, simply increasing the *number* of tools may not be the solution to achieving the sort of coherent and comprehensive evaluation strategy being promoted by many authors in the literature. Rather there are several key debates occurring in the literature, which would suggest that the current insufficiencies in our evaluation process have more to do with the conceptualization of the process than with the lack of tools to enact it. This paper is one of 24 papers commissioned for the Future of Medical Education in Canada Postgraduate (FMEC PG) Project. This paper summarizes the themes arising from the extensive number and variety of research papers, reviews and informed commentaries available in the literature. It describes the state-of-the-art regarding the current set of evaluation tools, but also highlights the key debates about what we can (and should be trying to) achieve with our evaluation tools and approaches.

Broadly, the review revealed three key issues that the authors believe are of critical importance and must be addressed in the near future:

1. At the largest level, the community would do well to be cautious in assuming that "the right" list of separate competencies will be a sufficient operational definition of what it is to be a physician. As we move forward toward a comprehensive and coherent assessment strategy, we might do well to stay open to other conceptualizations of physician development (such as "conceptualization of practice" and/or identity formation) and the implications of these for evaluating preparedness for practice in the field.

2. In addressing the weaknesses of the current in-training evaluation model, the community would do well to move beyond faculty development strategies that teach supervisors how to use the tools, and address the administrative, professional, and cognitive barriers that impede supervisors' ability to formally codify and document their expert assessments of their trainees.

3. In addressing the mandate of our assessment programs to offer meaningful feedback to trainees, the community would do well to find mechanisms to engage residents in the assessment process such that they are able to appreciate and incorporate corrective feedback into their professional development.

Introduction

Assessment has long played a central role in medical education at all levels of training. There are likely two broad reasons for this prominence. First, from a quality assurance perspective, assessment is seen as the primary mechanism by which both institutions (schools, universities and hospitals) and organizations (professional licensing and certifying bodies) can assure the public of acceptable levels of competence among their trainees and practitioners. Second, from an educational perspective, assessment is seen as a primary mechanism for providing feedback to trainees and practitioners for the purposes of improving performance. As Epstein and Hundert[1] have summarized, "Medical schools, postgraduate training programs and licensing bodies conduct assessments to certify the competence of future practitioners, discriminate among candidates for advanced training, provide motivation and direction for learning and judge the adequacy of training programs." (p.226).

This multifaceted set of goals for assessment has placed it in such a prominent role in education that assessment is often described as "the tail that wags the curriculum dog."[2,3] This is clearly true from the perspective of learners, who will orient and shape their leaning goals towards performing well on tests. Swanson and Case[3] powerfully articulate this effect on trainee learning activities with their colloquial assertion, "Grab students by the tests, and their hearts and minds will follow" (p.83, note 2). However this guiding effect of assessment also applies to curriculum developers, who often define the success of their programs by trainee performance on standardized tests, and who shape their curricular structures to address the requirements of "evidence" as specified in accreditation standards. David Leach[4] articulated this eloquently as the first principle in motivating the movement of the Accreditation Council for Graduate Medical Education (ACGME) from process-oriented accreditation standards to outcome-oriented standards, stating "we tend to improve that which we measure." (p.39)

Perhaps not surprisingly, given the explicit goal of the ACGME Outcomes Project[4], the shift to outcome-oriented accreditation standards has led, in recent years, to an even greater emphasis on the formal assessment of trainees at the postgraduate (as well as the undergraduate) level. And this movement is now being further fuelled by discussions of competency-based training models whereby trainees' progression through training is driven not by temporal considerations, but rather by accomplishment-based criteria. Obviously, the legitimacy of such a model is predicated on the capacity to effectively evaluate the requisite competencies, generating further pressure on assessment systems to fulfill this expectation.

The result of this ongoing (in fact, increasing) prominence of assessment as a cornerstone of both student learning and institutional practices in education is well described by Hodges[5] in the introduction of his literature review on assessment for the FMEC MD environmental scan project: "There is probably no aspect of medical education that is more discussed and debated than assessment. It has its own journals, conferences and grants. ... The literature on assessment in medical education is enormous." (p217). The goal of the current paper, therefore, is not to provide a comprehensive review of the assessment literature, but rather to summarize the themes arising from the extensive number and variety of literature reviews and informed commentaries that already exist.

