Numbers Encapsulate, Words Elaborate: Toward the Best Use of Comments for Assessment and Feedback on Entrustment Ratings

Shiphra Ginsburg, MD, PhD, Christopher J. Watling, MD, PhD, Daniel J. Schumacher, MD, PhD, MEd,
Andrea Gingerich, PhD, and Rose Hatala, MD, MSc

Abstract

The adoption of entrustment ratings in medical education is based on a seemingly simple premise: to align workplace-based supervision with resident assessment. Yet it has been difficult to operationalize this concept. Entrustment rating forms combine numeric scales with comments and are embedded in a programmatic assessment framework, which encourages the collection of a large quantity of data. The implicit assumption that more is better has led to an untamable volume of data that competency committees must grapple with. In this article, the authors explore the roles of numbers and words on entrustment rating forms, focusing on the intended and optimal use(s) of each, with a focus on the words. They also unpack the problematic issue of dual-purposing words for both assessment and feedback. Words have enormous potential to elaborate, to contextualize, and to instruct; to realize this potential, educators must be crystal clear about their use. The authors set forth a number of possible ways to reconcile these tensions by more explicitly aligning words to purpose. For example, educators could focus written comments solely on assessment; create assessment encounters distinct from feedback encounters; or use different words collected from the same encounter to serve distinct feedback and assessment purposes. Finally, the authors address the tyranny of documentation created by programmatic assessment and urge caution in yielding to the temptation to reduce words to numbers to make them manageable. Instead, they encourage educators to preserve some educational encounters purely for feedback, and to consider that not all words need to become data.

> Simplicities are enormously complex.
> Consider the sentence "I love you."
>
> —Richard O. Moore,
> *Writing the Silences*, 2010

The idea of entrustment ratings is seductive because it combines opportunity and economy. On a daily basis, supervisors make judgments to entrust trainees to carry out specific patient care tasks, with varying amounts of supervision. Why not harness these naturally occurring judgments to drive meaningful resident assessment? The idea has an irresistible simplicity: to align the construct of how supervisors "work with and make decisions about trainees in the workplace" with resident assessment.[1] And yet, it has been exceedingly difficult to operationalize such a seemingly straightforward concept.

A typical entrustment rating form combines a numeric scale with written comments. Including a numeric scale seems intuitive, as we are drawn to numbers for their simplicity, their ability to encapsulate, and their familiarity. Numbers have a veneer of objectivity that aligns with many aspects of our biomedical world. But we also recognize the limitations in using numbers alone, which can seem dry, sterile, and lacking in contextual details. Enter words, which hold the potential to justify, enrich, or supplant the numbers.[2–4] Words provide novel information that numbers cannot. But who will read the words, how will they be interpreted, and what will be done with them?

A brief aside on the use of the term "words." The word "narrative" has been used in many studies focused on written assessment comments,[2,5] but we will specifically avoid its use here. Narrative implies a story, and this fits well in research focused on in-training evaluation reports (ITERs), whose purpose is to integrate, synthesize, and document observations from a multiweek rotation.[4] The comments included with most entrustment ratings, however, are related to a single observed encounter and are by design much shorter—more like a text message than a story. For example, in psychiatry, 98% of completed entrustable professional activity (EPA) forms contained only a single comment.[6,7] Thus, throughout this article, we will refer to "comments" or "words" to avoid holding the words on entrustment rating forms to an unattainable standard of "narrative."

A variety of issues have arisen as we have implemented entrustment rating forms, not least of which is that we are drowning in a sea of data, both numbers and words, for each learner. Indeed, entrustment ratings are meant to be used within a system of programmatic assessment in which multiple "low-stakes" assessments are considered together by a competency committee, which makes summative decisions about trainees' progress. Programmatic assessment encourages the collection of a large quantity of data, with an implicit assumption that more is better. It may help us to manage and understand these data if we unpack the strengths and limitations of the numbers and the words. We need to consider that numbers and words have different affordances; for example, numbers encapsulate and

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

Correspondence should be addressed to Shiphra Ginsburg, Mount Sinai Hospital, 433-600 University Ave., Toronto, ON M5G 1X5, Canada; telephone: (416) 586-8671; email: shiphra.ginsburg@utoronto.ca.

