Entrustability Scales: Outlining Their Usefulness for Competency-Based Clinical Assessment

Janelle Rekman, MD, Wade Gofton, MD, MEd, Nancy Dudek, MD, MEd, Tyson Gofton, PhD, and Stanley J. Hamstra, PhD

Abstract

Meaningful residency education occurs at the bedside, along with opportunities for situated in-training assessment. A necessary component of workplace-based assessment (WBA) is the clinical supervisor, whose subjective judgments of residents' performance can yield rich and nuanced ratings but may also occasionally reflect bias. How to improve the validity of WBA instruments while simultaneously capturing meaningful subjective judgment is currently not clear. This Perspective outlines how "entrustability scales" may help bridge the gap between the assessment judgments of clinical supervisors and WBA instruments. Entrustment-based assessment evaluates trainees against what they will actually do when independent; thus, "entrustability scales"—defined as behaviorally anchored ordinal scales based on progression to competence—reflect a judgment that has clinical meaning for assessors. Rather than asking raters to assess trainees against abstract scales, entrustability scales provide raters with an assessment measure structured around the way evaluators already make day-to-day clinical entrustment decisions, which results in increased reliability. Entrustability scales help raters make assessments based on narrative descriptors that reflect real-world judgments, drawing attention to a trainee's readiness for independent practice rather than his/her deficiencies. These scales fit into milestone measurement both by allowing an individual resident to strive for independence in entrustable professional activities across the entire training period and by allowing residency directors to identify residents experiencing difficulty. Some WBA tools that have begun to use variations of entrustability scales show potential for allowing raters to produce valid judgments. This type of anchor scale should be brought into wider circulation.

Meaningful residency education occurs at the patient bedside, along with opportunities for situated in-training assessment. A necessary component of workplace-based assessment (WBA) is the clinical supervisor, whose subjective judgments of residents' performance can yield rich and nuanced ratings but may also on occasion reflect bias.[1,2] These biases often result in WBA tools demonstrating low reliability.[3–5] Various medical educators have proposed solutions—prioritization of objective tools,[6] standardization,[7] and analytic rating scales[8]—all of which have been largely ineffective. Interestingly, recent work suggests that raters do not so much disagree on trainee performance but, rather, on how to interpret the assessment scale or response format (e.g., how to "choose a number").[9] How to improve rater agreement in WBA and meaningfully capture subjective judgment is not currently clear. "Entrustability scales," which we define as behaviorally anchored ordinal scales based on progression to competence, reflect a judgment that has clinical meaning for assessors. These scales are unlikely to solve all of the problems associated with rater agreement with WBA; however, we believe that they have demonstrated the potential to be useful for clinical educators, such as ourselves, who must rate trainees within a competency-based assessment program.

As competency-based medical education (CBME) becomes more prevalent, the need to develop and deploy clinical assessment tools that reflect workplace reality becomes more and more critical.[10,11] Educators and evaluators have used variations of entrustability scales, sometimes referred to as "independence rating scales," in a variety of contexts and interdisciplinary specialties[12] for several years. Such scales are now appearing in many medical education workplaces, from the Australian operating room,[13] to the British acute care ward,[9] to the decisions based on progression through milestones, as promoted by the Accreditation Council for Graduate Medical Education (ACGME).[14]

The goal of this Perspective is to outline how entrustability scales may help bridge the gap between the assessment judgments of clinical supervisors and the WBA instruments available to them.

Entrustment

Learning to perform clinical tasks competently is a process all medical learners must navigate on their road to professional independence. Clinical preceptors are accountable for helping residents shoulder increasing responsibility and should continually be asking themselves if the resident is capable of completing a task independently.[15,16] Building trust and making entrustment decisions are complicated social interactions influenced by many competing factors in the workplace.[17] ten Cate[18] developed the idea of "entrustable professional activities" (EPAs) to make explicit the everyday judgments supervisors make regarding whether to trust a given

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

Correspondence should be addressed to Janelle Rekman, Division of General Surgery, Ottawa Hospital–Civic Campus, Room WM150b, Loeb Research Building–Main Floor, 725 Parkdale Ave., Ottawa, Ontario, Canada K1Y 4E9; telephone: (613) 798-5555 ext.10606; e-mail: jrekman@toh.on.ca.

