Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency

Shiphra Ginsburg, MD, MEd, Jodi McIlroy, PhD, Olga Oulanova, MA, Kevin Eva, PhD, and Glenn Regehr, PhD

Abstract

Purpose

The drive toward competency-based education frameworks has created a tension between competing desires—for quantified, standardized measures on one hand, and for an authentic representation of what it means to be a good doctor on the other. The purpose of this study was to better understand the tensions that exist between competency frameworks and faculty's real-life experiences in evaluating residents.

Method

Interviews were conducted with 19 experienced internal medicine attendings at two Canadian universities in 2007. Attendings each discussed a specific outstanding, average, and problematic resident they had supervised. Interviews were analyzed using grounded theory.

Results

Eight major themes emerged reflecting how faculty conceptualize residents' performance: knowledge, professionalism, patient interactions, team interactions, systems, disposition, trust, and impact on staff. Attendings' impressions of residents did not seem to result from a linear sum of dimensions; rather, domains idiosyncratically took on variable degrees of importance depending on the resident. Relative deficiencies in outstanding residents could be overlooked, whereas strengths in problematic residents could be discounted. Some constructs (e.g., impact on staff) were not competencies at all; rather, they seem to act as explanations or evidence of attendings' opinions. Standardized evaluation forms might constrain authentic depictions of residents' performance.

Conclusions

Despite concerted efforts to create standardized, objective, competency-based evaluations, the assessment of residents' clinical performance still has a strong subjective influence. Attendings' holistic impressions should not be considered invalid simply because they are subjective. Instead, assessment methods should consider novel ways of accommodating these impressions to improve evaluation.

Acad Med. 2010; 85:780–786.

Medical educators have struggled for decades with the question of how best to evaluate the clinical competence of residents. Many instruments and methods described in the literature have been developed with the goals of being objective and standardized. These instruments strive to minimize subjectivity as a source of construct-irrelevant variance—that is, to prevent the evaluators' subjective opinions from affecting the assessment. Interestingly, most evaluations of clinical performance, such as in-training evaluation reports (ITERs), objective structured clinical exams, and the mini-clinical evaluation exercise (mini-CEX),[1] still rely extensively on evaluators making judgments about trainees' behaviors. The dominant solution to this conundrum has been to try to mitigate these subjective effects through standardization, so that there is some consensus about what is being evaluated (e.g., what specific knowledge, attitudes, or skills are being assessed in a domain such as communication) and what constitutes various levels of performance (e.g., what is meant by such terms as "outstanding performance," "exceeds expectations," and "needs improvement"). These attempts at developing objective, standardized evaluations are driven by a desire to achieve the best possible reliability and validity, as a proxy for objectivity.[2]

At the same time, medical educators (and society) have moved toward the development of a more authentic representation of what it means to be a "good doctor." Some newer assessments, like the mini-CEX, were developed specifically to capture elements of performance more authentically by requiring direct observation of an actual patient encounter in the course of a resident's daily work. At an organizational level, the Canadian Medical Education Directions for Specialists (CanMEDS) project, which began in the mid-1990s, was an "initiative to improve patient care" by articulating a comprehensive definition of the core competencies required for medical practice.[3] Under this framework, physicians are expected to be communicators, collaborators, health advocates, managers, scholars, and professionals. Similarly, the Accreditation Council for Graduate Medical Education (ACGME) Outcomes Project outlined six major competencies, intended to serve as a framework for organizing residency

Dr. Ginsburg is associate professor of respirology and internal medicine, and clinician educator/researcher, Wilson Centre for Research in Education, University Health Network, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

Dr. McIlroy is assistant professor, Department of Medicine, University of Toronto, and affiliated scholar, Wilson Centre for Research in Education, University Health Network, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

Ms. Oulanova is a PhD candidate, Department of Adult Education and Counseling Psychology, Ontario Institute for Studies in Education, University of Toronto, Toronto, Ontario, Canada.