The Competence/Competency Conundrum – What are we trying to measure

No early 21st century discussion of assessment in postgraduate medical education can be meaningfully conducted without addressing the issue of the "competency" movement in North American medical education. The Canadian version of this movement (the CanMEDS roles) is being elaborated in FMEC PG commissioned paper 15: Integration of CanMEDS Expectations

and Outcomes, and the broader implications of the competency-based movement for postgraduate medical education are being addressed in FMEC PG commissioned paper 19: Innovations, Integration and Implementation Issues in Competency-based Training in Postgraduate Medical Education. Thus, we will not describe the CanMEDS roles or the broader implications of the competency-based movement here. However, it is worth noting that, for the last decade, the competency-based movement in North America has had two major influences on the direction and evolution of assessment development and research.

Most obviously, there has been a bloom of assessment tools whose explicit purpose is to address one or another of the ACGME competencies and/or CanMEDS roles. For example, one can find systematic reviews on evaluation tools to assess specific competencies such as: evidence-based practice[6]; professionalism[7,8]; and interpersonal skills / communication[9]. An effort to elaborate the details of these individual measures is beyond the scope of this document. However, in the next section we attempt to articulate a framework for understanding the various approaches embodied in these efforts.

Second, and more importantly for the purposes of this section, there has been a growing discussion about the advantages and potential dangers of adopting a competency-based approach to assessment, particularly in the postgraduate domain. That is, while many have argued for the systematic development of evaluation tools that address each of the competencies separately and have celebrated the achievements of these efforts to date (cf Green and Holmboe[10]), others have questioned this approach from both evidence-based and theoretical perspectives.

From an evidence-based perspective, some research has raised doubts about our ability to assess individual competencies. For example, based on a systematic review of the assessment literature, Lurie et al[11] concluded that "the peer reviewed literature provides no evidence that current measurement tools can assess the competencies independently of one another" (p301). These conclusions are consistent with a number of related studies that conclude that performance-based evaluation tools tend to load on a single, or at best two (cognitive and interpersonal), dimensions of performance.[12]

In addition to the data-driven concerns regarding the feasibility of separately evaluating the competencies, others have questioned the wisdom of even attempting to "anatomize clinical competence"[13], arguing that such an approach fails to capture residents' ability to integrate the various competencies into a coherent conceptualization of practice that allows them to perform the complex activity of patient care. As articulated by Leach[14] in his discussion of efforts to assess the ACGME competencies, "The relevance of the work is dependent on an integrated version of the competencies, whereas measurement relies on a speciated version of the competencies. The paradox cannot be resolved easily. The more the competencies are specified the less relevant to the whole they become."

This debate regarding the advantages and disadvantages of "anatomizing clinical competence" with our assessment tools is a critical conversation for assessment in postgraduate education. As described earlier, it is clear that assessment drives both trainee learning and institutional curriculum design. Thus, the question of whether we wish to focus assessment most heavily on accomplishment in each of the speciated competencies or on the capacity to integrate these competencies into a coherent conceptualization of practice (what has sometimes been described as the meta-competencies) has the potential to affect how our current trainees learn and ultimately understand what it is to be a physician. Hodges[15] has suggested that implicit within each construction of competence is the possibility of creating "hidden incompetences" that arise through neglect of that which is not emphasized. Thus, it is important to ask what is

missing in the speciated construction of evaluation, and what potential sources of incompetence are inherent within this model of evaluation.

Data Sources for the Assessment of Clinical Competence

As described earlier, an impressive number of individual evaluation tools have been introduced into the medical education literature, each of which has been demonstrated to have reasonable psychometric properties. It is beyond the scope of this report to describe them all here. However, it appears that all these tools can be classified into three broad approaches to documenting the assessment of clinical competence in postgraduate education: 1) formal, structured, circumscribed, *ex vivo* evaluations (such as written examinations, oral examinations, and the Objective Structured Clinical Examination – OSCE); 2) formal, structured, circumscribed, *in situ* observations of clinical performance (such as the Mini Clinical Evaluation Exercise – Mini-CEX); and 3) informal, unstructured, cumulative, *in situ* evaluations (such as the In-training Evaluation Report – ITER).

