

# The Power of Subjectivity in the Assessment of Medical Trainees

Olle ten Cate, PhD, and Glenn Regehr, PhD

## Abstract

Objectivity in the assessment of students and trainees has been a hallmark of quality since the introduction of multiple-choice items in the 1960s. In medical education, this has extended to the structured examination of clinical skills and workplace-based assessment. Competency-based medical education, a pervasive movement that started roughly around the turn of the century, similarly calls for rigorous, objective assessment to ensure that all medical trainees meet standards to assure quality of health care. At the same time, measures of objectivity,

such as reliability, have consistently shown disappointing results. This raises questions about the extent to which objectivity in such assessments can be ensured.

In fact, the legitimacy of “objective” assessment of individual trainees, particularly in the clinical workplace, may be questioned. Workplaces are highly dynamic and ratings by observers are inherently subjective, as they are based on expert judgment, and experts do not always agree—for good, idiosyncratic, reasons. Thus, efforts to

“objectify” these assessments may be problematically distorting the assessment process itself. In addition, “competence” must meet standards, but it is also context dependent.

Educators are now arriving at the insight that subjective expert judgments by medical professionals are not only unavoidable but actually should be embraced as the core of assessment of medical trainees. This paper elaborates on the case for subjectivity in assessment.

**F**or at least the last 50 years, “objectivity” has been an almost undisputed pursuit for those designing tests, including medical educators.<sup>1</sup> In 1961, De Groot<sup>2(p172)</sup> defined objectivity as judgment “without interference or even potential interference of personal opinions, preferences, modes of observation, views, interests or sentiments.” The search for objectivity in testing was effectively enabled with the introduction of multiple-choice questions (MCQs) in written assessments, as the growing numbers of students in schools and universities stimulated the need for more automatic scoring systems. MCQs offered opportunities to exclude the personal opinions of examiners when assessing student knowledge and, therefore, offered a much-needed response to disputes about fairness and standards.

In medical education, efforts toward objective testing soon extended to

the assessment of more sophisticated training goals beyond factual knowledge. Tests such as the patient management problem<sup>3</sup> and the triple jump exercise<sup>4</sup> were developed to “objectively” assess clinical reasoning and problem-solving skills, although these were plagued with issues of case and context specificity and, therefore, required many hours of testing to achieve appropriate levels of reliability.<sup>5</sup> More successfully, performance-based tests of clinical skills, exemplified by the objective structured clinical examination,<sup>4</sup> were integrated into many undergraduate and postgraduate medical education programs and national examinations.<sup>6,7</sup> With the introduction of mini-clinical evaluation exercises, clinical encounter cards, clinical work sampling, direct observation of procedural skills, and other tools, the search for objectivity also extended to workplace-based assessment.<sup>8</sup> More recently, the search for objective assessment in the clinical workplace was given impetus through the introduction of competency-based medical education (a pervasive movement that started roughly around the turn of the century), with its injunction to move away from unclear and local standards of clinical competence in time-based apprenticeship models toward transparent, structured, outcomes-oriented clinical performance assessment.<sup>9</sup>

There is no question that these innovations in testing have had a valuable impact on the assessment of knowledge and skill in medical education. In this paper, however, we wish to raise some concerns and doubts about this drive toward objectivity as the exclusive, or even best, mechanism for achieving clarity, transparency, fairness, and validity in assessment processes, particularly for undergraduate and postgraduate training in the clinical workplace. We will first suggest that objectivity may not, in fact, represent what current efforts are achieving. Rather, these purported efforts toward “objectivity” might better be understood as negotiating a “shared subjectivity,” a convergence on a single, but still socially constructed, perspective. Although such convergence might achieve consensus, this should not be mistaken for bias-free objectivity. Second, we will suggest that in many situations relevant to effective health care delivery, negotiating a single shared perspective among assessors fails to represent authentic practice, which represents a range of perspectives on competence and interpretations of performance from a multitude of stakeholders. We will end with a discussion of the implications of embracing subjectivity for determining the quality of assessment data and decision making about those being assessed.

Please see the end of this article for information about the authors.

Correspondence should be addressed to Olle (Th.J.) ten Cate, Center for Research and Development of Education, University Medical Center Utrecht, PO Box 85500, 3508 GA Utrecht, the Netherlands; telephone: (+31) 88-75-57010; e-mail: t.j.tencate@umcutrecht.nl.

*Acad Med.* 2019;94:333–337.