Acad Med. 2016;91:186–190.
First published online December 1, 2015
doi: 10.1097/ACM.0000000000001045

trainee with a specific task. EPAs are defined as essential responsibilities of the specialty that can be left, or *entrusted*, to a trainee. ten Cate designed EPAs to link competencies, such as those outlined by the CanMEDS roles and the ACGME domains,[14] to clinical practice.

Determining whether to entrust individual residents entails evaluating them against what they will actually do when practicing independently, or putting their abstract knowledge and generalized skills into a larger context.[19] Clinical instructors can (and do) make entrustment decisions at all milestones levels, not just in the final stages of training. To illustrate, we have listed the milestone levels for the Patient Care (PC2) competency within the practice domain of "Care for Diseases and Conditions" of the ACGME milestones for general surgery. The narrative scale progresses through four levels, beginning with "This resident recognizes and manages common post-operative problems … with the assistance of senior residents or staff members who are physically present."[14] The scale then progresses through "This resident recognizes and manages common post-operative problems … with the assistance of senior residents or staff members who are available for consultation, but not physically present" and "This resident recognizes and manages complex postoperative problems … independently."[14] Finally, the Level 4 competency milestone reads, "This resident can lead a team and provide supervision in the evaluation and management of complex post-operative problems."[14] The PC2 milestones provide ongoing opportunities to assess learners during their training; however, a final, decisive "ready for independent practice" entrustment judgment is necessary for each learner to complete his or her residency training,[20,21] and entrustability scales may help with such judgments.

Entrustability Scales: A Species of Construct-Aligned Scales

It is crucial for frontline educators to feel an assessment tool captures their true appraisal of a resident. Crossley and Jolly[22] have suggested that effective assessment tools have *construct alignment*, which means that the tool reflects the expertise and priorities of the evaluator. In a recent review of in-training assessment, van der Vleuten and Verhoeven[5] note that the value of assessment instruments depends more on the users (raters) than on the instruments themselves. Rather than asking raters to make assessments against abstract scales, such as skill level according to postgraduate year or ranking within cohort, construct-aligned scales provide evaluators with a standardized assessment measure that is structured around the way they already make day-to-day decisions.

ten Cate and Scheele[19] assert that the construct of medical education assessment, agreed on by clinician–raters across specialties, is competency progression or *entrustment*. Accordingly, a tool aligned to the construct of competency progression would incorporate an expert's perception of the trainee's increasing clinical ability.[23] Entrustability scales are a species of construct-aligned anchor scales because they align with the expertise and priorities of clinician–educators. The Zwisch scale,[24] a behaviorally anchored ordinal (1–4) scale used to grade the degree of guidance necessary during a technical procedure, is a great simple example. At the lowest end of the scale is "show and tell," a stage at which each step is outlined for the trainee, and at the top, "supervision only," at which the supervisor's presence is warranted only to ensure patient safety. In between these two levels, a trainee progresses through a stage of active help and a stage of passive help. Another example of an entrustability scale that assesses a trainee's ability to complete named surgical procedures is the O-SCORE (Table 1).[25]

An entrustability scale may seem to apply more naturally to procedural skills; however, examples of the effectiveness of entrustability scales are also available for assessing nontechnical skills and more complex tasks such as "managing an interdisciplinary team" and "taking a detailed history."[26] These results demonstrate the potential of such scales for addressing a wide range of clinical performance.