The assessment of knowledge and technical skills is increasingly being addressed in formal, structured assessment contexts outside the clinical setting. The technology of multiple choice questions, for example, is a long standing tradition in medical education and is well honed to determine the level of knowledge that an individual possesses with high reliability and validity presuming best practices are followed (see, for example, the manual by Case and Swanson, downloadable at *http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf*). Similarly, the Objective Structured Clinical Examination (OSCE), first introduced by Harden and Gleeson in 1979[16], is now over three decades old. Over the past 30 years, the OSCE has been adapted to assess resident level performance in a variety of domains, and has spawned a number of specialty-based derivatives (such as the OSATS[17,18], SAMSS[19], MISTELS[20], and MOMS[21]). With the OSCE's ongoing evolution, and the advent of increasingly sophisticated simulation contexts, there is increasing confidence in our ability to assess many aspects of clinical performance in *ex vivo* settings. Ongoing issues in the use of these simulated environments for both teaching and assessment purposes (such as the role of fidelity in transferability and generalizability to real world environments) is being addressed more extensively in FMEC PG commissioned paper 18: Simulation in Postgraduate Medical Education). However, it is clear from the literature that the reliability of these formal structured *ex vivo* evaluation tools is well established and, increasingly, there is evidence of their validity.

Educators have also developed tools for the formal, structured, circumscribed assessments of performance in the clinical context (for an extensive review of these tools, see Kogan et al[22]). An iconic example of such a tool is the Mini-Clinical Evaluation Exercise (Mini-CEX). [23] The Mini-CEX requires the trainee to interact with a real patient while being explicitly observed and assessed (with documentation of this assessment) by a clinician supervisor. Since its introduction in 1995[23], the Mini-CEX has gained popularity and has been adapted to evaluate specific domains and competencies such as professionalism. Research has been aimed at determining the number of assessment events needed to obtain reliable measures of trainee performance in a given clinical domain. We would note, however, that the enactment of an explicit circumscribed assessment activity in the real clinical context might be considered simply a natural extension of the fidelity issue raised in the OSCE discussion above. That is, such assessment activities have the fidelity of a real clinical case, but they also represent an explicit evaluation moment that will undoubtedly affect the behaviour of the trainee being observed, thereby running the risk of remaining at the "shows how" level of Miller's pyramid rather than moving to the "does" level. It also necessarily focuses on "in the moment" behaviours, and, therefore, may not be well suited to the assessment of competencies that involve integration of patient care and management over time.

Despite the advances in clinical assessment described above, the most common form of clinical evaluation continues to be the In-Training Evaluation Report (ITER). This model of evaluation involves the informal, unstructured observation of trainees over a more extended period and the accumulation of impressions derived from these observations into an integrated (often summative) assessment. The purported strengths of these more integrative, comprehensive evaluations are a perfect complement to the purported weaknesses of the circumscribed evaluation contexts described above: they are based on the observation of naturalistic performance embedded in the real practice setting over an extended observation period. Thus, there are good reasons to believe that the ITER should be an excellent assessment tool. In practice, however, the ITER has been problematic as a mechanism for discriminating between residents,[24,25] and, in particular, has been a weak tool for identifying learners who are experiencing clinical difficulties.[26,27] Despite efforts to improve the scales on which the ITERs are based,[28] and despite various efforts to train faculty to use the scales more effectively,[28] the ITER continues to be problematic as a tool to describe and discriminate resident performance. Because of its central role in clinical evaluation, understanding potential reasons for its poor psychometric properties has been a topic of research in medical education and related health professional fields. The core findings of this literature are described in the next section.

"Barriers" to Effective Documentation of Clinical Performance

Despite all the tools available for direct observation of clinical performance, there continue to be difficulties in discriminating between residents and, in particular, in identifying residents in potential difficulty. Although some argue that the problem of unreliable assessments revolves around insufficient training of faculty[10], the evidence that additional training is an effective remedy for these issues is questionable at best.[30,31,32] Several studies have demonstrated that there is a problem with documenting unsatisfactory performance of trainees[27,28,33], with a survey by the Association of American Medical Colleges (AAMC) of faculty at 10 schools reporting that "unwillingness to record negative evaluations" was rated as a problem by 74.5% of respondents.[34]

It is worth looking, therefore, at some of the mechanisms that might be interfering with a supervisor's ability or willingness to confidently document a trainee's level of clinical competence on an evaluation form. A review of the literature seems to point to three broad areas of difficulty. These might be described as: 1) administrative and political issues; 2) personal and relational issues and 3) cognitive issues.