First published online October 16, 2018  
doi: 10.1097/ACM.0000000000002495

Copyright © 2018 by the Association of American Medical Colleges

## The Myth of Objectivity

Objectivity, from a positivist, classical test theory perspective, suggests that for each desired learner quality to be measured, a true score exists. With any existing assessment tool, the derived score will deviate from this true score (the “measurement error”). However, in domains such as medicine, in which students must learn to solve problems rather than produce undisputed answers, very often the objectivity of true scores or standards can be questioned. For example, if “objective” is defined as “precluding personal bias of the assessor,” then it could be argued that even large-scale MCQ tests are not objective. Indeed, every test question is created by an individual, often an expert, and represents a value judgment regarding what material is worth testing and sometimes even what the best answer is. Different experts are very likely to differ in their opinions in this regard, as is acknowledged in some test formats.<sup>10,11</sup> Conversations to arrive at a test blueprint, determining the topics to be included and/or the weighting of topics that are included, are seldom straightforward. Similarly, standard setting often requires a highly complex negotiation among experts, not only regarding how much a minimally competent candidate should know but also how to ensure that a test does not fail an inappropriately high proportion of candidates (e.g., if the expert-determined standard on a national licensing examination were found to fail half the candidates, there would undoubtedly be a strong tendency to adjust standards to bring the failure rate in line with expectations). Even the answers to questions (such as the most likely diagnosis) may be subject to negotiation, and consequently some recent test models have tried to incorporate a variety of expert opinions in the scoring rubric.<sup>10</sup> Thus, even in the purest tests of knowledge, the best approximation of objectivity is often simply a (grudging) consensus among a numerical majority of experts, resulting in what might, therefore, be considered a (negotiated) shared subjectivity rather than objectivity.

This negotiation of shared subjectivity becomes even more obvious in rater-based assessments. The consistent demonstration of psychometric weaknesses,<sup>12–15</sup> even when preceptors

rate the same performance,<sup>16–19</sup> has led to numerous efforts at rater training, including frame-of-reference training,<sup>20</sup> to negotiate a common perspective. Interestingly, the relative lack of success for such training efforts has led some to simply exclude “inherently inconsistent” raters to ensure a common perspective and a perceived reliability among the remaining raters.<sup>16</sup> It is hard to argue that this approach excludes subjectivity—at best, it masks subjectivity behind a constructed consensus.

Not only do raters cause problems for the notion of objectivity, so does the context. Context specificity (i.e., the observation that an individual’s performance on a particular problem or in a particular situation is only weakly predictive of the same individual’s performance on a different problem or in a different situation<sup>21</sup>) is a commonly recognized thorn in the side of psychometricians trying to assess performance. Indeed, Norcini has been credited with suggesting that context specificity is “the one fact of medical education.”<sup>22(p1220)</sup> Given the widespread prevalence of this “fact,” perhaps it is time to suggest that competence does not reside in the individual but, rather, in the individual’s interaction with a highly variable context.<sup>23,24</sup> Further, Gingerich<sup>25</sup> has suggested that the judgment of clinical competence is an inherently social activity and that social judgments are necessarily interpretations of the performance. If so, then a perceiver’s (rater’s) interpretation of an individual’s performance *is a part of the context*, and multiple perceivers means multiple contexts. For example, one perceiver may experience a performance as reassuring (confident and knowledgeable), and another may see it as off-putting (controlling and arrogant), but each experience of the performance is “true” for that perceiver. Thus, the most appropriate conclusion to draw from variations in assessment for a given performance is not that there is noise in the ratings and a problematic lack of objectivity but, rather, that the performance can be perceived in importantly different ways, so *there is no single “objective” truth about the performance (much less, the performer)*. This is a constructivist view, rather than a positivist one, which, as mentioned above, would suggest that there is only one truth.