Dr. Eva is associate professor, Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada.

Dr. Regehr is professor, Department of Surgery, and associate director, Centre for Health Education Scholarship, University of British Columbia Faculty of Medicine, Vancouver, BC, Canada.

Correspondence should be addressed to Dr. Ginsburg, Mount Sinai Hospital, 433-600 University Avenue, Toronto, Ontario, Canada, M5G 1X5; e-mail: shiphra.ginsburg@utoronto.ca.

curricula.[4] The project was also meant to assist programs to develop "useful, reliable, and valid methods for assessing attainment of the competencies."

Despite these goals, a recent systematic review of the literature found no assessment methods that can reliably measure the competencies separately from one another as independent constructs.[5] The authors concluded that it is not that the competencies themselves are "wrong" but that assessment measures do not correspond neatly with the framework. In addition, some of the competencies (like systems-based practice) are so dependent on other individuals and external forces that it may not be possible to evaluate a resident separate from the system in which the resident is functioning. It may be that medical educators have blurred the distinction between using competencies as an educational framework to organize and guide learning, and attempting to translate them directly into evaluation tools.

With this in mind, we sought to better understand the apparent tensions that exist between competency frameworks and faculty's experience in the day-to-day evaluation of residents. As part of a larger study to develop a novel evaluation method, we interviewed internal medicine attending physicians to determine what they actually consider when forming opinions about the performance of residents they supervise on a clinical teaching unit. We report here the results of a qualitative analysis of these interviews.

Method

Participants and interviews

Potential participants included all clinical faculty at two Canadian universities (University of Toronto and McMaster University) who had at least two years of experience in teaching and evaluating residents in internal medicine. Sampling was purposive, in that we initially targeted faculty in general internal medicine who attended on the general medical wards at any of our five main teaching hospitals, as they would likely have the most experience in the areas we were exploring. We then targeted faculty from other divisions of internal medicine but who attended in the inpatient setting (e.g., subspecialty inpatient teaching units). Our goal was to interview 20 faculty (15 from the University of Toronto and 5 from McMaster University); we based this number in part on other considerations and goals relevant to the larger study. However, on the basis of our anticipation of a moderate homogeneity of our sample, we expected to reach theoretical saturation on the themes relevant to this aspect of the study with this sample size. We obtained approval from the research ethics boards at both schools.

Faculty attendings were invited to participate by e-mail. Each attending was interviewed for 30 to 60 minutes by the same trained research assistant according to a script developed by the research group. One pilot interview was conducted to test the script; some refinements were made, and that interview was not used in our analysis. During the interviews, attendings were asked to describe (without mentioning names) first a specific outstanding resident they had supervised, then a problematic resident, and finally an average resident. These descriptions could be about any aspect of performance, and there was no attempt to encourage discussion of any particular area. However, descriptions had to be of actual residents rather than generalized opinions. Probes were used where necessary to promote specific descriptions of behaviors (e.g., if the attending stated that the resident was "very professional," the research assistant would ask, "How was that displayed?" or "What did you observe that led to that opinion?"). Probes were also used where necessary to identify areas in which excellent residents revealed deficiencies and problematic residents showed strength. The interviews were audiotaped and transcribed verbatim, with any potentially identifying features removed.

Analysis

Analysis of the interviews began alongside data collection to ensure the interviews were effectively eliciting the types of descriptions we had anticipated and to determine when theoretical saturation had been reached.[6] This occurred after 15 interviews were done at the first university and 4 at the second, resulting in a final sample of 19 interviews that were analyzed using grounded theory. We chose grounded theory for this analysis because we were attempting to develop a theoretical framework to describe how faculty actually thought—and talked—about their residents.[7] Each researcher read the initial transcripts during the open coding process. We then met repeatedly as a group and refined the coding using constant comparison, where categories were further defined, merged, or deleted. Agreement was achieved through consensus, and discussions proceeded until the coding structure was deemed stable. It was then entered into NVivo software, which was used by the research assistant to code all 19 transcripts.[8]

Results

The 19 interviews resulted in 158 pages of text for analysis. The participants were all members of departments of internal medicine at the two universities. Eleven were general internists, with two of those also identifying as geriatricians. The remainder were from respirology (3), cardiology (2), nephrology (2), and infectious diseases (1). There were 11 men and 8 women. Attendings' discussion and descriptions of average (as opposed to outstanding or problematic) residents were quite brief, perhaps because they were discussed at the end of the interviews, and did not contribute meaningfully to our understanding. Therefore, we have not included those data in this analysis.