Dudek et al[35] described several administrative issues that faculty identified as reasons for their "failure to fail" residents that they felt were in potential difficulty clinically. Dominant among these reasons was a sense among the faculty that there were not well established administrative structures to address the situation once identified, including a perceived lack of remediation structures to support the trainees. Consistent with this perception, Hauer et al[36] reported that, based on a survey of 122 US medical schools, institutions with greater "trust in their remediation process" were more likely to enforce consequences of poor performance. Additionally, faculty expressly identified a perceived (or anticipated) lack of administrative support when facing trainee appeals, which resulted in concerns for their own time and credibility during the process. In fact, Cleland et al[37] reported that, far from feeling supported, supervisors may feel a sense of pressure from the institution to pass a low performing trainee. While such concerns expressed by the faculty might be dismissed either as misperceptions or as reflecting a lack of professionalism for their unwillingness to "do the right thing" despite the associated work, it is unlikely that either assurances of available support or challenges to their professionalism are meaningful administrative responses, and Dudek et al [35] suggest the elaboration of explicit administrative support structures to redress these administrative concerns expressed by faculty.

The second set of barriers to authentically documenting trainee performance identified in the literature is related to personal and relational issues. Cleland et al[37], for example, reported supervisors' sense of conflict between their role as a supportive, caring educator and their responsibility to the public and the institution as a gate-keeper of quality. As Bogo et al[38] describe in the social work literature, attempting to maintain these two roles simultaneously is experienced as a "collision of values". They suggest that the educator role is more generally consistent with supervisors' sense of themselves as educators and mentors, with this role feeling particularly dominant in the early stages of a supervision when the supervisors are focused on developing an educational alliance with the trainee. When potential difficulties become apparent, it feels too late to invoke the gate-keeper role, especially since early concerns are usually not well documented[35]. Related to this, Cleland et al[37] also reported that supervisors expressed concerns for the reputation of the trainee and were hesitant to document a negative evaluation if others had not expressed such concerns before. Consistent with this reported lack of confidence in their opinion, supervisors in Dudek et al's[35] study reported being less likely to document failing performance if they were unable to find supporting evidence for their judgements from colleagues. These findings suggest that the lack of success in faculty development around evaluation may be, in part, because the training focuses around "what defines competent" (something the supervisors may be particularly adept at already) rather than on preparing them to manage the conflicting values of the educator and evaluator roles that will arise when faced with poorly performing trainees.

The final set of issues relates to the mechanisms of documentation available to supervisors. That is, there appears to be a discrepancy between the way supervisors seem to construct their cognitive representation of trainees and the way in which they are expected to document this representation. As Lavine et al[39] described, the difference between the high performing and the problematic trainee in a supervisor's mind is often more related to the "when" and "why" than to the "what" of trainee activities. That is, supervisor impressions of a trainee do not necessarily reflect trainee behaviours *per se*, but rather the supervisor's attributions of these behaviours to the trainee's underlying motivation and ability to prioritize activities (such as student learning versus patient care). Consistent with this description, Ginsburg et al[40] reported that, when supervisors are asked to "tell the story" of excellent or problematic residents they have supervised, the factors they often describe do not necessarily map well onto the typical dimensions on evaluation forms. For example, supervisors seem to incorporate a trainee's potential for, and evidence of, improvement to date as well as their current level of performance into their impressions of the trainee. In short, for many (albeit not all) aspects of performance, it is okay to not be great if the trainee understands what great is supposed to look like and is progressing in that direction. More broadly, the literature seems to point to a number of issues that faculty seem to orient around when thinking about a trainee's performance including: 1) the extent to which the trainee can be trusted to manage emergent situations, recognize and react appropriately when she is over her head, and provide an accurate picture of the situation; 2) the extent to which the trainee is flexible and adaptable to changing situations; 3) the extent to which the trainee makes the supervisor's life easier or harder; and 4) the extent to which the trainee "just seems to get it". Importantly for our evaluation systems, such constructions of one's trainee may not translate well to a set of behaviourally anchored five-point rating scales of discrete competencies. As articulated by Lurie in summarizing the results of his systematic review[11] (personal communication repeated here with permission): "We found that the 6 ACGME competencies (which came into being as a result of a complex political process) do not directly correspond to anything that has been empirically measured among trainees. Thus, we concluded that, while the competencies provide a necessary framework for organizing assessments of trainees, it does not appear possible to 'measure the competencies' as naturally-occurring psychological constructs." And consistent with the negative findings of Lurie

et al[11], Crossley and colleagues[41] found that scales designed to reflect dimensions of "developing clinical sophistication" and "independence" (constructs they believe are aligned with supervisors' natural organization of trainee performance) enhance reliability and validity of clinical evaluations made by supervisors, at least in the context of circumscribed *in situ* assessment exercises such as the Mini-CEX.