Assessment in the clinical setting complicates the notion of “objectivity” still further. In the clinical context, assessment of trainees implies an evaluation of their readiness to be entrusted with care,<sup>26</sup> and therefore, the assessment of learners and decisions around patient care are inextricable.<sup>27–29</sup> As medical trainees work under the supervision of a practitioner, the evaluation of their readiness to engage in patient care involves a continuous balancing of the benefits and risks for both the learner and patient.<sup>30,31</sup> That is, when a medical specialist who is charged with the care of a patient is simultaneously evaluating a trainee’s capacity to participate in the care of that patient, the interest of providing the best care will, and likely should, affect the judgment. In this regard, an intimate and personal acquaintance with the learner is needed to develop the confidence needed to make meaningful entrustment decisions.<sup>32,33</sup> These moment-by-moment, ad hoc entrustment decisions must, by definition, be subjective and situation specific. Applying a criterion of “objectivity” to clinical assessment also carries with it the assumption that judgments can always be expressed as documentation that can be shared and understood acontextually. In practice, preceptors’ gut feelings or intuitions about trainees likely have an important role in guiding decisions about their readiness to practice alone.<sup>27</sup> Similar to expert judgments about patients, these intuitions might be shared meaningfully among other experts with similar experiences but are likely to lose some of their essence when formalized in documented words or numbers.<sup>34,35</sup> Thus, to arrive at a summative decision about a given individual, either to guide further training or to determine readiness for certification,<sup>36</sup> it is necessary for a team, such as a clinical competency committee, to examine the breadth of these subjective assessments over some time period and to negotiate, in light of the complex patterns of data and informed by their own personal knowledge and experience, until the team feels comfortable in making a coherent collective (rather than “objective”) determination.<sup>37,38</sup>

## The Power of Embracing Subjectivity

To this point, we have argued that the pursuit of objectivity is problematic because of the “single truth” that it implies. In this section, we will suggest, consistent with other recent authors,<sup>39,40</sup> not merely that subjectivity cannot be avoided but that, in fact, it should be embraced. We previously pointed out that there might be multiple legitimate perspectives on a single performance and that each of these perspectives might be “true” in the experience of the individual perceiver. If so, the effort to average those perspectives to find the “signal in the noise” or to try to negotiate a single common perspective among perceivers is problematic not only in its representation of the individual but also in its preparation of the individual for effective future performance. The popularity and widespread acceptance of multisource feedback (MSF) as a legitimate approach to building a valid image of a trainee exemplifies our point.<sup>41,42</sup> It is *because* of the differences between assessors, not despite them, that MSF is so useful.

Adaptability to the context is a particularly important feature of a skillful practitioner, and as previously suggested, assessors are part of that context. What the community (patients, health professionals, hospitals, etc.) would like to see in a high-quality practitioner is the ability (and propensity) to monitor his or her impact on other individuals in an interaction and, when needed, to modify his or her behaviors in ways that accommodate the feedback received about his or her actions as experienced by each person. If others perceive one’s actions or approach positively, then continuing in this way is appropriate. However, if it becomes apparent that others are finding one’s actions or approach off-putting, then it becomes necessary to adapt appropriately. To effectively monitor and accommodate in this way, it is critical for the actor to *know* that some people find his or her style off-putting (or see his or her actions or approach as arrogant) so that he or she can be alert to this concern and adapt if and when this sort of reaction is being perceived. Thus, in contrast with an assessment process that suggests that there is just one best way to act in a particular situation, a more appropriate message to relay to the trainee might be the various ways in which his

or her behavior was interpreted so that he or she can be alert to these sorts of interpretations and respond accordingly, in a situationally appropriate way. The fact that learners often react to “inconsistent” feedback with frustration, therefore, might be an important signal that they believe there is a single objectively correct way to act. This frustration suggests that these learners are currently not well prepared for the variability in interpretations of their behavior that they will face in clinical practice.

Interestingly, embracing subjectivity not only offers the possibility of richer feedback by defensibly representing differing perspectives on performance across assessments, but it also enables better thinking about the value and defensibility of the moment-by-moment judgments being made by preceptors to enable ad hoc entrustment. Using Crossley and colleagues<sup>43</sup> notion of construct-aligned scales, assessment is shifting away from statements about the individual being assessed and focusing instead on the level of participation that the preceptor feels comfortable allowing for a certain learner at a certain moment.<sup>44–46</sup> Entrusting learners with clinical tasks implies an assessment of perceived risk, as the anticipated level to which the learner will be able to perform the task is weighed against the patient’s safety in that particular context.<sup>29–31,47</sup> Importantly, this shift in focus empowers the preceptor to probe and document his or her subjective experience rather than forcing him or her to document a context-free inference about the learner in the guise of objectivity. Ironically, therefore, the move to subjectivity as a framing of assessment in the workplace places the preceptor in a substantially more defensible position with regard to his or her documentation. A learner might legitimately question the “objective truth” in statements such as “below average” or “meets expectations” or challenge the fairness of differences in “objective scores” given to him or her as compared with different learners. However, it is difficult for a learner to challenge a statement such as “I am just not comfortable with you performing this procedure,” “I’ll not have you lead that patient conversation on your own yet,” or “I’m now comfortable leaving the operating room while you complete this part of the procedure.”<sup>46</sup> In other words, at the level of an individual assessment

of a single performance, documentation of the preceptor’s *subjective* experience is the only truly defensible proposition. “Objective truth” statements are always open to being questioned. Even descriptions of expected behavior at different developmental stages<sup>48,49</sup> can, at best, be a suggested reference for raters; they can never serve as “objective” milestones.<sup>50</sup>