Acad Med. 2021;96:S81–S86.
doi: 10.1097/ACM.0000000000004089

words elaborate. Numbers and words speak to different audiences ranging from learners and supervisors to programs and society. Numbers and words reflect fundamentally different philosophical positions, with numbers reflecting a positivist or postpositivist view of the world and words more aligned with constructivism.[8] Confounding the duality of numbers versus words is a duality of purposes that the entrustment rating forms are meant to serve—summative assessment and developmental feedback—which are sometimes at odds.

In this article, we will focus on entrustment rating form comments, examining the purposes, strengths, and limitations of numbers and words, with an emphasis on the words. We will highlight key challenges in these ratings and suggest some forward directions that may rescue us from the sea of data and bring us to shore.

Who Needs Numbers and for What Purpose?

We first turn our attention to the numbers on entrustment-supervision scales. Entrustment rating forms commonly use a 4- or 5-point scale with each ascending number tied to an anchor that represents a discrete supervisory judgment or decision ordered from most to least amount of supervision provided.[1,9–13] Numbers are not actually required for entrustment scales, but when they are used, they should be thought of as succinctly representing a shorthand code for a particular supervisory decision that was made; that is, the numbers serve as a label but not as a count or measure.[14]

The numbers on entrustment scales can be used to efficiently document a supervisory decision that was made in the moment or record a proclamation of which level of supervision should be used in the future.[1,15] A number can serve as a data point for the program, administration, and competency committees, and as an "aide memoire" for the trainee. When used for summative purposes, the number concisely documents proof of merit.[16] Numbers also lend themselves to mathematics that can reliably combine, filter, and summarize the data points to provide precise numerical representation of a large dataset, although treating entrustment-supervision numbers in this way is controversial.[1] Unfortunately, the process to translate a numerical representation back into high-stakes supervisory, progression, and/or competence decisions is less obvious. These computations use numbers stripped bare of the contextual details of the trainee's engagement with the activities, patients, and supervisors, and their perceived responses to those interactions, feedback, and outcomes. It may be this barrenness that makes us wary of relying on numbers without words and cautions us against limiting feedback solely to ratings.[17] As is discussed in the following section, written comments offer information that numbers (and their anchors) on scales cannot.

Who Needs Comments and for What Purpose?

The comments on entrustment rating forms can serve a variety of functions. If we start from the learner's perspective, comments can provide developmental feedback. Learners can use specific, actionable comments to create learning goals and reflect on their progress in meeting these goals as training progresses.[18] From a supervisor's perspective, comments can justify and provide context to support a rating or decision. Comments can also be used to capture what may not be represented in the numeric scales, such as certain aspects of professional behavior.[3,5] Comment boxes on rating forms can be used to send messages to programs, sometimes by using "coded" language meant to allow trainees to save face.[19,20] From a program perspective, comments can help identify learners in difficulty earlier than the scores alone and can change summative decisions when used in combination with scores.[21–23] In aggregate, comments become part of a portfolio or formal record that can be used for high-stakes decision making.[24] Clearly, some of these purposes are in conflict, as it is difficult for supervisors to provide constructive feedback in writing without considering the potential downstream effect their words might have when it comes to decision making.[20,25–27]

Written comments have been shown to be reliable and valid when it comes to decision making in both ITER and OSCE settings.[2,28,29] One study of ITER comments found that the comments have higher reliability than numeric scores, while requiring less data.[2] Written comments are good at providing contextual detail related to a particular rating, and longer, more specific comments can make residents feel more valued.[26,27] We have less data specifically on EPA comments, but some reports suggest that these comments are more specific and behavioral than what is usually reported on ITERs,[18] and they may capture information that is not otherwise included in the scales.[30] So while we urge caution in extrapolating from ITER "narrative" data to entrustment rating comments, early results suggest they may be similarly useful.

Of course there are numerous critiques of written comments as well, including that comments are too vague and nonspecific to be useful and that they have not been shown to lead to learning improvement.[31,32] In comparison to numbers, comments are often derided as being "too subjective," even though assigning a numeric score to an observed performance is also a subjective act.[33] Concerns have been expressed that in moving "beyond psychometrics," we may have swung the pendulum too far in the wrong direction and have begun to undervalue appropriate use of numeric scores.[34] Finally, written assessment comments can reproduce or promote implicit bias that can be harmful to certain groups, such as women or under-represented minorities.[35,36]

Is It a Problem That the Comments and the Scores Are Doing Something Different?