Thus, it appears that if we are to effectively address the lack of valuable documentation of resident performance in the clinical context, it will likely require a multi-pronged approach. This multi-pronged approach will have to include: 1) developing effective administrative structures that explicitly support and reward faculty for identifying and documenting resident difficulties; 2) preparing faculty for the personal and professional challenges attendant with documenting and discussing their concerns with trainees; and 3) developing assessment tools that allow faculty to effectively document their impressions of residents in a manner that is consistent with the manner in which they represent resident competence cognitively. Without addressing these issues in a meaningful way, it is unlikely that we will be able to improve the value of clinical assessment.

Credibility and Value of Assessment in the Eyes of Participants

Finally, it is worth noting that the vast majority of the research papers on assessment in medical education focus on the gate-keeping role of measurement instruments. By contrast, there is relatively little discussion of the explicit formative value of the assessment process. It is well established in other literatures that one's perception of the feedback given is critical to one's ability to incorporate it and use it effectively[42,43,44,45]. Thus, it is worth exploring the perceptions of participants regarding the value of current assessment instruments for shaping professional development.

In a pair of studies by Watling and colleagues[46,47], there appeared to be a discrepancy between attending physicians' and residents' opinions regarding the value of the evaluation process. That is, faculty supervisors seemed to note a high level of engagement with the ITER process. Although they acknowledged that the assessment process was fragmented due to a lack of time to interact with the trainees, as well as an inconsistency in both the goals of assessment and the standards of achievement, they nonetheless felt that the assessment process was an important reflection of the profession's social accountability as well as their responsibility to the residents. In short, faculty supervisors believed their assessments played a strong role in guiding residents' professional development.[46] In contrast, a widespread negative opinion of the evaluation process was observed amongst residents, with many viewing the ITER as being of limited importance to their professional development due to a lack of engagement with the process.[47] Consistent with Watling et al's findings,[46,47] a survey of surgeons and residents on clinical performance feedback found that 90% of surgeons felt that they were successful in providing effective feedback, while only 16.7% of residents agreed.[48]

Finding mechanisms to effectively engage residents in the evaluation process, therefore, is critical. Watling et al[47] identify several potential mechanisms for enhancing this engagement, perhaps the most important of which is the presence of a trusted and engaged faculty member who has a long-standing professional relationship with the resident and understands both the resident's own set of learning objectives and her own perception of how she might be able to achieve them. Embodied in this description is the recognition that there may be an important difference between *data* (which is often the output of a structured evaluation process and often imparted to a resident with little contextual richness and, therefore, with little meaning) and *feedback*, which is likely to be most effective when treated as a highly personal process that must be negotiated carefully and thoughtfully in the context of a trusting relationship. Thus, our

documented evaluation systems may have the potential to form the basis of effective feedback, but they should likely not be considered feedback in and of themselves.

Summary

In the field of medical education, there is a massive literature in the domain of assessment and an (almost literally) innumerable set of assessment instruments, each designed to evaluate some aspect of clinical performance. These instruments can be classed into three broad categories based on the context of the assessment process: 1) formal, structured, circumscribed, *ex vivo* evaluations; 2) formal, structured, circumscribed, *in situ* observations of clinical performance; and 3) informal, unstructured, cumulative, *in situ* evaluations. Each is likely to have an important role in the sort of coherent and comprehensive evaluation strategy promoted by many authors in the literature. Yet, despite the number of instruments introduced into the literature, even the most optimistic reviewers of this literature seem to acknowledge that the clinical assessment "toolbox is not overflowing with perfect instruments"[10(p789)].

Rather than promoting a further proliferation of instruments aimed at individual competencies, however, our conclusions from this review of the literature suggest a set of larger issues that the community will have to grapple with if the national assessment agenda is to fulfill its dual mandate of assuring competence and providing meaningful feedback for shaping professional development of our trainees. Broadly, three key issues have been identified that we believe are of critical importance and must be addressed in the near future:

1. At the largest level, the community would do well to be cautious in assuming that "the right" list of separate competencies will be a sufficient operational definition of what it is to be a physician. As we move forward toward a comprehensive and coherent assessment strategy, we might do well to stay open to other conceptualizations of physician development (such as "conceptualization of practice" and/or identity formation) and the implications of these for evaluating preparedness for practice in the field.

2. In addressing the weaknesses of the current in-training evaluation model, the community would do well to move beyond faculty development strategies that teach supervisors how to use the tools, and address the administrative, professional, and cognitive barriers that impede supervisors' ability to formally codify and document their expert assessments of their trainees.

3. In addressing the mandate of our assessment programs to offer meaningful feedback to trainees, the community would do well to find mechanisms to engage residents in the assessment process such that they are able to appreciate and incorporate corrective feedback into their professional development.

