Reconceptualizing Variable Rater Assessments as Both an Educational and Clinical Care Problem

Jennifer R. Kogan, MD, Lisa N. Conforti, MPH, William F. Iobst, MD, and Eric S. Holmboe, MD

Abstract

The public is calling for the U.S. health care and medical education system to be accountable for ensuring high-quality, safe, effective, patient-centered care. As medical education shifts to a competency-based training paradigm, clinician educators’ assessment of and feedback to trainees about their developing clinical skills becomes paramount. However, there is substantial variability in the accuracy, reliability, and validity of the assessments faculty make when they directly observe trainees with patients. These difficulties have been treated primarily as a rater cognition problem focusing on the inability of the assessor to make reliable and valid assessments of the trainee.

The authors’ purpose is to reconceptualize the rater cognition problem as both an educational and clinical care problem. The variable quality of faculty assessments is not just a psychometric predicament but also an issue that has implications for decisions regarding trainee supervision and the delivery of quality patient care. The authors suggest that the frame of reference for rating performance during workplace-based assessments be the ability to provide safe, effective, patient-centered care. The authors developed the Accountable Assessment for Quality Care and Supervision equation to remind faculty that supervision is a dynamic, complex process essential for patients to receive high-quality care. This fundamental shift in how assessment is conceptualized requires new models of faculty development and emphasizes the essential and irreplaceable importance of the clinician educator in trainee assessment.

We wrote this article to reconceptualize the rater cognition problem as both an educational and a clinical care problem.

We will propose that the variable quality of faculty assessments is more than a psychometric predicament but is also a problem that has implications for decisions regarding trainee supervision and quality of care. We will highlight how valid assessments of trainees (most often residents and fellows but sometimes medical students also) at the point of care are fundamental to achieving high-quality patient care. We will argue that if an assessment tool does not take into consideration the outcome of safe, effective, patient-centered care, the utility of that assessment is null and void. We will propose that the frame of reference and anchor of workplace-based assessments be the ability to provide safe, effective, patient-centered care. We will offer an equation, which we call the Accountable Assessment for Quality Care and Supervision (AAQCS) equation, to remind faculty that supervision is a dynamic, complex process essential for patients to receive high-quality care. We will recommend aligning rater training with programs that simultaneously allow clinical supervisors to practice the patient care and communication skills underpinning high-quality care. Finally, we will argue for the indispensable role of the clinician educator in trainee assessment.

Introduction

After the first 18 to 24 months of medical school, physicians’ training occurs primarily through clinical experiences in which trainees deliver patient care under the supervision of more experienced physicians. Trainees begin as novices and, through clinical encounters and structured, supervised educational experiences, practice, acquire, and ultimately apply competencies in the care of patients. The ultimate goal is for trainees to achieve the competence necessary to enter practice as unsupervised professionals. Clinical supervision provides monitoring, guidance, and feedback about skill development in the context of the trainee’s care of patients. In 2008, the Institute of Medicine (IOM) highlighted the importance of supervision and called for closer supervision of trainees. The intent was to reduce errors, lower patient mortality, and improve quality of care while enhancing residents’ education.

During training, learners ideally should move through graduated levels of supervision based on their ability to care for patients, and the intensity of supervision should be informed by valid workplace-based assessment. Workplace-based assessments are conducted in...
authentic situations (i.e., day-to-day practice) and evaluate multiple, essential competencies in an integrated, simultaneous fashion.7 Direct observation of real-life patient encounters (e.g., watching a resident with a patient) is one of the primary mechanisms by which to collect information about trainees’ skills.7,8 It also provides essential content and context for feedback about clinical skills.12

Workplace-based assessments are particularly important given the shift toward competency-based medical education in which core competencies are articulated and subsequently measured.2 A key impetus for competency-based education is the public’s calls for physician accountability, improved health care quality, and enhanced supervision and assessment of trainees.2,5,13 Foundational to competency-based education are multiple formative assessments of trainees, which inform evaluation and learning.1