Benefits of Entrustability Scales

Raters find increased meaning in their assessment decisions due to construct alignment

The cornerstone of entrustability scales is that they are deliberately aligned with day-to-day assessments of competency and independence in the setting of clinical education.[22] Rather than requiring an attending physician to translate his or her assessment into an ordinal category such as 4 for "above average," or 2 for "below average," entrustability scales acknowledge the categorical judgment—to entrust or not to entrust—that raters have used successfully in the workplace.[2] Rating scale error is partially due to a rater failing to correctly translate implicit categorical (interval) judgments into the ordinal judgments traditionally required by abstract scales.[27] By reverse engineering descriptors to fit raters' existing categorical schemas, entrustability scales can increase assessment reliability.[22] For example, when investigators asked anesthesia attending physicians to judge whether a trainee required direct, indirect, or distant supervision with a case, the physicians were much more reliable in their assessments than when they were asked to judge what is expected of a trainee at different stages of training (only 9 assessments were required versus 50 to reach agreement).[13]

Moreover, entrustability scales reflect a judgment that already has meaning for evaluators in the context of clinical education.[28] To balance trainee learning and patient safety, clinician–educators must decide when trainees may act without supervision. Because entrustability scales are based on this real-world judgment—not an abstract model of ideal trainee performance—actual day-to-day decisions to entrust underlie the assessment. That is, this real-world, practical judgment makes the assessment meaningful. For example, anesthesiologists are accustomed to thinking, "Can I leave my resident alone to do this task?" Formalizing this judgment on an assessment form is less onerous for and more meaningful to both the rater and the trainee.[13] In another study, when investigators compared an evaluation of residents' technical skills using procedure-based assessment (PBA, a type of entrustability scale) versus an evaluation using the OSATS (Objective Structured Assessment of Technical Skills), they found that the former, using the PBA, was much more reliable.[29]

Behavioral-based rating scales have a long tradition,[30] and their narrative wording seems to be easier for raters to interpret

Table 1

The Ottawa Surgical Competency Operating Room (O-SCORE) Scale[a]: An Entrustability-Aligned Anchor Scale

Level	Descriptor
1	"I had to do" (i.e., requires complete hands on guidance, did not do, or was not given the opportunity to do)
2	"I had to talk them through" (i.e., able to perform tasks but requires constant direction)
3	"I had to prompt them from time to time" (i.e., demonstrates some independence, but requires intermittent direction)
4	"I needed to be there in the room just in case" (i.e., independence but unaware of risks and still requires supervision for safe practice)
5	"I did not need to be there" (i.e., complete independence, understands risks and performs safely, practice ready)

[a]The authors adapted the scale from Gofton W, Dudek N, Wood T, Balaa F, Hamstra S. The Ottawa surgical competency operating room evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–407.

because they offer a ready-made rich description of the construct, compared with scales that include only numbers and just one or two words (e.g., "average"). Crossley and colleagues[9] demonstrated this in a study comparing assessment tools in which they changed only the anchors, not the actual scale. The version of the scale with narrative, construct-aligned anchors showed greater reliability, suggesting that the poor reliability of WBA tools may not, as traditionally assumed, be due to differences in rater assessment but, instead, to different interpretations of poorly aligned scales.

Changing the culture of giving and receiving assessment

Resident physicians are professional students, and to get where they are, they likely have a history of scoring well on written and other tests; however, to advance toward competency in residency, these adult learners must also receive constructive critique on the areas in which they must improve. Frequently, however, medical assessment results are skewed toward the top of the scale, and most residents receive "above average" scores.[31] Medical educators have proposed many reasons for this phenomenon including the lack of remediation options and the desire of attending physicians to preserve a positive working relationship. Physician raters may also be hesitant to assign low scores if these require more justification or could lead to legal action.[31]

On a practical level, clinical raters are unlikely to tell their residents that their performance was "unsatisfactory" or even "below the expected level"—and even more unlikely to fail them.[13] However, the reference standard of an entrustability anchor scale is a workplace-based decision by the evaluator about when a trainee may safely perform independently. Entrustability scales naturally focus feedback on a trainee's readiness for independent practice rather than on a trainee's deficiencies or his or her ranking with respect to peers.[25] Basing evaluations on the external reference of safe independent practice overcomes two of the most common weaknesses inherent in WBA scales—central tendency and leniency bias[1]—and creates freedom for the assessor to use all categories/numbers on the scale.[9]