Analysis of the transcripts resulted in the identification of eight major domains, or themes, that together reflect what faculty attendings consider when forming opinions about their residents: knowledge, professionalism, patient interactions, team interactions, systems, disposition, trust, and impact on staff. Definitions and examples of these domains can be seen in Table 1, and the frequencies with which each was mentioned are presented graphically in Figure 1. As apparent in the table and figure, each was used in approximately equal frequencies in descriptions of both outstanding and problematic residents, with some exceptions that we will address. In addition, the domains discussed could reflect either positive or negative examples within that category (e.g., excellent versus poor knowledge base) regardless of what type of resident

Table 1

Major Themes Arising From Interviews in Which Faculty Attendings in Internal Medicine Discussed Excellent and Problematic Residents, University of Toronto and McMaster University, 2007*

Major theme	Definition	Positive examples	Negative examples
Knowledge			
Knowledge base	Scope or depth of knowledge, use of evidence, etc	Encyclopedic knowledge; good understanding of pathophysiology; surprised you with what they knew	Weak knowledge base; superficial answers to questions; surprising gaps identified
Knowledge translation	Ability to use knowledge, clinical judgment, ability to prioritize	Had an ability to prioritize; evidence-based approach to care; can apply the knowledge they have	Unable to prioritize; had difficulty coming up with a differential diagnosis or management plan
Professionalism			
Work ethic	Overall approach to work in general, reliability, dependability, etc	Came early and stayed late; responsible; punctual; raised the bar; took ownership of patients	Treated work as a 9–5 job; punched the clock; gave too much of herself to work
Work skills	Concrete aspects of performance, including organizational and time management skills	Kept workload under control; "OCD" about details; managed time well	Took too long to assess patients; had poor handwriting;
Response to feedback	Including seeking out feedback, responding, making change	Sought out and integrated feedback to change behavior; responded to feedback	Was defensive or dismissive of feedback; seemed to listen but nothing changed
Conduct	Dress, appearance, language	Polite; professional; friendly; courteous	Made derogatory comments about patients; dressed in sloppy manner
Patient interactions	Including communication, rapport, empathy	Developed great relationships; patients felt relaxed around him; took an interest in patients as people	Was often unclear or confusing to patients; was abrupt; staff had to step in to correct
Team interactions			
Leadership skills	Supervision, teaching, managing the team	Was a role model for team; embraces opportunities to teach; inspires confidence; adjusts supervision according to trainees' abilities	Micromanaged the juniors; bossed others around; did not ask for others' opinions before instituting plans
Team player	Including team relationships, helpfulness, support	Took on fair share of work; offered to help others	Disturbed or disrupted team dynamics; was rude or dismissive to team members; went behind a colleague's back
Allied health interactions	Relationships, communication	Included allied health workers in patient care; treated allied health workers as colleagues	Treated allied health workers as employees; did not respect opinions of non-MDs
Systems	Knowledge of and ability to work within system	Knew who to call; knew how to get things done; understood how to best use resources	Didn't recognize role of internal medicine in the hospital; didn't appreciate how our system works
Disposition			
Attitudes	Perceived attitudes	Curiosity; passion; energy; enthusiasm	Complained about work; was overly quiet; seemed disengaged
Personality	Apparent characteristics of the resident	Intelligent; bright; smart	Lazy; arrogant; unpleasant
Trust			
Believability	Credibility, reliability, honesty	Admitted errors; staff knew whatever the resident said would be correct	Found mistakes that weren't disclosed; staff felt need to double-check everything
Discernment	Ability to see or discern the borders/limits of what one is able to do	Knew when to ask for help; knew when to call; was good at self-assessment and knowing what they needed to look up	Called too frequently; waited too long to call; did not have insight into limitations
Impact on staff	How the resident affected the staff supervisor	Was fun to work with; made my life easier; their approach matched my approach	Caused nightmares; created more work than other residents; I was happy when the rotation was over