References

1. Epstein RM, Hundert EM. Defining and assessing professional competence. JAMA. 2002 Jan 9;287(2):226-35.
2. Hargreaves, A. (1989). Curriculum and assessment reform. Milton Keynes: Open University Press.
3. Swanson DB, Case SM. Assessment in basic science instruction: directions for practice and research. Adv Health Sci Educ Theory Pract. 1997;2(1):71-84.
4. Leach DC. Building and assessing competence: the potential for evidence-based graduate medical education. Qual Manag Health Care. 2002 Fall;11(1):39-44.
5. Hodges BD. Assessment and medical education: Major Trends and issues for the future of medical education in Canada. Future of Medical Education in Canada Project – Undergraduate Medical Education Environmental Scan Project, 217-228.
6. Shaneyfelt T, Baum KD, Bell D, Feldstein D, Houston TK, Kaatz S, Whelan C, Green M. Instruments for evaluating education in evidence-based practice: a systematic review. JAMA. 2006 Sep 6;296(9):1116-27
7. Lynch DC, Surdyk PM, Eiser AR. Assessing professionalism: a review of the literature. Med Teach. 2004 Jun;26(4):366-73.
8. Veloski JJ, Fields SK, Boex JR, Blank LL. Measuring professionalism: a review of studies with instruments reported in the literature between 1982 and 2002. Acad Med. 2005 Apr;80(4):366-70
9. Duffy FD, Gordon GH, Whelan G, Cole-Kelly K, Frankel R, Buffone N, Lofton S, Wallace M, Goode L, Langdon L; Participants in the American Academy on Physician and Patient's Conference on Education and Evaluation of Competence in Communication and Interpersonal Skills. Assessing competence in communication and interpersonal skills: the Kalamazoo II report. Acad Med. 2004 Jun;79(6):495-507.
10. Green ML, Holmboe E. Perspective: the ACGME toolbox: half empty or half full? Acad Med. 2010 May;85(5):787-90.
11. Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. Acad Med. 2009 Mar;84(3):301-9.
12. Scheuneman AL, Carley JP, Baker WH. Residency evaluations. Are they worth the effort? Arch Surg. 1994 Oct;129(10):1067-73.
13. Huddle TS, Heudebert GR. Taking apart the art: the risk of anatomizing clinical competence. Acad Med. 2007 Jun;82(6):536-41.
14. Leach DC. ACGME e-Bulletin. Accreditation Council for Graduate Medical Education, 2006.
15. Hodges B. Medical education and the maintenance of incompetence. Med Teach. 2006; 28(8):690-696.
16. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ. 1979 Jan;13(1):41-54.

17. Martin JA, Regehr G, Reznick R, MacRae H, Murnaghan J, Hutchison C, Brown M. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg. 1997 Feb;84(2):273-8.

18. Lentz GM, Mandel LS, Lee D, Gardella C, Melville J, Goff BA.Testing surgical skills of obstetric and gynecologic residents in a bench laboratory setting: validity and reliability. Am J Obstet Gynecol. 2001 Jun;184(7):1462-8

19. Friedlich M, MacRae H, Oandasan I, Tannenbaum D, Batty H, Reznick R, Regehr G. Structured assessment of minor surgical skills (SAMSS) for family medicine residents. Acad Med. 2001 Dec;76(12):1241-6.

20. Fraser SA, Klassen DR, Feldman LS, Ghitulescu GA, Stanbridge D, Fried GM. Evaluating laparoscopic skills: setting the pass/fail score for the MISTELS system. Surg Endosc. 2003 Jun;17(6):964-7. Epub 2003 Mar 28.

21. Mackay S, Datta V, Chang A, Shah J, Kneebone R, Darzi A. Multiple Objective Measures of Skill (MOMS): a new approach to the assessment of technical ability in surgical trainees. Ann Surg. 2003 Aug;238(2):291-300.

22. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 2009 Sep 23;302(12):1316-26.

23. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med. 1995 Nov 15;123(10):795-9.

24. Turnbull J, van Barneveld C. Assessment of Clinical Performance: In-training Evaluation. In: Norman GR, van der Vleuten CPM, Newble DI (eds). International Handbook of Research in Medical Education. London: Kluwer Academic Publishing, 2002:793-810.

25. Gray JD. Global rating scales in residency education. Acad Med. 1996; 71(1):S55-S63.

26. Dudek NL, Marks MB, Regehr G. Failure to fail: The perspectives of clinical supervisors. Acad Med. 2005; 80(10 Suppl):S84-7.

27. Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? Teach Learn Med. 1993; 5(1):10.

28. Speer AJ, Solomon DJ, Ainsworth MA. An innovative evaluation method in an internal medicine clerkship Acad Med. 1996; 71(1 Suppl):S76-8

29. Holmboe ES, Hawkins RE, Huot SJ. Effects of training in direct observation of medical residents' clinical competence: A randomized trial. Ann Intern Med. 2004; 140(11):874-881

30. Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. Medical Education. 1980;14:345-349.

31. Williams RG, Klamen DA, McGaghie WC. Cognitive, Social and Environmental Sources of Bias in Clinical Performance Ratings. Teaching & Learning in Medicine. 2003;15:270-292.

32. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: a randomized, controlled trial. Journal Of General Internal Medicine. 2009;24:74-79.

33. Cohen GS, Henry NL, Dodd PE. A self-study of clinical evaluation in the McMaster clerkship. Med Teach. 1990;12:265–72.

34. Tonesk X, Buchanan RG. An AAMC pilot study by 10 medical schools of clinical evaluation of students. J Med Educ. 1987;62:707–18.

35. Dudek NL, Marks MB, Regehr G. Failure to fail: The perspectives of clinical supervisors. Acad Med. 2005; 80(10 Suppl):S84-7.

36. Hauer KE, Teherani A, Kerr KM, Irby DM, O'Sullivan PS. Consequences within medical schools for students with poor performance on a medical school standardized patient comprehensive assessment. Acad Med. 2009 May;84(5):663-8.

37. Cleland JA, Knight LV, Rees CE, Tracey S, Bond CM. Is it me or is it them? Factors that influence the passing of underperforming students. Med Educ. 2008 Aug;42(8):800-9.

38. Bogo M, Regehr C, Power R, Regehr G. When values collide: Field instructors' experiences of providing feedback and evaluating competence. The Clinical Supervisor. 2007;26(1/2):99-117.

39. Lavine E, Regehr G, Garwood K, Ginsburg S. The role of attribution to clerk factors and contextual factors in supervisors' perceptions of clerks' behaviors. Teach Learn Med. 2004 Fall;16(4):317-22.

40. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. Acad Med. 2010 May;85(5):780-6.

41. Crossley J, Johnson G, Booth J, Wade, W. Good question, good answer: Construct alignment improves the performance of workplace based assessment scales. Medical Education, in press.

42. Yorke M. Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. Higher Education. 2003;45:477-501.

43. Higgins R, Hartley P, Skelton A. Getting the message across: The problem of communicating assessment feedback. Teaching in Higher Education, 2001;6(2):269–274.

44. Lawler EE. The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology. 1967;51:369–381.

45. Ilgen, DR, Fisher CD, Taylor M. Consequences of individual feedback on behaviour in organizations. Journal of Applied Psychology. 1979;64:349–371.

46. Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L. An exploration of faculty perspectives on the in-training evaluation of residents. Acad Med. 2010

47. Watling CJ, Lingard L. Toward meaningful evaluation of medical trainees: the influence of participants' perceptions of the process. Adv Health Sci Educ Theory Pract. 2010 Feb 9. [Epub ahead of print] PMID: 20143260

48. Sender Liberman A, Liberman M. Steinert Y, McLeod P, Meterissian S. Surgery residents and attending surgeons have different perceptions of feedback. Med Teach. 2005;27(5):470–472.

Appendix 1: About the Authors

Glenn Regehr obtained his PhD in cognitive psychology from McMaster University, and during the last year of his PhD, he trained as a research associate in medical education at McMaster University Medical Centre. Currently, he is Professor (Department of Surgery) and Associate Director of Research for the Centre for Health Education Scholarship in the Faculty of Medicine at the University of British Columbia. He also holds a cross appointment with the UBC Faculty of Education, he maintains cross appointments with the University of Toronto Faculties of Education, Medicine, Nursing and Dentistry, and he is associate faculty with the Faculty of Medicine at the University of Ottawa. He regularly consults to a variety of health professional regulatory bodies across Canada and the United States regarding models of continuing professional development. Recent awards include the National Board of Medical Examiners Hubbard Award (2007) and the Medical Council of Canada Outstanding Achievement Award (2008) for his contributions to the evaluation of clinical competence.