Entrustability scales also help clinical raters make nonpejorative assessments based on narrative descriptions that reflect real-world judgments. This not only increases the likelihood of an honest assessment (i.e., it allows for a more valid assessment of the construct of interest) but also helps trainees interpret the assessment as a representation of their progress toward safe independent practice, rather than as a comparison with their peers or with an abstract construct, such as their year of training. Investigators studying the O-SCORE tool found that residents were comfortable receiving lower scores when the scale was worded in terms of entrustment.[25] Perhaps residents are more comfortable hearing assessments such as "I don't think you are ready to do this lumbar puncture yet. First look up the anatomy and let me show you" (a score of 1 on an entrustability anchor scale) or "You did a good job of explaining that the patient has a wound infection and needs antibiotics. But I needed to remind you to explain the side effects and to tell the patient to call the nurse if the infection appears to be worsening" (a score of 3)—rather than, for example, "You scored below average" or even, simply, "Your performance was average." Importantly, entrustability scales also focus attention on the practical goal of independent practice rather than the personal or psychological goals of approval and competitive excellence. Given CBME's reliance on formative feedback to guide residents from milestone to milestone,[32] fostering a culture of accepting low-stakes daily assessments is essential.[33]

Entrustment and milestone progression

Entrustability scales fit into milestone measurement by providing a consistent measure across the entire training cycle; they eliminate the moving goalposts of comparison to peer group. Further, entrustability scales allow individual residents to strive for independence throughout their training period and, over time, to achieve greater degrees of entrustment.[19] Each resident starts at 1 (beginner, requiring consistent attending physician modeling) and strives toward the goal of 5 (competent, autonomous) regardless of his or her peers' performance.

That said, a key aspect of a competency-based assessment program should be the timely identification of residents who are progressing slowly. Entrustability scales streamline this process by providing a way to track learners over time and assess stages during which they are "falling off the curve."[13] Promising research, performed among social work trainees, has shown that assessment forms using narrative descriptions of performance (similar to entrustability-anchored descriptions) may be better at identifying borderline performance than traditional forms.[34] Even residents who progress normally in the entrustability growth curve will show variability over time that reflects contextual differences. Compared with cohort measurement, which would measure only changes in standing, this achievement model unambiguously reflects individual learning. Given that the goal of an assessment program is to obtain the best assessment of the construct of interest by aligning the language of assessment with the actual performance of residents (not their

rankings), narratively based entrustability scales show potential.

Rater training and entrustability scales

Whenever a new WBA tool is implemented, concerns over rater bias and subjectivity should prompt discussions about how much rater training is necessary.[1] Indeed, evidence shows that rater training can increase proper use of an assessment tool[35]; however, these training sessions are resource intensive, time consuming (often requiring two to four hours of a rater's time), and difficult to implement outside of a study environment.[36] For this reason, their potential long-term feasibility is questionable. Crossley and Jolly's[22] work on construct-aligned scales—along with the success of other entrustability scales in helping raters understand what they are being asked to rate—suggests that traditional rater training programs may no longer be necessary. Nonetheless, when education leaders use this type of scale, especially the first time, they should monitor for raters who fail to use the external reference as intended (i.e., who revert to comparing within peer group instead of evaluating for independent competence). A method for creating targeted rater reorientation, called frame-of-reference training, could be attempted by monitoring regularly for outlying raters likely to benefit from realignment to scale principles. Pugh and colleagues[37] applied one such method to train raters who were assessing a procedure-skills-based objective structured clinical exam.

Context Complexity and Entrustability Scales

A limitation that entrustment-aligned tools share with all WBA tools is their inability to completely account for context complexity.[13] Although assessments are specific to a particular resident performing a specific task, completing a lumbar puncture on a slim patient is different from completing one on a patient with a larger body habitus—even though both procedures are covered by the same EPA. Indeed, because many contextual factors influence entrustment decisions,[17] we recommend multiple assessments over time—ideally completed by several raters—to limit the impact of contextual variations on overall trainee assessment. Additionally, tools using entrustability scales benefit from a space for raters to leave narrative comments.[38] These narrative comments support learning by giving the resident detailed explanations and contextual examples of performance, and they help the individual(s) ultimately responsible for collating results of multiple WBAs to make more informed decisions.