Issues Associated With Workplace-Based Assessment

Despite their importance, workplace-based assessments are beset with validity concerns.3,4 A primary issue is high interrater reliability (i.e., faculty) variability of scores, leading to poor interrater reliability.4,14 Assessors themselves explain as much as 40% of this variability.16–18 Rating errors stem from leniency, stringency, halo, and central tendency effects, as well as anchoring and contrast bias.20–22 The implication of these errors and biases is that different raters rate the same encounter differently23,24 and make different pass/fail decisions.24

The etiology of high interrater variability is likely multifactorial. Multiple complex social and psychological processes are involved when faculty make judgments about a trainee.5,13 First, raters use different frames of reference (i.e., what they compare the trainee’s behaviors against), with themselves being the most common.22 For example, faculty benchmark residents’ performance to their own practice styles. Comparing a resident’s performance with that of residents at a similar training level is also common and becomes problematic if faculty are uncertain about what skills should be expected at a particular stage of training, or what constitutes competence.23,25 Assessment criteria often develop experientially over time, may be weighted differently by different assessors, and may uniquely focus on variable and different aspects of performance.21 Consequently, the aspects of performance that assessors regard as useful for determining quality are inconsistent between assessors.21,25

When observing trainees with patients, faculty also make many inferences (deriving what seem to be logical conclusions from premises assumed to be true).20,21 For example, imagine that a resident, while standing with crossed arms, delivers bad news to a patient. The faculty member may infer from this posture that the resident lacks empathy and humanism. However, alternative explanations are possible (e.g., the trainee has never delivered bad news and is uncomfortable). Faculty make inferences about residents’ funds of knowledge, competence, work ethics, prior experience, emotions, intentions, personalities, and even culture.20,25 Although making inferences is unavoidable and not necessarily bad, this becomes problematic when inferences are not recognized and validated.23 Therefore, inference in direct observation further influences the validity of assessment.

A third potential source of high interrater variability might relate to the clinical skills of faculty. Supervising physicians’ clinical skills may influence how they rate residents in the clinical setting.20,23,26 The clinical skills of faculty are variable and sometimes deficient.22–26 If quality medical care underpins clinical skills assessment, faculty must know and be able to accurately identify and assess evidence-based skills that characterize or promote such care. However, many of these skills, such as patient-centered communication, are minimally taught or emphasized during medical school and postgraduate training.31 Many faculty (save recent medical school graduates or those seeking additional training) have not received formal training in important communication skills such as delivering bad news, using empathy as a communication tool, and developing patient partnerships.32 Additionally, few opportunities exist for practicing physicians to learn and intentionally practice such skills.31

Historically, various approaches have been used to tackle the quandary of idiosyncratic ratings and high interrater variability. A common strategy has been to try to “fix the rating form.”23,24 A second strategy, rater training, has had minimal impact.35,26 A third strategy has used statistical methods (sampling across contexts and assessors) to ensure the reproducibility or generalizability of assessments.4 Practically, this entails observing the trainee in multiple clinical encounters over time using multiple different observers; a reliable, valid assessment of the trainee can be made with seven to nine assessments.27 It’s Not Just a Rating Problem

To date, high interrater variability has primarily been treated as a rater cognition or psychometric problem that affects trainee assessment.24,29 We submit that there is another stakeholder missing from the discussion of ratings: the patient.