While numbers offer an appealing shorthand for representing learner performance, comments promise a more elaborate picture. Holmboe et al called for a better balance between the quantitative and the qualitative elements of assessment information, noting that numbers are but a code, incomplete without attachment to the meaning and nuance that only words can offer.[14] As ten Cate and Regehr note, entrustment decisions made on the frontlines of the clinical learning environment inherently necessitate a judgment of perceived risk by the supervisor.[33] This, they argue, advantages comments, noting: "documentation of the preceptor's *subjective* experience is the only truly defensible proposition."[33] Comments add subtlety and substance to the bluntness of numbers by providing the rationale for an assessment rating, highlighting the contextual caveats related to a particular observed performance, and articulating the experience of supervising that performance. By enriching the assessment data available, words improve perceptions of fairness, bolster defensibility, and facilitate group decision making in competence committee settings.[37] This potential is most readily realized, however, when numbers and words are aligned both philosophically and around purpose.[8] Using Kane's validity framework,[38] in this instance, the intended use of the entrustment rating form (both numbers and words) is to provide a judgment regarding whether the resident can be entrusted with that task in the future. The number provides a readily recognized and easily processed label to classify the observed performance, and the words explain and justify the choice of that label. Simple.

But entrustment rating forms are deeply embedded in programmatic assessment, and programmatic assessment expects more from words. Programmatic assessment aims not only to assess learner performance but also to stimulate learner development. The words on entrustment rating forms must somehow serve both aims. They must explain and rationalize a judgment or decision on the one hand, while offering feedback and coaching to motivate improvement on the other. If we consider Kane's validity framework as another way of conceptualizing this problem, there is not a single intended use for the words.[38] As currently operationalized within programmatic assessment, the words serve a dual purpose: a) to be used summatively to contribute toward promotion decisions for learners (which aligns with the intended use for the numbers) and b) to provide developmental feedback to learners.

This double-barreled expectation complicates things immensely. Schut et al point out that rich, narrative feedback is critical to harnessing the developmental aims of programmatic assessment.[39] But can the same words that justify a rating also constitute "rich narrative feedback"? Probably not. Tavares and colleagues, in a controlled study of videotaped OSCE performances, reported that raters engage with assessment tasks similarly, whether they are intended as summative or formative, mainly because they consider all assessment tasks as summative.[40] Thus, it may be a leap to assume that even if assessors do write similar words regardless of the intended purpose of their comments, that their words can be equally *effective* as both assessment and feedback, or that they will be interpreted the same way. One recent study found that the same words may take on different meanings when considered for different purposes, for example, when appearing on an assessment form versus a reference letter.[20] Schut and colleagues, in fact, have problematized this tension, suggesting that learning may be stymied unless the developmental purpose of assessment tasks receives careful attention.[41] In the absence of such careful attention, learners tend to perceive all observations as judgment. Learners may interact with comments differently depending on whether they believe their purpose is to pass judgment or to aid development.

Reconciling Tensions: Aligning Words With Purpose

How can we remain true to the simplicity of capturing supervisory decisions in entrustment-based assessment moments, while encouraging a developmental mindset? In 2 studies of entrustment-based assessments in the Canadian context, residents perceived that the required volume and summative intent of EPA assessments led to a "tick-box" exercise by both residents and faculty, increased the volume of feedback at the expense of lower-quality feedback conversations, and strained the resident–supervisor relationship.[42,43] Residents also perceived that verbal feedback had greater value and utility than numbers and written comments, so much so that they sometimes circumvented the form-filling exercise to engage in learning conversations.[42] This leads to some interesting potential directions forward which we explore below, including separating assessment activities from developmental ones and considering different roles for spoken versus written words.

If we wish to harness the developmental potential of words in our entrustment-based assessments, we must carefully consider whether dual purposing for both assessment and feedback is the best direction forward.[25] Dual purposing may end up not serving either intended use well.[25,44,45] At a conceptual level, dual purposes seem an impossibility as different inferences in the validity argument come into play to support different intended uses.[38] However, some thought leaders in our field see dual purposing as a possibility—that an assessment does not have to be "either-or" but rather can be "both-and."[46] Sorting out this conundrum is crucial.