Kevin Eva began his appointment as Senior Scientist in CHES and Associate Professor, Director of Educational Research and Scholarship in the Department of Medicine in July 2010. After completing his PhD in Cognitive Psychology in 2001 he became a faculty member in the Department of Clinical Epidemiology and Biostatistics and joined the Program for Educational Research and Development at McMaster University. Kevin is Editor-in-Chief for the journal Medical Education and sits on four other editorial boards. He has been an Affiliated Scholar at the Wilson Centre of the University of Toronto since 1999 and maintains additional appointments as Associate Professor in the School of Health Education at Maastricht University (The Netherlands) and as Visiting Professor at Bern University (Switzerland). Recent awards include the Association of Faculties of Medicine in Canada-GlaxoSmithKline Young Educators Award.

Shiphra Ginsburg is an Associate Professor, Department of Medicine (Respirology), and an Adjunct Scientist, Wilson Centre for Research in Education, both at the University of Toronto. She is also the Director of Education Scholarship and the Director of the Clinician-Educator Training Program for the Department of Medicine, and the co-Director of the Centre for Faculty Development's new certificate course "CoFER" (Core Foundations in Education Research). Her primary research program focuses on issues of understanding and evaluating professionalism in medical education. Other research interests include the evaluation of clinical competence, professional identity formation, education scholarship, and qualitative methodology. Dr. Ginsburg participates in professionalism initiatives at the local, national and international levels, and serves as Deputy Editor at the journal Medical Education, and as a Fellow at the Editorial Board of Academic Medicine. She is the current Kimball Scholar at the American Board of Internal Medicine. In her roles for the department, Dr. Ginsburg enjoys mentoring other faculty members in the development of their own research and scholarship.

Yasmin Halwani is a general surgery resident in the Department of Surgery and a Clinical Educator Fellow in the Centre for Health Education Scholarship at the University of British Columbia. She completed her medical degree at McGill University Throughout her training, she has had an interest in education, and worked on the Core Surgery Resident Education Committee. She is committed to improving the quality of resident education and sits on the General Surgery ADHD and Journal Club committees. She is working with Dr. Ravi Sidhu on implementing a tool to improve intraoperative assessment of residents. She is currently completing her Masters of Adult Education at UBC.

Ravi Sidhu is an Associate Professor in the Department of Surgery at the University of British Columbia. He joined UBC in 2004 after completing his surgical training, an MEd degree, and a two-year education research fellowship at the University of Toronto. He currently holds the position of Director of Clinical Educator Fellowship Program at CHES. His clinical practice in vascular surgery is based at St. Paul's Hospital. Ravi's academic and administrative duties include many different spheres of education scholarship. His research program is based on assessment of postgraduate trainees and practicing physicians and focuses particularly on technical skills. He also functions at the Core Surgery and Vascular Surgery Program Director at UBC, the Director of Postgraduate Education for the Department of Surgery, the Chair of the Education Committee for the Canadian Society for Vascular Surgery, and the Chair of the Research Committee of the Association for Surgical Education.

Appendix 2: Annotated Bibliography

Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33(3):206-14.

This paper provides an excellent summary of the current state of the art for the criteria by which individual assessment tools can and should be judged including: (1) validity or coherence, (2) reproducibility or consistency, (3) equivalence, (4) feasibility, (5) educational effect, (6) catalytic effect, and (7) acceptability. Additionally, the paper describes four further aspects that should be taken into account as the criteria for such tools are refined and improved: (1) the perspectives of patients and the public, (2) the intimate relationship between assessment, feedback, and continued learning, (3) systems of assessment, and (4) accreditation systems. The authors emphasize very strongly the issues of educational and catalytic effect arguing the need for an increasing role for assessment in the learning process.

Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA. 2009 Sep 23;302(12):1316-26.

This paper provides a very comprehensive listing and description of the various tools for the assessment of clinical skills using direct observation that are represented in the literature to date. This systematic review covers 55 reported tools, 21 assessing medical students, 33 assessing postgraduates and fellows and 2 assessing the full continuum of learners. The authors conclude that although there are many tools described in the literature, validity evidence and descriptions of educational outcomes resulting from the assessments are rare.

van der Vleuten CP, Schuwirth LW. Assessing professional competence: from methods to programmes. Med Educ. 2005 Mar;39(3):309-17.

This is an often cited paper that describes several important concepts related to meaningful and authentic assessment. In particular, the authors stress the need not only to use multiple methods of assessment but, importantly, to frame the assessment in the context of an “assessment program” rather than as an isolated set of separate methods. They also promote the situation of assessment in the real world workplace, the use of global and holistic assessments rather than breaking down competency into small units, and the reliance on professional judgement as a basis for the assessment process.