## Implications and Future Directions

Acknowledging and celebrating the reemergence of subjectivity in assessment, Hodges<sup>40</sup> has described health professions education as moving into a “post-psychometric era.” We agree that there is a growing and appropriate challenge of the psychometric premise of objectivity and its attendant construction of variability in assessment as simply noise masking a single “true” signal about an individual or performance. However, we wish to strongly suggest that this should *not* lead to the return of a “pre-psychometric” mind-set about data and assessment. We cannot ignore, and, in fact, must build on, the lessons of the psychometric era. It is important to remember that the psychometric pursuit of objectivity included, in no small part, an effort to achieve fairness in assessment. Lessons from the past repeatedly demonstrate that unfettered subjectivity can easily lead to the (implicit or explicit) systematic disadvantaging and even outright exclusion of individuals from different social groups. However, as noted by Gould,<sup>51</sup> lessons from the past also suggest that the development of “objective” measures has not infrequently produced similar results. Seeking fairness in assessment remains an important goal. But learners should realize that fairness results from the interaction of ability (observed behavior) with context (including the expert rater and the circumstances), making comparisons among learners challenging and inherently less transparent (“I saw you [learner A] doing very well with an easy case” versus “I saw you [learner B] struggling with a difficult case” could lead to a similar rating, but could make learner A feel that he or she was being treated unfairly). Nonetheless, it is important to remember that not all data are created equal, and as educators move toward embracing subjectivity in assessment, they must develop new

ways to determine the legitimacy and meaningfulness of the judgments supervising clinicians make, whether for summative or formative purposes. Clinicians, as content experts, should receive support and guidance on how to explain entrustment decisions to learners.

There are already some intriguing explorations in this direction for the assessment of individual pieces of written work (as reviewed and elaborated on by Kuper<sup>52</sup>). In considering how to translate these ideas into the clinical realm, one promising direction is to deeply explore what makes professionals trust their colleagues as practitioners.<sup>53</sup> Expert judgment, although fraught with subjectivity, is unavoidable, but its quality increases with experience. As Hodges<sup>40(p37)</sup> has argued, clinical assessment of trainees might best be likened to clinical judgment: “With experience, expert clinicians become more rapid and more accurate in their recognition of patterns. There is no reason to believe that this process does not also operate in education.” Yet, this process need not be treated as a “black box.” We are hopeful that by unpacking the process by which ad hoc entrustment decisions are being made, researchers can develop opportunities to inform and shape these practices in positive ways.

Increasingly, it is being recognized that uniqueness among individual practitioners is not something to be avoided. In fact, recent models of patient safety have suggested that “everyday performance variability provides the adaptations that are needed to respond to varying conditions, and hence is the reason why things go right. Humans are consequently seen as a resource necessary for system flexibility and resilience.”<sup>54(p4)</sup> This suggests that much of the context-dependent ability that comprises expert clinical performance may legitimately vary among practitioners and requires subjective judgment. Educators must explore how to compile subjective data to compare across people or against some standard for the purposes of high-stakes decision making. This will require enough data to be able to discern patterns and interpret individual data points in context. This does not necessarily mean discarding outliers or averaging such that varying opinions are lost in the summative representation of the individual (as happens with the

use of “central tendency” statistics) but, rather, interpreting the variability of data according to each data point’s context and giving the data their weight based on their importance rather than on their consistency.

*Funding/Support:* None reported.

*Other disclosures:* None reported.

*Ethical approval:* Reported as not applicable.

**O. ten Cate** is professor of medical education and senior scientist, Center for Research and Development of Education, University Medical Center Utrecht, Utrecht, the Netherlands; ORCID: <https://orcid.org/0000-0002-6379-8780>.

**G. Regehr** is professor, Department of Surgery, and associate director of research, Centre for Health Education Scholarship, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada; ORCID: <http://orcid.org/0000-0002-3144-331X>.