* Themes derived through a grounded theory analysis of interview transcripts.

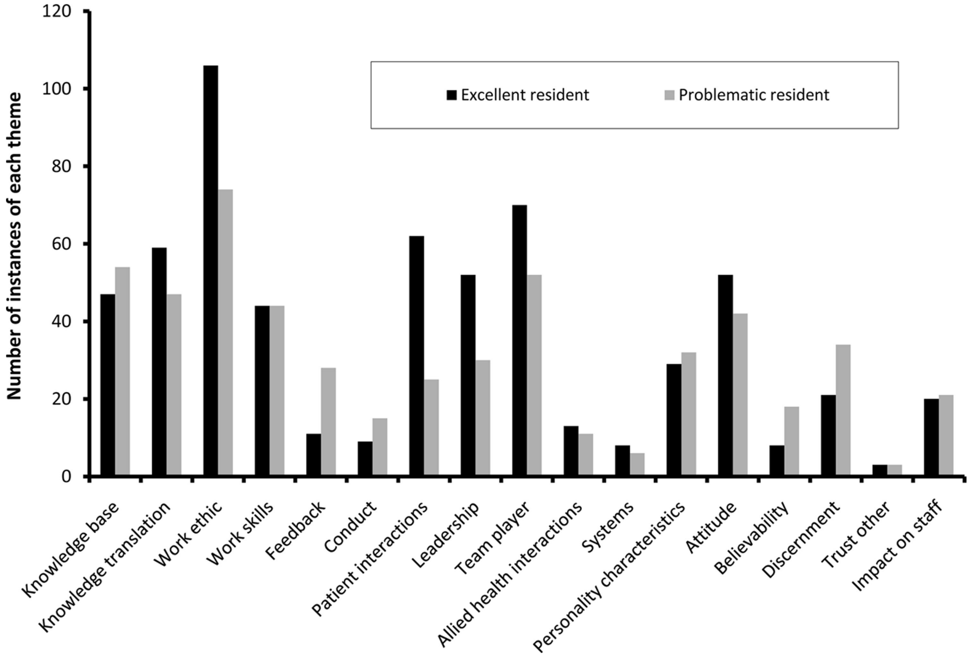

Figure 1 Graphic depiction of the frequency of themes that arose from a grounded theory analysis of transcripts from interviews in which faculty attendings in internal medicine discussed excellent and problematic residents, University of Toronto and McMaster University, 2007. (See Table 1 for information about the themes named at the bottom of the figure.)

they were describing (e.g., outstanding versus problematic).

Domains of performance and how they were discussed

Our first major finding related to the nature of the domains of competence discussed and how they were incorporated into the overall impression of the resident. In terms of *what* they discussed, attendings frequently commented on traditional domains of resident performance, such as knowledge, professionalism, patient and team interactions, and ability to work within a system (see Table 1 for examples). However, in their individual descriptions, attendings did not discuss every domain for every resident, and their discussions did not follow any set order. Rather, they began by discussing what was most relevant to their opinion of that resident's performance, which was different for each resident.