We could stop dual purposing our forms and instead align the numbers and the words around the explicit purpose of assessment. Written words would then be used to justify and provide context for the numerical rating. The intended audience would be the program, and the entrustment rating forms would be solely for assessment. In this approach, we would harness *the encounter* as the developmental opportunity but not *the form*. In other words, not every aspect of the encounter would have to serve an assessment focus and, if the resident and supervisor engaged in a rich, unrecorded learning conversation, the spoken words could be used developmentally. If the resident wished, they could write notes for themselves—souvenirs of the conversation they could keep privately and draw on later to stimulate self-reflection and development. This approach would not get around the issue of learners engaging in staged performances during assessment contexts,[47,48] nor would it address the potential problems that might ensue if there is a disconnect between what is discussed in person and what is recorded on a form, but it would open a space for learning conversations.[49]

As an alternative within this single-purpose approach, we could consider employing separate assessors for the entrustment ratings, thus making a clear distinction between assessment encounters (conducted by assessors) and feedback encounters (conducted by supervisors).[16,50] This separation could free up supervisors to truly be coaches to the residents, engaged in direct observation and feedback conversations to foster resident development.[51] Both of these approaches recognize the need for culture change. Supporting meaningful learning conversations requires more than training teachers in assessment and feedback or encouraging learners to adopt a growth mindset. It also requires the deliberate adoption of organizational strategies that set the stage for effective learning conversations to occur.[52]

Which brings us to a third possible course of action, in which we leverage the separate strengths of numbers and words to explicitly dual purpose and embody a “both-and” approach. Numbers would remain focused on assessment, but words would be used both for assessment and for feedback, with programs having to clearly specify the purpose of the words and their intended use. To optimize the dual purposes that comments can serve, both would have to be distinct on the frontline assessment forms that supervisors complete. Dory and colleagues have shown that simple “nudge interventions” on assessment reports, such as putting the comment box first instead of last, can increase the level of detail and actionability of the comments that teachers provide. Building on that work, we could include prompts on entrustment rating forms that address purpose and intended audience. Different prompts could be used to elicit either assessment comments (those intended to explain observations or justify ratings) or feedback comments (those intended to coach or shape continued development). The intended audience for the former is the competency committee, but these comments would be visible to the trainee. The latter comments would be intended for the trainee, and we argue these could, and likely should, be hidden from the competency committee. Coaching works best when learners feel safe to be vulnerable,[53,54] and providing them with data for their eyes only may cement that sense of safety.

If supervisors can feel confident that their words are intended solely to support learner development, they might feel less constrained and therefore write more honest, critical comments. Their comments, in turn, might more meaningfully inform next steps in development, without learners harboring concerns that they will be used to render a higher-stakes decision about performance. Achieving this clarity of purpose would require system change. It would not be sufficient to simply tell teachers or learners that a particular set of comments should be treated as assessment or as feedback, as evidence suggests this would not have much effect on either the perceived stakes of the interaction or on the ratings and comments produced.[40,44] Achieving these dual purposes means that the 2 sets of comments may not be perfectly aligned with one another. But alignment isn’t the goal. Decision makers need words that allow them to understand learner performance so they can make trustworthy and defensible decisions. Learners need words that are tailored to their developmental trajectory and that support continued improvement. Both purposes could be well served, with separate words for each.

Programmatic Assessment and the Tyranny of Documentation

A final problem to grapple with is what to do with the sheer volume of comments that are produced for each learner. Within a programmatic assessment framework, “massive information” is gathered over time[55] and necessitates a system to make meaning from a variety of data sources. Competency committees may find it difficult to make decisions based on numeric data alone, yet may struggle to read and interpret dozens or hundreds of comments per resident. In response to this problem, some authors have explored reducing words to numbers, through methods such as natural-language processing. If what one is looking for is a “signal” or “code” that identifies to the program those learners who need more attention, then the appeal of numeric scores for this purpose is obvious, as numbers can be summarized efficiently and can act as a “first-pass filter” that can help focus subsequent review.[34] Following this logic, several researchers have attempted to do the same with words, using computer algorithms to screen for or predict learners in difficulty. In one study, keyword algorithms identified more residents in difficulty than the numeric scores suggested,[56] but the overall feasibility and utility of this approach have not been established. Furthermore, reducing words to numbers loses all the potential of what the words might offer in terms of context and description. We strongly advocate to let words be as they are, to use the numbers to search for signals in the data, and to read the words that accompany those signals to provide context for the assessment encounters they describe.