A patient is not just a “variable” in a multivariable psychometric equation. A clinical supervisor must be able to accurately observe how a trainee provides care during a patient encounter to ensure that the patient receives high-quality care in that care episode. Accurate observation means identifying the presence or absence of appropriate clinical skills and important patient outcomes. For example, an attending observing a resident counseling a patient about starting a medication must correctly ascertain whether the elements of patient-centered, informed decision making are present and whether the patient receives the care he or she needs.38 If all this does not occur, the attending may have failed to appropriately supplement the resident’s clinical skills to ensure that the patient receives the needed care in that encounter. So while a composite judgment of a trainee across encounters and assessors helps to achieve “good” psychometric assessment criteria of the trainee, it fails to address the ability of a clinical supervisor to assess effective patient care in the context of a single clinical encounter or across encounters. Therefore, a rating does not simply represent how well (or not) the trainee did. Rather, in our example, the observations that informed the judgment and rating should also direct the attending’s decision making about the kind of care that the attending must contribute to the patient encounter. Observation, therefore, informs decisions about immediate and future supervision.
Clinical supervisors must balance trainee independence required for learning with adequate supervision for patient safety.\textsuperscript{39} Negotiating this balance requires point-of-care competence assessment—that is, assessment occurring at the time of and in the setting of clinical care.\textsuperscript{19} Faculty observation of, for example, a resident–patient interaction informs decisions about supervision—not just for that particular encounter but also for encounters going forward.\textsuperscript{39} An important trigger for more intensive oversight or supervision is the concern about a resident’s competence or ability to effectively handle a specific clinical situation.\textsuperscript{40} Therefore, if an attending inaccurately identifies how well or poorly the resident carries out patient care, the attending may not make high-quality decisions about the resident’s ability to give unsupervised care in the future.

Accurate assessment is, therefore, fundamental to high-quality decisions about supervision. Interrater variability is no longer just a rater cognition problem affecting the assessment of trainees. It risks being a clinical care problem for patients. Herein lies the conundrum. From a psychometric viewpoint, educators know that decisions about a trainee should not be based on a single observation, and many training programs have systems in place to obtain multiple evaluations per trainee. However, faculty must be able to use a single observation to determine what care they need to provide to a patient in a given encounter. A single observation may also inform decisions about the intensity of ongoing supervision (keeping in mind case and context specificity). Errors in these assessments run the risk that a patient, or future patients, will receive inadequate care.

In our current educational system, faculty cannot observe all patient encounters in their entirety. Faculty rely heavily on trainees’ self-assessments for when they need to ask for additional help.\textsuperscript{40} This implies that trainees can know the limits of their abilities or can self-monitor in action.\textsuperscript{44} Self-assessments can be informed by accurate, valid, external feedback.\textsuperscript{42} If faculty make low-quality observations of trainees, the subsequent feedback to the trainee will be of low utility. Consequently, trainees do not obtain the external feedback that could better guide their requests for future supervision. This further jeopardizes the quality of patient care that the trainee provides in the future. The end result is a continual dysfunctional cycle of poor education and poor clinical care.

### Solutions Going Forward

We propose three potential strategies to address this rating conundrum: (1) redefine workplace-based assessments’ frame of reference and rating scale anchors, (2) link the focus of in-training assessment with the assessment of high-quality care to inform supervision decisions, and (3) align rater training with clinical skills education—that is, by having programs that simultaneously allow clinical supervisors to learn and practice clinical skills that underpin high-quality care.

**Redefine workplace-based assessments’ frame-of-reference and rating scale anchors**

Many workplace-based assessment tools have not articulated what frame of reference (e.g., the standard by which to assess performance, such as criterion based or normative) should be used when assessing trainees. Most tools have anchors such as “unsatisfactory,” “satisfactory,” or “superior,” or “doesn’t meet expectations,” “meets expectations,” and “exceeds expectations,” with little guidance as to what informs that decision.\textsuperscript{44} Furthermore, one might ask, “Satisfactory for whom?”