More interestingly, a domain could take on variable importance, depending on other areas of performance for that resident. Each of the themes could be discussed in either positive or negative terms, but this was not necessarily dependent on the type of resident being discussed. For example, in their descriptions of excellent residents, 14 attendings brought up 20 examples of deficiencies in performance, although in almost all cases this information was not offered spontaneously (it arose after the interviewer asked whether that particular outstanding resident had any areas that needed to be improved on). Interestingly, despite such comments as "To be outstanding you have to have outstanding knowledge base, I think. You can be outstanding in everything else but if you don't know enough internal medicine you can't," these relative deficiencies were most often in the area of knowledge base or knowledge translation (n = 9). We heard comments such as "The knowledge base was extensive enough. It wasn't as extensive as some other residents, but it was better applied than in residents who had greater knowledge." Attendings found knowledge problems "easy" to deal with: "It's just an issue of sitting down and spending time." Furthermore, because knowledge itself was seen as being easily accessible ("You don't know what it is, you Google it, you go on any of the online resources—most people have them on a handheld"), it was not considered by most to be a true marker of who is excellent.

It is important to note that it was not just deficits in knowledge that could be overlooked. Attendings also commented on relative deficiencies in otherwise excellent residents' work skills (e.g., could improve on efficiency or dictations), personality or attitude (such as being too quiet or lacking in self-confidence), and discernment (waiting too long to call for advice). Interestingly, three attendings brought up concerns about excellent residents who seemed "too invested" in their work and at risk of burning out. These attendings admitted that they liked this quality because it made their lives easier, but they did express concern for their otherwise excellent trainees.

In discussing problematic residents, 14 attendings mentioned 26 examples of areas in which the residents excelled, but in contrast to their discussions of excellent residents, they usually did not need to be probed for these. In nine interviews, an area of strength was actually noted as one of the first three domains mentioned. Sixteen of these comments related to residents having either an excellent knowledge base (or "book knowledge") or being very bright and intelligent. They also noted residents who had pleasant personalities (n = 5) or good work skills (n = 4).

In sum, attendings seemed to overlook, or excuse, deficiencies in residents they thought of as being outstanding, whereas competence or even excellence in some domains did not "save" other residents from being thought of as problematic. Attendings' impressions did not result from a linear sum of dimensions; further, what was weighted most or least heavily in any one description seemed to be variable and idiosyncratic.

Relative prominence of themes

Our second finding relates to the relative frequencies of the themes, as depicted in Figure 1. (Although the validity of counting comments in qualitative research is controversial, we included numbers here to provide an illustration of the relative prominence of the domains discussed.) Work ethic was by far the most frequently used code in the entire data set and was especially prominent when attendings discussed excellent residents. As one put it, "Some people come in and they just look for work. They love it. Tell them about a case and they are just all over it." Another stated, "He was available, he would always respond. He was proactive in anticipating problems. He did not wait for them to happen; he expected them to develop." Comments about patient communication and leadership were also much more common in discussions of excellent residents, whereas issues of trust and residents' response to feedback arose much more frequently in discussions of problematic residents.

"Noncompetency" constructs

Our third major finding was that attendings elaborated several constructs that affected their opinions of residents that were not in fact competencies at all. Consider, for example, the theme of disposition. Attendings frequently commented on residents' attitudes and personality characteristics, as typified by this explanation of why one attending thought a resident was problematic: "So this person had a demeanor, more reserved, quiet, she seemed disengaged in rounds. She didn't seem to enjoy what we were doing. She didn't want to be there." Or consider these phrases used to describe what made certain residents outstanding: "He had a sense of humor"; "She was just a really nice person"; and "He was not artificial. He was very down to earth." These comments were generally offered as explanations for an attending's opinion about why a particular resident was outstanding or problematic. Similarly, the theme of impact on staff evolved to capture comments attendings made in which their opinion of a resident was shaped by how that resident affected the faculty member's life. These themes were discussed with equal frequency for both outstanding and problematic residents. Again, these comments did not describe a particular area of performance or competency but, rather, were offered as support or as explanation for attendings' stated opinions.

Finally, although attendings certainly discussed elements of residents' performance that can be considered competencies, they themselves did not use the word "competency" at all, and the term "CanMEDS" was not mentioned once during any interview.