## References

- 1 Van der Vleuten CP, Norman GR, De Graaff E. Pitfalls in the pursuit of objectivity: Issues of reliability. *Med Educ.* 1991;25:110–118.
- 2 De Groot AD. Methodology [in Dutch; reprinted in 1994 and 2008]. Assen, the Netherlands: Van Gorcum; 1961.
- 3 McGuire C, Babnott D. Simulation technique in the measurement of problem solving skills. *J Educ Meas.* 1967;4:1–10.
- 4 Powles A, Wintrup N, Neufeld V, Wakefield J, Coates G, Burrows J. The triple jump exercise: Further studies of an evaluative technique. In: Proceedings of the 20th Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges; 1981:74–79.
- 5 Van der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract.* 1996;1: 41–67.
- 6 Reznick RK, Blackmore D, Dauphinée WD, Rothman AI, Smees S. Large-scale high-stakes testing with an OSCE: Report from the Medical Council of Canada. *Acad Med.* 1996;71(1 suppl):S19–S21.
- 7 Tamblyn R, Abrahamowicz M, Dauphinée D, et al. Physician scores on a national clinical skills examination as predictors of complaints to medical regulatory authorities. *JAMA.* 2007;298:993–1001.
- 8 Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE guide no. 31. *Med Teach.* 2007;29:855–871.
- 9 Gruppen LD, Ten Cate O, Lingard LA, Teunissen PW, Kogan JR. Enhanced requirements for assessment in a competency-based, time-variable medical education system. *Acad Med.* 2018;93(3 suppl):S17–S21.
- 10 Charlin B, Roy L, Brailovsky C, Goulet F, van der Vleuten C. The script concordance test: A tool to assess the reflective clinician. *Teach Learn Med.* 2000;12:189–195.
- 11 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of

- script concordance test scores. *Med Educ.* 2013;47:1175–1183.
- 12 Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students’ clinical skills and behaviors in medical school. *Acad Med.* 1999;74:842–849.
- 13 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med.* 2003;14:37–41.
- 14 Cacamese SM, Elnicki M, Speer AJ. Grade inflation and the internal medicine subinternship: A national survey of clerkship directors. *Teach Learn Med.* 2004;19:343–346.
- 15 Albanese MA. Challenges in using rater judgements in medical education. *J Eval Clin Pract.* 2000;6:305–319.
- 16 Newble DI, Hoare J, Sheldrake PF. The selection and training of examiners for clinical examinations. *Med Educ.* 1980;14:345–349.
- 17 Elliot DL, Hickam DH. Evaluation of physical examination skills. Reliability of faculty observers and patient instructors. *JAMA.* 1987;258:3405–3408.
- 18 Noel GL, Herbers JE Jr, Caplow MP, Cooper GS, Pangaro LN, Harvey J. How well do internal medicine faculty members evaluate the clinical skills of residents? *Ann Intern Med.* 1992;117:757–765.
- 19 Clauser B, Subhiyah R, Nungester R, Ripkey D, Clyman S, McKinley D. Scoring a performance-based assessment by modeling the judgments of experts. *J Educ Meas.* 1995;32:397–415.
- 20 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. *J Gen Intern Med.* 2009;24:74–79.
- 21 Eva KW. On the generality of specificity. *Med Educ.* 2003;37:587–588.
- 22 Colliver JA. Educational theory and medical education practice: A cautionary note for medical school faculty. *Acad Med.* 2002;77:1217–1220.
- 23 ten Cate O, Snell L, Carraccio C. Medical competence: The interplay between individual ability and the health care environment. *Med Teach.* 2010;32:669–675.
- 24 Regehr G. It’s NOT rocket science: Rethinking our metaphors for research in health professions education. *Med Educ.* 2010;44:31–39.
- 25 Gingerich A. What if the “trust” in entrustable were a social judgement? *Med Educ.* 2015;49:750–752.
- 26 Ten Cate O. Entrustment as assessment: Recognizing the ability, the right, and the duty to act. *J Grad Med Educ.* 2016;8:261–262.
- 27 ten Cate O. Trust, competence, and the supervisor’s role in postgraduate training. *BMJ.* 2006;333:748–751.
- 28 Kogan JR, Conforti LN, Iobst WF, Holmboe ES. Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Acad Med.* 2014;89:721–727.
- 29 Ten Cate O. Entrustment decisions: Bringing the patient into the assessment equation. *Acad Med.* 2017;92:736–738.
- 30 Ten Cate O. Managing risks and benefits: Key issues in entrustment decisions. *Med Educ.* 2017;51:879–881.