Paradoxically, the problem that programmatic assessment was trying to solve (not enough assessment data) has now created a new problem: the tyranny of documentation. If we require every interaction with a learner to turn a feedback moment into an assessment moment that must be recorded, we undermine the developmental relationship between supervisor and resident. An assessment-dominant context removes the safe spaces for learning and risks guiding residents toward inauthentic staged performances for assessment purposes, which are of low value for meaningful feedback.[47] Perhaps we should question the assumption that “more is better” and consider that less data—but the right kind of data—may serve our purposes equally well.

Conclusion

As a community, we have implemented entrustment rating forms with both numbers and words, with perhaps insufficient attention to the purpose and intended use of each. Because of the ambiguity as to whether the words are for feedback or assessment, the higher-stakes purpose tends to dominate and the default has been to treat words as assessment, which surely represents a lost opportunity for meaningfully helping trainees take the next steps in their ongoing development. Shifting our energies toward achieving clarity of purpose, and experimenting with different approaches (single purpose, dual purpose) to understand whether and how assessment and feedback may be achieved through words, may be a helpful way forward. Entrustment rating forms are linked with programmatic assessment, and we need to foreground the ways in which programmatic assessment can be implemented to encourage developmental purposes—of which the most important may be focusing on longitudinal, trusting relationships between learners and supervisors as the context in which learning conversations can occur.[39,57] Words have enormous potential to elaborate, to contextualize, and to instruct. To realize this potential, we must be crystal clear about their intended use and work toward aligning *how and when* we collect words with *why*. We encourage educators to preserve some educational encounters purely for feedback and to consider that not all words need to become data.

Funding/Support: S. Ginsburg is supported by a Canada Research Chair in Health Professions Education.

Other disclosures: None reported.

Ethical approval: Reported as not applicable

S. Ginsburg is professor of medicine, Department of Medicine, Sinai Health System and Faculty of Medicine, University of Toronto, scientist, Wilson Centre for Research in Education, University of Toronto, Toronto, Ontario, Canada, and Canada Research Chair in Health Professions Education; ORCID: http://orcid.org/0000-0002-4595-6650.

C.J. Watling is professor and director, Centre for Education Research and Innovation, Schulich School of Medicine & Dentistry, Western University, London, Ontario, Canada; ORCID: https://orcid.org/0000-0001-9686-795X.

D.J. Schumacher is associate professor of pediatrics, Cincinnati Children's Hospital Medical Center and University of Cincinnati College of Medicine, Cincinnati, Ohio; ORCID: https://orcid.org/0000-0001-5507-8452.

A. Gingerich is assistant professor, Northern Medical Program, University of Northern British Columbia, Prince George, British Columbia, Canada; ORCID: https://orcid.org/0000-0001-5765-3975.

R. Hatala is professor, Department of Medicine, and director, Clinical Educator Fellowship, Center for Health Education Scholarship, University of British Columbia, Vancouver, British Columbia, Canada; ORCID: https://orcid.org/0000-0003-0521-2590.