Discussion

The development of competency frameworks was intended to serve as a structure for the education of residents, with the overall goal of ensuring competence in all domains considered essential for medical practice. Once these frameworks were developed, it was intended that evaluations would follow to ensure that residents met requirements for each competency. Evaluating the clinical competence of residents is complicated, partly because of competing desires: for feasible, reliable, quantified assessment tools, on the one hand, and, on the other, for an authentic representation of what it means to be a good doctor. Developing assessment instruments to evaluate these "core competencies" has been difficult, as recently reported by Lurie et al.[5] It seems the individual competencies cannot be evaluated separately from one another, and most assessments probably measure a single construct (or several that do not map neatly onto the framework, as supported by our findings). From in-depth interviews with experienced faculty attendings, we think we have found some insights as to why these efforts have not been successful.

One possible reason for these difficulties relates to a growing recognition that many of the desired competencies are in some ways socially determined. For example, an individual's performance related to the ACGME competencies of practice-based learning or systems-based practice is dependent on interactions with other people and the environment. An individual's contribution cannot be easily teased out.[5] Perhaps more important, however, an underlying presupposition still seems to exist that there is a "true score" within an individual (his or her knowledge and skills, for example) that can be measured accurately once the right tools are found. As van der Vleuten, Norman, and Graaf[2,9] argued in two key articles nearly 20 years ago, in our drive toward objectivity we seem to have assumed that subjective measures are inherently unreliable and that reducing bias by removing human inferences from the judgment process would improve precision of evaluations.

That may be true for certain situations (like written exams to test knowledge), but as van der Vleuten and colleagues stressed, the choice of assessment method should be determined by the educational context or by the purpose of the testing situation, not by a blind desire to be as objective or standardized as possible. Perhaps some of the difficulties in evaluating competencies in a clinical setting arise from the fact that the starting point is usually the competency one wants to assess, rather than the context in which it is being observed. For our context—the clinical teaching units in internal medicine—it might make more sense to start with what attendings (i.e., evaluators) actually observe, experience, and can comment on.

Second, others have suggested that faculty supervisors conceptualize trainees' performance according to a set of meta-competencies, within which they consider an individual's performance. For example, Bogo et al[10] found that, as supervisors discussed their outstanding and problematic social work trainees, they would elevate—or discount—the relative importance of a particular domain, depending on their overall opinion of a given trainee. But this is not simply a manifestation of the halo effect, as supervisors did not rate their students as outstanding in all areas. They acknowledged deficits, as our supervisors did, but discounted them. In Bogo and colleagues' study,[10] these descriptions were framed as "but statements"; for instance, an exemplary student's skills in a particular area needed work *but* the supervisor excused it, believing it was simply the result of a lack of formal training in that area. This can be explained by attribution theory, as the supervisor in this example attributed the deficiency to a lack of training (as opposed, say, to laziness or incompetence, which might not have been overlooked).[11] Thus, as supported by our data, a weakness does not necessarily preclude a learner from being considered outstanding. As a corollary to this process, attendings were often dismissive of adequate (or even well-developed) areas of performance in learners they think of as problematic. Thus, consistent with research comparing scores from checklists versus global ratings,[12] the overall impression of the resident is far from a simple linear addition of the various dimensions being assessed, and even a weighting of these dimensions would be unlikely to adequately capture the supervisor's sense of the resident as a clinician-in-training.

This phenomenon was further illustrated by our finding that, when discussing trainees, attendings led with what they found particularly salient about that individual and what was most important in their judgment of that person. This is perhaps not surprising, because in contrast to Bogo's studies, we did not prompt attendings to discuss residents according to a competency framework or in reference to predetermined categories. We explicitly encouraged them to discuss residents' performance in their own language, the way they would speak, for example, with their colleagues. They did not, therefore, address every construct for every resident. In contrast, evaluation instruments are usually designed so that the competencies are presented in a set order, giving approximately equal visual space to each. This order may reveal the residency program's implicit beliefs about the relative importance of each competency, and the equal spacing implies that each should be considered equally for each resident. Our findings suggest that this visual rhetoric is inconsistent with the way faculty actually conceptualize and express their opinions about the performance of their residents. As a result, authentic depictions of residents' performance may be constrained.