Conclusions

Identifying feasible assessment tools, and confirming a rater's belief that a tool actually allows authentic assessment of a resident, often seem like competing interests.[39] Entrustability scales can make formative feedback more meaningful for raters and trainees alike[13] while also increasing the reliability of assessments.[9] Frontline clinicians do not want their judgments of residents' abilities to get "lost in translation,"[40] so a tool that helps them avoid this problem would add value to the clinical learning environment. Likewise, a tool that helps residents focus on an end goal (rather than on a grade) has the potential to increase the amount of well-constructed, actionable feedback they receive. Entrustability scales (or independence-aligned scales[9] or construct-aligned scales) show great potential for synchronizing actual clinical rater judgments with specific anchor scale measures in the competency-based environment. We believe that entrustability scales are valuable assessment tools that residency program directors and clinical instructors should adopt more widely.

Funding/Support: During the period of this project, J. Rekman received salary support through the University of Ottawa, Department of Surgery Surgeon Scientist Clinical Investigator Program. Otherwise, the authors report no external funding source for this study.

Other disclosures: None reported.

Ethical approval: Reported as not applicable.

J. Rekman is a general surgery resident and master's in health professions education student, University of Ottawa, Ottawa, Ontario, Canada.

W. Gofton is an orthopedic surgeon, University of Ottawa, Ottawa, Ontario, Canada.

N. Dudek is associate professor, Department of Medicine, University of Ottawa, Ottawa, Ontario, Canada.

T. Gofton is Wissenschaftlicher Mitarbeiter, Department of Philosophy, Eberhard Karls Universität, Tübingen, Germany.

S.J. Hamstra is vice president, Milestones Research and Evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois.

References

1 Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. Teach Learn Med. 2003;15:270–292.
2 Yeates P, O'Neill P, Mann K, Eva K. Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. Adv Health Sci Educ Theory Pract. 2013;18:325–341.
3 Carline JD, Wenrich M, Ramsey PG. Characteristics of ratings of physician competence by professional associates. Eval Health Prof. 1989;12:409–423.
4 Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998;73:1294–1298.
5 van der Vleuten C, Verhoeven B. In-training assessment developments in postgraduate education in Europe. ANZ J Surg. 2013;83:454–459.
6 Hodges B. Assessment in the post-psychometric era: Learning to love the subjective and collective. Med Teach. 2013;35:564–568.
7 Govaerts MJ, van der Vleuten CP, Schuwirth LW, Muijtjens AM. Broadening perspectives on clinical performance assessment: Rethinking the nature of in-training assessment. Adv Health Sci Educ Theory Pract. 2007;12:239–260.
8 Driessen E, Scheele F. What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. Med Teach. 2013;35:569–574.
9 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.
10 Embo M, Driessen E, Valcke M, van der Vleuten CP. Integrating learning assessment and supervision in a competency framework for clinical workplace education. Nurse Educ Today. 2015;35:341–346.
11 Bok HG, Teunissen PW, Favier RP, et al. Programmatic assessment of competency-based workplace learning: When theory meets practice. BMC Med Educ. 2013;13:123.
12 Durkin GJ. Development and implementation of an independence rating scale and evaluation process for nursing orientation of new graduates. J Nurses Staff Dev. 2010;26:64–72.
13 Weller JM, Jones A, Merry AF, Jolly B, Saunders D. Investigation of trainee and specialist reactions to the mini-clinical evaluation exercise in anaesthesia: Implications for implementation. Br J Anaesth. 2009;103:524–530.
14 Accreditation Council for Graduate Medical Education; American Board of Surgery. The general surgery milestone project July 2015. http://www.acgme.org/acgmeweb/portals/0/pdfs/milestones/surgerymilestones.pdf. Accessed October 22, 2015.
15 Mulder H, Ten Cate O, Daalder R, Berkvens J. Building a competency-based workplace curriculum around entrustable professional activities: The case of physician assistant training. Med Teach. 2010;32:e453–e459.