If quality clinical care is the assessment end point, then assessment of trainees should be based on those competencies needed to achieve high-quality care, thereby enabling medical educators to prepare trainees for independent practice.\textsuperscript{44} Therefore, we propose that the education community strive to make the frame of reference of workplace-based assessments the patient rather than the more commonly used frames of reference such as self, trainee level, or a gestalt or gut feeling.\textsuperscript{25}

In *Crossing the Quality Chasm: A New Health System for the 21st Century,*\textsuperscript{44} the IOM articulated the six aims of quality in medical care: it should be safe, effective, efficient, patient centered, timely, and equitable. Reconceptualizing medical education from the desired end point of clinical care quality situates patient care at the center of the assessment discussion rather than making it a peripheral, discretionary topic.\textsuperscript{43,45,46} Therefore, we propose that the definition of *satisfactory* when doing workplace-based assessments be what is required to achieve high-quality patient care (as defined by the IOM and evidence-based best practices) in the unsupervised setting. For example, if the nine-point version of the mini-CEX was used,\textsuperscript{21,47,54} which is defined as “satisfactory,” would be synonymous with “competent care.” This would be the minimal floor for any patient cared for in a training setting. In other words, the scale’s midpoint, by definition, must equate to the trainee’s ability to provide safe, effective, patient-centered care in independent practice.

This reconceptualization represents a significant shift in our training culture. Although faculty may believe that they take these domains into consideration, rater assessment research suggests otherwise (ratings are idiosyncratic, normative, inflated, and fail to align with criterion-referenced quality of care).\textsuperscript{21,23,47} For example, our experience in leading faculty development sessions on trainee assessment revealed that many faculty see patient-centered skills as “icing on the cake” rather than as necessary for competent care, despite the fact that these skills are linked to health outcomes.\textsuperscript{48}

Redefining *satisfactory* aligns with the transition toward criterion-based competency assessment.\textsuperscript{49} Competency-based assessment tools are emerging in which the rating system is based on the trainee’s readiness for safe, independent practice.\textsuperscript{50} Aligning rating scales to the construct of clinical independence, or entrustability, may improve score reliability and assessor discrimination, reduce assessor disagreement, and be more cognitively aligned with the reality map of the assessors (i.e., resonate with raters’ experiences of what they can assess).\textsuperscript{51} Knowing whether trainees are achieving important developmental milestones on time is clearly important. However, by consistently applying the lens of high-quality patient care, we also keep the focus on what is most affected by their competencies: the patient.

What are potential implications for using this frame of reference (the patient) and this anchor (*satisfactory*) is the
ability to provide competent care unsupervised? Many trainees would be rated “unsatisfactory,” unlike the situation now, where grade inflation is endemic. Both faculty and trainees would need to recognize that many ratings would fall in the lower end of the scale. This would necessitate a shift in the educational culture of what unsatisfactory means—no small feat. Second, educators or trainees might see “competent” as the goal, rather than proficiency, expertise, or mastery. Although “competent” may anchor the midpoint of any scale, we believe it is important to emphasize that being competent is not the aspiration, it is the floor. Articulating and defining aspirational, superior clinical skills will be necessary. For example, “superior” might reflect proficient, expert performance defined as intuitive clinical care and fluid performance that happens unconsciously and automatically, no longer depending on explicit knowledge.

Link in-training assessment with the assessment of high-quality care to inform supervision decisions

The utility elements of assessments are validity, reliability, educational impact, practicability, acceptability, and cost-effectiveness, and the utility value of an assessment becomes null and void if any of the utility factors become zero. For example, an assessment would not have utility if it had no validity, even if there was good reliability, impact, practicability, acceptability, and cost-effectiveness.

We suggest that there is a missing factor in the current utility equation: the ability to assess the patient care outcome. If an assessment tool fails to consider the outcome of safe, effective, patient-centered care, we submit that the utility index for that assessment tool is also null and void. Although validity has been framed in terms of quality of patient care, that framing does not necessarily ensure that the patient receives good care. The flaw in previous conceptions of assessment is that the patient is treated only as a contextual variable, something to adjust for. We believe that an assessment tool cannot be agnostic to what happens to the patient. The patient, and the care that the patient receives, must be part of what is assessed in an observation.