Another critical theme that arose in our analysis was a resident's impact on the attending. This theme arose in nearly every interview and was discussed in both positive and negative terms. Although we were initially somewhat hesitant to code this theme, its prominence in the data set made it difficult to ignore. In addition, this phenomenon has been reported elsewhere, including Bogo and colleagues'[10] research and several studies on supervisors' failure to fail underperforming trainees.[13] Of course, we do not propose adding an item on "impact on staff" on residents' ITERs, but its prominence suggests that it should not be ignored as an important contributing factor to supervisors' opinions of residents.

Returning to the concerns of van der Vleuten et al about pitfalls in the pursuit of objectivity, and in light of our current findings, it seems that, in the setting of clinical teaching units, a more subjective approach to evaluation may actually be desirable. In an effort to objectify in this setting, we risk the loss of authenticity. We measure what we think is important, simple, and feasible, but we may have stripped away too much and may not be capturing the essence of what it means to be a good doctor (or at least a good resident, in the eyes of the attending). We agree, therefore, with the assertions of Lurie et al[5] and Bogo et al[10] that competency frameworks may best be thought of as "outside the realm of evaluation"; they are certainly very useful in guiding education, but they may not be the best place to start from for evaluation purposes. It is not that the competency frameworks are unimportant in assessment, but evaluation is more subtle than a sum of the various dimensions. Further, as Hodges[14] has suggested, any model of education and evaluation may result in hidden "side effects." By overemphasizing what we explicitly choose to measure and count, we may fail to recognize—or in some cases may even create—incompetence. Thus, as with any educational model, we should not ignore the potential unintended consequences of competency-based evaluation.

The issues described in the preceding paragraphs cannot be resolved with simple tweaks to the evaluation forms. Differentially weighting the scales, for example, which is often suggested, will not work because it is not the case that one competency is always more important than another. The relative importance of a domain depends not only on the particular individual being described, as discussed above, but also on the particular evaluator, as it has also been shown that idiosyncrasies exist in terms of what individual faculty attendings value.[15] Further, the act of abstracting from observations to interpretations and then translating into numbers on scales has been shown to be problematic, with a resulting loss of authenticity.[16] Promising research in social work has found that evaluations using standardized narrative descriptions of residents' performance, written in the language that clinical supervisors actually use, may be better at picking up borderline performance than traditional, structured evaluation forms.[17] A similar strategy is currently being studied in internal medicine.

As with any qualitative research based on the grounded theory tradition, our findings are meant to generate a new theoretical or conceptual framework for understanding, rather than to test or attempt to disprove a hypothesis. Our findings do have credibility in that the data were gathered and analyzed according to traditional qualitative methods,[18] but because we only included internal medicine faculty, we do not know whether this framework would apply to other medical domains, other types of rotations (e.g., ambulatory), or in other institutions (although we found no differences between the two programs we studied).

Conclusions

Our study reinforces and adds evidence to the growing concern regarding pitfalls in the pursuit of objectivity, by showing that assessment of residents' performance in the clinical setting is still, despite concerted efforts to promote standardized competency frameworks, heavily influenced by the subjective. But this should not be considered a failure. Along with others, we have shown that, as faculty attendings, we cannot separate ourselves as human beings from the role we play as supervisors. Whether it is our demonstrated overreliance on person factors and underappreciation of the situation[19,20] or the subjective opinions and emotional reactions we have about our learners,[13,21] what affects us as human beings affects us as evaluators. Further, as suggested by Leach,[22] the relevance of evaluation 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."

Our faculty, all experienced evaluators, do seem to form integrative impressions of their residents and do not seem to use speciated versions of the competencies when forming their opinions. Rather than rejecting these opinions as "too subjective," we feel that evaluation instruments should accommodate these impressions to improve evaluation in this setting.