16 ten Cate O. Trust, competence, and the supervisor's role in postgraduate training. BMJ. 2006;333:748–751.
17 Hauer KE, Ten Cate O, Boscardin C, Irby DM, Iobst W, O'Sullivan PS. Understanding trust as an essential element of trainee supervision and learning in the workplace. Adv Health Sci Educ Theory Pract. 2014;19:435–456.
18 ten Cate O. Entrustability of professional activities and competency-based training. Med Educ. 2005;39:1176–1177.
19 ten Cate O, Scheele F. Competency-based postgraduate training: Can we bridge the gap between theory and clinical practice? Acad Med. 2007;82:542–547.
20 Cogbill TH, Swing SR. Development of the educational milestones for surgery. J Grad Med Educ. 2014;6(1 suppl 1):317–319.
21 Swing SR, Beeson MS, Carraccio C, et al. Educational milestone development in the first 7 specialties to enter the next accreditation system. J Grad Med Educ. 2013;5:98–106.
22 Crossley J, Jolly B. Making sense of work-based assessment: Ask the right questions, in the right way, about the right things, of the right people. Med Educ. 2012;46:28–37.
23 Wijnen-Meijer M, Van der Schaaf M, Booij E, et al. An argument-based approach to the validation of UHTRUST: Can we measure how recent graduates can be trusted with unfamiliar tasks? Adv Health Sci Educ Theory Pract. 2013;18:1009–1027.
24 George BC, Teitelbaum EN, Meyerson SL, et al. Reliability, validity, and feasibility of the Zwisch scale for the assessment of intraoperative performance. J Surg Educ. 2014;71:e90–e96.
25 Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa surgical competency operating room evaluation (O-SCORE): A tool to assess surgical competence. Acad Med. 2012;87:1401–1407.
26 Warm EJ, Mathis BR, Held JD, et al. Entrustment and mapping of observable practice activities for resident assessment. J Gen Intern Med. 2014;29:1177–1182.
27 Macrae CN, Bodenhausen GV. Social cognition: Thinking categorically about others. Annu Rev Psychol. 2000;51:93–120.
28 Kennedy TJ, Regehr G, Baker GR, Lingard L. Point-of-care assessment of medical trainee competence for independent clinical work. Acad Med. 2008;83(10 suppl):S89–S92.
29 Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: A prospective observational study of the methodology. Health Technol Assess. 2011;15:i–xxi, 1.
30 Schwab DP, Heneman HG, DeCotiis TA. Behaviorally anchored rating scales: A review of the literature. Pers Psychol. 1975;28:549–562.
31 Dudek NL, Marks MB, Regehr G. Failure to fail: The perspectives of clinical supervisors. Acad Med. 2005;80(10 suppl):S84–S87.
32 Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. Med Teach. 2010;32:676–682.
33 Schuwirth L, Ash J. Assessing tomorrow's learners: In competency-based education only a radically different holistic method of assessment will work. Six things we could forget. Med Teach. 2013;35:555–559.
34 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;43:327–344.
35 Roch SG, Woehr DJ, Mishra V, Kieszczynska U. Rater training revisited: An updated meta-analytic review of frame-of-reference training. J Occup Organ Psychol. 2012;85:370–395.
36 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.
37 Pugh D, Hamstra SJ, Wood TJ, et al. A procedural skills OSCE: Assessing technical and non-technical skills of internal medicine residents. Adv Health Sci Educ Theory Pract. 2015;20:85–100.
38 Driessen E, van der Vleuten C, Schuwirth L, van Tartwijk J, Vermunt J. The use of qualitative research criteria for portfolio assessment as an alternative to reliability evaluation: A case study. Med Educ. 2005;39:214–220.
39 Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. Acad Med. 2010;85:780–786.
40 Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the "black box" differently: Assessor cognition from three research perspectives. Med Educ. 2014;48:1055–1068.