References

1 ten Cate O, Schwartz A, Chen HC. Assessing trainees and making entrustment decisions: On the nature and use of entrustment-supervision scales. Acad Med. 2020;95:1662–1669.
2 Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: A quantitative reliability analysis of qualitative data. Acad Med. 2017;92:1617–1621.
3 Hanson JL, Rosenberg AA, Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. Front Psychol. 2013;4:668.
4 Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. Acad Med. 2016;91:1359–1369.
5 Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H. Competencies "plus": The nature of written comments on internal medicine residents' evaluation forms. Acad Med. 2011;86(10 suppl):S30–S34.
6 Young JQ, McClure M. Fast, easy, and good: Assessing entrustable professional activities in psychiatry residents with a mobile app. Acad Med. 2020;95:1546–1549.
7 Diller D, Cooper S, Jain A, Lam CN, Riddell J. Which emergency medicine milestone sub-competencies are identified through narrative assessments? West J Emerg Med. 2019;21:173–179.
8 Tavares W, Kuper A, Kulasegaram K, Whitehead C. The compatibility principle: On philosophies in the assessment of clinical competence. Adv Health Sci Educ Theory Pract. 2020;25:1003–1018.
9 Weller JM, Castanelli DJ, Chen Y, Jolly B. Making robust assessments of specialist trainees' workplace performance. Br J Anaesth. 2017;118:207–214.
10 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.
11 Chen HC, van den Broek WE, ten Cate O. The case for use of entrustable professional activities in undergraduate medical education. Acad Med. 2015;90:431–436.
12 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.
13 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.
14 Holmboe ES, Yamazaki K, Hamstra SJ. The evolution of assessment: Thinking longitudinally and developmentally. Acad Med. 2020;95(11 suppl):S7–S9.
15 van Enk A, ten Cate O. "Languaging" tacit judgment in formal postgraduate assessment: The documentation of ad hoc and summative entrustment decisions. Perspect Med Educ. 2020;9:373–378.
16 Prentice S, Benson J, Kirkpatrick E, Schuwirth L. Workplace-based assessments in postgraduate medical education: A hermeneutic review. Med Educ. 2020;54:981–992.
17 Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: The do's, don'ts and don't knows of direct observation of clinical skills in medical education. Perspect Med Educ. 2017;6:286–305.
18 Young JQ, Sugarman R, Holmboe E, O'Sullivan PS. Advancing our understanding of narrative comments generated by direct observation tools: Lessons from the psychopharmacotherapy-structured clinical observation. J Grad Med Educ. 2019;11:570–579.
19 Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: Faculty interpretations of narrative evaluation comments. Med Educ. 2015;49:296–306.
20 Ginsburg S, Kogan JR, Gingerich A, Lynch M, Watling CJ. Taken out of context: Hazards in the interpretation of written assessment comments. Acad Med. 2020;95:1082–1088.
21 Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? Teach Learn Med. 1993;5:10–15.
22 Guerrasio J, Cumbler E, Trosterman A, Wald H, Brandenburg S, Aagaard E. Determining need for remediation through postrotation evaluations. J Grad Med Educ. 2012;4:47–51.
23 Lefebvre C, Hiestand B, Glass C, et al. Examining the effects of narrative commentary on evaluators' summative assessments of resident performance. Eval Health Prof. 2020;43:159–161.
24 Schumacher DJ, Michelson C, Poynter S, et al; APPD LEARN CCC Study Group. Thresholds and interpretations: How clinical competency committees identify pediatric residents with performance concerns. Med Teach. 2018;40:70–79.
25 Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. 2019;53:76–85.
26 Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: Residents' interpretations of written assessment comments. Med Educ. 2017;51:401–410.
27 Patel R, Drover A, Chafe R. Pediatric faculty and residents perspectives on in-training evaluation reports (ITERs). Can Med Educ J. 2015;6:41–53.
28 Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA. Using In-Training Evaluation Report (ITER) qualitative comments to assess medical students and residents: A systematic review. Acad Med. 2017;92:868–879.
29 Wilby KJ, Govaerts MJB, Dolmans DHJM, Austin Z, van der Vleuten C. Reliability of narrative assessment data on communication skills in a summative OSCE. Patient Educ Couns. 2019;102:1164–1169.
30 Colbert-Getz JM, Lappe K, Northrup M, Roussel D. To what degree are the 13 entrustable professional activities already incorporated into physicians' performance schemas for medical students? Teach Learn Med. 2019;31:361–369.
31 Jackson JL, Kay C, Jackson WC, Frank M. The quality of written feedback by attendings of internal medicine residents. J Gen Intern Med. 2015;30:973–978.
32 Lye PS, Biernat KA, Bragg DS, Simpson DE. A pleasure to work with: An analysis of written comments on student evaluations. Ambul Pediatrics. 2001;1:128–131.
33 ten Cate O, Regehr G. The power of subjectivity in the assessment of medical trainees. Acad Med. 2019;94:333–337.
34 Pearce J. In defence of constructivist, utility-driven psychometrics for the 'post-psychometric era'. Med Educ. 2020;54:99–102.
35 Rojek AE, Khanna R, Yim JWL, et al. Differences in narrative language in evaluations of medical students by gender and under-represented minority status. J Gen Intern Med. 2019;34:684–691.
36 Mueller AS, Jenkins TM, Osborne M, Dayal A, O'Connor DM, Arora VM. Gender differences in attending physicians' feedback to residents: A qualitative analysis. J Grad Med Educ. 2017;9:577–585.
37 Valentine N, Durning S, Shanahan EM, Schuwirth L. Fairness in human judgement in assessment: A hermeneutic literature review and conceptual framework. Adv Health Sci Educ. 2020.
38 Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. Med Educ. 2015;49:560–575.
39 Schut S, Heeneman S, Bierer B, Driessen E, van Tartwijk J, van der Vleuten C. Between trust and control: Teachers' assessment conceptualisations within programmatic assessment. Med Educ. 2020;54:528–537.
40 Tavares W, Young M, Gauthier G, St-Onge C. The effect of foregrounding intended use on observers' ratings and comments in the assessment of clinical competence. Acad Med. 2020;95:777–785.
41 Schut S, Maggio LA, Heeneman S, van Tartwijk J, van der Vleuten C, Driessen E. Where the rubber meets the road—An integrative review of programmatic