Acknowledgments: The authors wish to thank the clinical faculty who participated in this study.

Funding/Support: This study was funded by the Medical Council of Canada.

Other disclosures: None.

Ethical approval: This study was approved by the institutional review boards of the University of Toronto and McMaster University.

Previous presentations: The abstract of an earlier version of this article was presented at the Association of American Medical Colleges/Research in Medical Education Annual Meeting, October 2008, San Antonio, Texas.

References

1 Chaudhry SI, Holmboe ES, Beasley BW. The state of evaluation in internal medicine residency. J Gen Intern Med. 2008;23:1010–1015.

2 van der Vleuten CP, Norman GR, Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. Med Educ. 1991;25:110–118.

3 Frank JR. The CanMEDS 2005 Physician Competency Framework. Ottawa, Ontario, Canada: The Royal College of Physicians and Surgeons of Canada; 2005.

4 Accreditation Council for Graduate Medical Education. Outcomes Project. Available at: http://www.acgme.org/outcome/project/proHome.asp. Accessed January 21, 2010.

5 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;84:301–309.

6 Creswell JW. Data collection. In: Qualitative Inquiry and Research Design. Thousand Oaks, Calif: Sage Publications; 1998:109–135.

7 Kennedy TJ, Lingard L. Making sense of grounded theory in medical education. Med Educ. 2006;40:101–108.

8 NVivo Qualitative Data Analysis Program. Version 8. Melbourne, Australia: QSR International Pty. Ltd.; 2008.

9 Norman GR, van der Vleuten CP, Graaff E. Pitfalls in the pursuit of objectivity: Issues of validity, efficiency and acceptability. Med Educ. 1991;25:119–126.

10 Bogo M, Regehr C, Woodford M, Hughes J, Power R, Regehr G. Beyond competencies: Field instructors' descriptions of student performance. J Soc Work Educ. 2006;42:579–593.

11 Ross L, Nisbett RE. The Person and the Situation: Perspectives of Social Psychology. Boston, Mass: McGraw Hill; 1991.

12 Hodges B, Regehr G, McNaughton N, Tiberius R, Hanson M. OSCE checklists do not capture increasing levels of expertise. Acad Med. 1999;74:1129–1134.

13 Ilott I, Murphy R. Feelings and failing in professional training: The assessor's dilemma. Assess Eval Higher Educ. 1997;22:307–316.

14 Hodges B. Medical education and the maintenance of incompetence. Med Teach. 2006;28:690–696.

15 Williams RG, Klamen DL. See one, do one, teach one—Exploring the core teaching beliefs of medical school faculty. Med Teach. 2006;28:418–424.

16 Ginsburg S, Regehr G, Mylopoulos M. From behaviours to attributions: Further concerns regarding the evaluation of professionalism. Med Educ. 2009;43:414–425.

17 Regehr G, Bogo M, Regehr C, Power R. Can we build a better mousetrap? Improving the measures of practice performance in the field practicum. J Soc Work Educ. 2007;43:327–343.

18 Corbin J, Strauss AL. Criteria for Evaluation. In: Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory. 3rd ed. Thousand Oaks, Calif: Sage Publications, Inc.; 2008:297–312.

19 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;16:403–408.

20 Rees CE, Knight LV. Viewpoint: The trouble with assessing students' professionalism: Theoretical insights from sociocognitive psychology. Acad Med. 2007;82:46–50.

21 Cleland J, Knight LV, Rees CE, Tracey S, Bond CF. Is it me or is it them? Factors that influence the passing of underperforming students. Med Educ. 2008;42:800–809.

22 Leach DC. Six competencies, and the importance of dialogue with the community. ACGME e-Bulletin. August 2006:3. Available at: http://www.acgme.org/acWebsite/bulletin-e/ebu_index.asp. Accessed January 21 2010.