We believe that medical educators can improve the validity of a tool’s assessments while concomitantly ensuring safe, effective, patient-centered care by changing the frame of reference and accountability goals of the assessment. Although assessment is integral to learning, assessment tools must help clinical supervisors determine what a trainee can do independently. Supervisors can then use the assessment to inform what they must contribute to the care of the patient and in what capacity.

To highlight this idea conceptually, we developed the AAQCS equation (portrayed in Figure 1). Modeled on the utility index equation, the AAQCS underscores that what a trainee does (which is a function of that individual’s level of competence in a particular context) combined with an appropriate level of supervision (a function of the supervisor’s competence in context) must equate to safe, effective, patient-centered care. The “equation” is not something that is mathematically computed into a “score” but, instead, exemplifies that it is the interaction between trainee performance and supervision that produces, or ensures, high-quality patient care. A faculty member doing workplace-based assessment often dually serves as an assessor of a trainee and as the clinical supervisor responsible for the patient’s care. Anchoring assessment tools in high-quality care (defined as safe, effective, patient centered) reminds faculty who supervise trainees that their assessments reflect both the care that a trainee is independently able to provide the patient in a particular encounter and the quality of the care rendered. If care does not meet the standard of satisfactory (i.e., safe, effective, patient centered), faculty would be prompted to contribute to the encounter so that the care delivered rises to the minimal standard of competent (and hopefully aspirational) care. The AAQCS could also guide trainees’ self-assessment in action and clarify the standards against which they should be interpreting and judging their performance.

Clinical supervisors need to determine whether a resident can be entrusted with the tasks or activities critical to the profession. Currently, the intensity of clinical supervision is predominantly a “one size fits all” approach using the resident’s training level or time as a proxy for competence. Often, on the basis of this assumption, faculty assign tasks at higher levels of responsibility than they believe suitable. For example, in U.S. outpatient continuity clinics, attending physicians are required to see residents’ patients prior to the conclusion of the patient’s office visit only for the first six months of residents’ training. However, this Medicare time policy is often discordant with residents’ actual achievement of the expected milestones that would allow less direct supervision after six months of training. Not all residents reach the same level of competence in all activities at exactly the same time, and not all are ready for indirect supervision after six months of training. When satisfactory is defined as

![Figure 1 A graphic portrayal of the Accountable Assessment for Quality Care and Supervision equation (AAQCS) developed by the authors. The AAQCS underscores that what a trainee does (which is a function of that individual’s level of competence in a particular context) combined with an appropriate level of supervision (a function of the supervisor’s competence in context) must equate to safe, effective, patient-centered care. The “equation” is not something that is mathematically computed into a “score” but, instead, exemplifies that it is the interaction between trainee performance and supervision, as described above, that produces, or ensures, high-quality patient care. The purpose of the equation is to remind faculty that supervision is a dynamic, complex process essential for patients to receive high-quality care.](attachment:Figure 1.png)
the ability to provide high-quality care unsupervised, there is an inherent shift to deliberate, competency-based, informed supervision. The intensity of supervision is then grounded in the trainee’s skill level in a given clinical context.63,64

To implement such an approach will be daunting, yet doing so is imperative, given calls for public accountability. Because direct observation is infrequent, strategies ensuring patient safety in nonobserved encounters are needed. Developing a method for prioritizing the many different assessments/observations that could be performed to evaluate trainee competence is essential.65 Preferentially focusing observation early or when new skills are expected—for example, with simulation—could provide a realistic understanding of trainee skills.66,67 Given the discontinuity in clinical training,68 enhanced communication about learners’ skills has greater importance. For trainees, instilling a culture that promotes them to self-monitor and empowers them to request direct observation is needed.69

Align rater training with clinical skills education
If we want to train learners to be competent in providing high-quality care, faculty must be familiar with and trained in the requisite skills. As previously described, faculty frequently use themselves as a frame of reference when assessing residents.25 This is problematic if the faculty are not providing high-quality care.71–73 Therefore, many practicing physicians who supervise and assess trainees will need to improve or learn new clinical skills and practice them to reach proficiency.32