assessment in health care professions education. Perspect Med Educ. 2021;10:6–13.

42 Branfield Day L, Miles A, Ginsburg S, Melvin L. Resident perceptions of assessment and feedback in competency-based medical education: A focus group study of one internal medicine residency program. Acad Med. 2020;95:1712–1717.

43 Martin L, Sibbald M, Brandt Vegas D, Russell D, Govaerts M. The impact of entrustment assessments on feedback and learning: Trainee perspectives. Med Educ. 2020;54:328–336.

44 Schut S, Driessen E, van Tartwijk J, van der Vleuten C, Heeneman S. Stakes in the eye of the beholder: An international study of learners' perceptions within programmatic assessment. Med Educ. 2018;52:654–663.

45 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.

46 Govaerts MJB, van der Vleuten CPM, Holmboe ES. Managing tensions in assessment: Moving beyond either-or thinking. Med Educ. 2019;53:64–75.

47 LaDonna KA, Hatala R, Lingard L, Voyer S, Watling C. Staging a performance: Learners' perceptions about direct observation during residency. Med Educ. 2017;51:498–510.

48 Gaunt A, Patel A, Rusius V, Royle TJ, Markham DH, Pawlikowska T. 'Playing the game': How do surgical trainees seek feedback using workplace-based assessment? Med Educ. 2017;51:953–962.

49 Tavares W, Eppich W, Cheng A, et al. Learning conversations: An analysis of the theoretical roots and their manifestations of feedback and debriefing in medical education. Acad Med. 2020;95:1020–1025.

50 Voyer S, Cuncic C, Butler DL, MacNeil K, Watling C, Hatala R. Investigating conditions for meaningful feedback in the context of an evidence-based feedback programme. Med Educ. 2016;50:943–954.

51 Cavalcanti RB, Detsky AS. The education and training of future physicians: Why coaches can't be judges. JAMA. 2011;306:993–994.

52 Watling CJ. Unfulfilled promise, untapped potential: Feedback at the crossroads. Med Teach. 2014;36:692–697.

53 Molloy E, Bearman M. Embracing the tension between vulnerability and credibility: 'intellectual candour' in health professions education. Med Educ. 2019;53:32–41.

54 Watling CJ, LaDonna KA. Where philosophy meets culture: Exploring how coaches conceptualise their roles. Med Educ. 2019;53:467–476.

55 van der Vleuten CP, Schuwirth LW, Driessen EW, Govaerts MJ, Heeneman S. 12 Tips for programmatic assessment. Med Teach. 2014:1–6.

56 Tremblay G, Carmichael PH, Maziade J, Gregoire M. Detection of residents with progress issues using a keyword-specific algorithm. J Grad Med Educ. 2019;11:656–662.

57 Schut S, van Tartwijk J, Driessen E, van der Vleuten C, Heeneman S. Understanding the influence of teacher–learner relationships on learners' assessment perception. Adv Health Sci Educ. 2019;25:441–456.