- 31 Damodaran A. Trust and risk: A model for medical education. *Med Educ.* 2017;51:892–902.
- 32 Hirsh DA, Holmboe ES, ten Cate O. Time to trust: Longitudinal integrated clerkships and entrustable professional activities. *Acad Med.* 2014;89:201–204.
- 33 Boscardin CK, Wijnen-Meijer M, Cate OT. Taking rater exposure to trainees into account when explaining rater variability. *J Grad Med Educ.* 2016;8:726–730.
- 34 Billett SR. Securing intersubjectivity through interprofessional workplace learning experiences. *J Interprof Care.* 2014;28:206–211.
- 35 Gigerenzer G. *Gut Feelings: The Intelligence of the Unconscious.* New York, NY: Penguin Group; 2007.
- 36 Ten Cate O, Hart D, Ankel F, et al; International Competency-Based Medical Education Collaborators. Entrustment decision making in clinical training. *Acad Med.* 2016;91:191–198.
- 37 Hauer KE, Cate OT, Boscardin CK, et al. Ensuring resident competence: A narrative review of the literature on group decision making to inform the work of clinical competency committees. *J Grad Med Educ.* 2016;8:156–164.
- 38 Hauer KE, Chesluk B, Iobst W, et al. Reviewing residents' competence: A qualitative study of the role of clinical competency committees in performance assessment. *Acad Med.* 2015;90:1084–1092.
- 39 van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach.* 2012;34:205–214.
- 40 Hodges B. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach.* 2013;35:564–568.
- 41 Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof.* 2003;23:4–12.
- 42 Alofs L, Huiskes J, Heineman MJ, et al. User reception of a simple online multisource feedback tool for residents. *Perspect Med Educ.* 2015;4:57–65.
- 43 Crossley J, Johnson G, Booth J, Wade W. Good questions, good answers: Construct alignment improves the performance of workplace-based assessment scales. *Med Educ.* 2011;45:560–569.
- 44 Weller JM, Misur M, Nicolson S, et al. Can I leave the theatre? A key to more reliable workplace-based assessment. *Br J Anaesth.* 2014;112:1083–1091.
- 45 Mink RB, Schwartz A, Herman BE, et al; and the Steering Committee of the Subspecialty Pediatrics Investigator Network (SPIN). Validity of level of supervision scales for assessing pediatric fellows on the common pediatric subspecialty entrustable professional activities. *Acad Med.* 2018;93:283–291.
- 46 Weller JM, Castanelli DJ, Chen Y, Jolly B. Making robust assessments of specialist trainees' workplace performance. *Br J Anaesth.* 2017;118:207–214.
- 47 Holzhausen Y, Maaz A, Cianciolo AT, Ten Cate O, Peters H. Applying occupational and organizational psychology theory to entrustment decision-making about trainees in health care: A conceptual model. *Perspect Med Educ.* 2017;6:119–126.
- 48 Frank J, Snell L, Sherbino J, eds. *CanMEDS 2015 Physician Competency Framework.* Ottawa, Ontario, Canada: Royal College of Physicians and Surgeons of Canada; 2015. <http://www.royalcollege.ca/rcsite/documents/canmeds/canmeds-full-framework-e.pdf>. Accessed October 2, 2018.
- 49 Lowry BN, Vansaghi LM, Rigler SK, Stites SW. Applying the milestones in an internal medicine residency program curriculum: A foundation for outcomes-based learner assessment under the next accreditation system. *Acad Med.* 2013;88:1665–1669.
- 50 Hawkins RE, Welcher CM, Holmboe ES, et al. Implementation of competency-based medical education: Are we addressing the concerns and challenges? *Med Educ.* 2015;49:1086–1102.
- 51 Gould S. *The Mismeasure of Man.* New York, NY: Norton; 1981.
- 52 Kuper A. Literature and medicine: A problem of assessment. *Acad Med.* 2006;81:128–137.
- 53 Gingerich A, Daniels V, Farrell L, Olsen SR, Kennedy T, Hatala R. Beyond hands-on and hands-off: Supervisory approaches and entrustment on the inpatient ward. *Med Educ.* 2018;52:1028–1040.
- 54 Hollnagel E, Wears R, Braithwaite J. *From safety-I to safety-II: A white paper.* <https://www.england.nhs.uk/signuptosafety/wp-content/uploads/sites/16/2015/10/safety-1-safety-2-white-papr.pdf>. Published 2015. Accessed September 20, 2018.