We have a unique opportunity to leverage the rater assessment problem by providing opportunities for clinical supervisors to learn and/or practice the very skills they are assessing. For example, there are best practices for patient-centered communication,70 informed decision making,71 breaking bad news,71 motivational interviewing,72 the evidence-based physical exam,73 effective use of the electronic medical record during a patient visit,74 and assessment of the geriatric patient.75,76 Reviewing and practicing these frameworks with faculty may provide an opportunity to not only define best practices but also to raise the floor for the care we provide. How to best accomplish this remains uncertain, but it would inevitably require a national faculty development curriculum.32

A Fundamental Shift Is Needed in Trainee Assessment
To be professionally accountable and attain the public’s trust, the onus is on us, as medical educators, to make good assessment decisions. The interrater variability of work-based assessments is not just an educational issue but also a patient care and safety issue. Medical education and health care delivery are intertwined. As such, part of the evaluation of trainees should consider the quality of patient care.77

When embedded in patient care, direct observation is still one of the best methods to assess how trainees integrate multiple skills in the real-world setting. However, its effectiveness is profoundly dependent on the faculty members’ skills. Done well, direct observation can be a quality improvement intervention for both the trainee and the patient. When observation is grounded in concepts of high-quality care, there is the potential to improve the trainee’s skills and competence (i.e., to catalyze the trainee’s additional learning). Simultaneously, identifying gaps in the trainee’s ability to provide safe, effective, patient-centered care informs decisions regarding supervision and entrustment, which can influence patient safety.

Now is the time to elevate direct observation from trainee assessment alone to trainee assessment that improves the quality of care patients receive in real time. Doing so offers opportunities to realize improved patient safety and decreased costs. We recognize this is a potentially controversial, fundamental shift in how assessment is currently conceptualized. Yet, there are real costs for not adopting this conceptualization of assessment. This approach will require faculty development—both for developing teaching skills and perhaps learning or refreshing clinical skills and best clinical practices.

Trainee assessment, clinical care supervision, and faculty members’ own clinical skills are inextricably linked to the process of ensuring high-quality patient care in the training environment. These interrelationships reinforce the essential and indispensable role of clinical faculty in the training and patient care missions. Nonphysicians can assess trainees’ skills, but they cannot provide clinical supervision. Intrinsically, clinician educators are fundamental to workplace-based assessment because they can be held accountable to ensuring that an individual patient and the population of patients that the trainee will see in the future receive high-quality care. Clinician educators are vital because they can bridge potential gaps in care, thereby elevating care quality. Without outstanding, capable clinicians who are also exceptional educators, our medical education system will fail to produce physicians who will meet the needs of our society. Now is the time to ensure that our health care and medical education systems support clinician educators to acquire the assessment and clinical skills they need to ensure that outcome.

Funding/Support: American Board of Internal Medicine.

Other disclosures: Dr. Holmboe receives royalties from Mosby-Elsevier.

Ethical approval: Reported as not applicable.

Dr. Kogan is associate professor, Department of Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania.

Ms. Conforti is research associate for academic programs, American Board of Internal Medicine, Philadelphia, Pennsylvania.

Dr. Iobst is vice president of academic affairs and clinical affairs and vice dean, Commonwealth Medical College, Scranton, Pennsylvania. During the preparation of this article, he was vice president of academic affairs, American Board of Internal Medicine, Philadelphia, Pennsylvania.

Dr. Holmboe is senior vice president of milestones development and evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois. During the preparation of this article, he was chief medical officer, American Board of Internal Medicine, Philadelphia, Pennsylvania.

References
Perspective


Pauw DS, Wernich MD, Curtiss JR, Carlene JD, Ramsey PG. Ability of physicians to recognize physical findings associated with HIV infection. JAMA. 1995;274:1380–1382.


