Cracking the code

Cracking the code: residents' interpretations of written assessment comments

Shiphra Ginsburg,[1] Cees PM van der Vleuten,[2] Kevin W Eva[3] & Lorelei Lingard[4]

CONTEXT Interest is growing in the use of qualitative data for assessment. Written comments on residents' in-training evaluation reports (ITERs) can be reliably rank-ordered by faculty attendings, who are adept at interpreting these narratives. However, if residents do not interpret assessment comments in the same way, a valuable educational opportunity may be lost.

OBJECTIVES Our purpose was to explore residents' interpretations of written assessment comments using mixed methods.

METHODS Twelve internal medicine (IM) postgraduate year 2 (PGY2) residents were asked to rank-order a set of anonymised PGY1 residents (n = 48) from a previous year in IM based solely on their ITER comments. Each PGY1 was ranked by four PGY2s; generalisability theory was used to assess inter-rater reliability. The PGY2s were then interviewed separately about their rank-ordering process, how they made sense of the comments and how they viewed ITERs in general. Interviews were analysed using constructivist grounded theory.

RESULTS Across four PGY2 residents, the G coefficient was 0.84; for a single resident it was 0.56. Resident rankings correlated extremely well with faculty member rankings (r = 0.90). Residents were equally adept at reading between the lines to construct meaning from the comments and used language cues in ways similarly reported in faculty attendings. Participants discussed the difficulties of interpreting vague language and provided perspectives on why they thought it occurs (time, discomfort, memorability and the permanency of written records). They emphasised the importance of face-to-face discussions, the relative value of comments over scores, staff-dependent variability of assessment and the perceived purpose and value of ITERs. They saw particular value in opportunities to review an aggregated set of comments.

CONCLUSIONS Residents understood the 'hidden code' in assessment language and their ability to rank-order residents based on comments matched that of faculty. Residents seemed to accept staff-dependent variability as a reality. These findings add to the growing evidence that supports the use of narrative comments and subjectivity in assessment.

Medical Education 2017: 51: 401–410
doi: 10.1111/medu.13158

[1]Department of Medicine, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[2]Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands
[3]Department of Medicine, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada
[4]Department of Medicine, Schulich School of Medicine and Dentistry, Western University, London, Ontario, Canada

Correspondence: Shiphra Ginsburg, Mount Sinai Hospital, 600 University Avenue, Room 433, Toronto, Ontario M5G 1X5, Canada. Tel: 00 1 416 586 8671;
E-mail: shiphra.ginsburg@utoronto.ca

INTRODUCTION

Assessment in medical education is evolving. Although most assessment is still based on numeric scores, there is increasing interest in the value that written comments might provide, especially in residency education.[1–4] Indeed, the literature involving assessment comments often uses the term 'narrative', reflecting a sense that the comments are intended to convey a story to the reader.[3,5,6] Written comments may be better than scores at identifying learners in difficulty[7] and have the advantage of being able to describe to residents both their strengths and areas in which they require improvement.[8] However, narrative comments are also commonly critiqued for containing language that is frustratingly vague and lacking in meaning.[9,10]

Despite this critique, recent research suggests that it is possible to use comments in a reliable way. For example, one study reported that faculty attendings were reliably able to rank-order residents in postgraduate year 1 (PGY1s) in internal medicine (IM) based on comments alone.[11] Another found that comments in PGY1 could predict performance in PGY3.[12] A further study reported that faculty attendings were adept at 'reading between the lines' to construct meaning from otherwise apparently vague language.[13] These participants were able to make sense of what their colleagues were expressing in their narratives, suggesting a shared understanding of an implicit, or hidden, code in assessment language. The high reliabilities seen in these studies provide some validity evidence for the use of narrative comments in the assessment of residents.[14] However, the potential for comments to be useful to residents (thereby indicating consequential validity) is unknown because we do not know whether residents interpret or 'decode' written comments in the same way as their attendings.

In one study investigating residents' perceptions of their ward assessments, Watling *et al.* reported widespread dissatisfaction.[15] Many residents – from multiple specialties – viewed the workplace-based assessment strategy known as an in-training evaluation report (ITER) as being of 'limited importance to their professional development'.[15] The dominant theme identified in these interviews was the importance of engagement, on the part of both the faculty member and the resident him or herself. Similar findings have been reported by others in obstetrics training,[16] in clinical clerkships,[17,18] and in a recent review of the literature related to workplace-based assessment more generally.[19] These studies do not focus solely on comments and consider broader aspects of the assessment process, but they do suggest that learners may not perceive a great deal of educational value in their assessments. To our knowledge, no studies have explored residents' understanding or perceptions of the written comment portion of such assessments. This is important because it may help us to understand why residents perceive little value, which may in turn offer guidance for faculty and instrument development. If residents are misinterpreting what their attending physicians are trying to convey, or discounting or devaluing the messages the comments contain, then aspects of validity are at stake and formative advantages that might be gained will be suboptimal.

To address this gap, our study sought to explore residents' interpretations of written assessment comments. The research questions we addressed include: (i) Can PGY2 residents rank-order PGY1 residents based on comments alone as reliably as faculty members? (ii) How do they make sense of what is written? (iii) What do they see as the purpose of ITERs, and do they believe they are useful? Towards these goals, our study used mixed methods, as described below, including a quantitative, experimental component and interviews conducted and analysed using a constructivist grounded theory approach.

METHODS

This study is part of a programme of research that uses a multi-phase, mixed-methods design,[12,13,20] involving multiple projects conducted over time and linked together by a common purpose: to better understand interpretations, uses and barriers inherent in using qualitative comments to inform trainee assessment.[21] Within each individual study, methodological choices are made with the pragmatic goal of using 'a method and philosophy that attempts to fit together the insights provided by qualitative and quantitative research into a workable solution'.[22,23] Our pragmatic approach therefore included an experimental protocol that generated rank-order data, analysed by quantitative methods, as well as interview transcripts, analysed using elements of constructivist grounded theory (as described in the Analysis section below).

The protocol used here replicates a study design used previously for faculty participants.[12] As in the faculty study, the materials consisted of all ITER comments collected over an entire academic year for 48 PGY1 residents in IM who graduated in 2011. There were 56 residents in that year, 48 of whom had eight or more ITERs with comments. A document that included the year's worth of comments was created for each resident. These documents were placed into 12 packages of documents for 15–16 residents such that each PGY1 resident appeared in four packages and no two packages were the same. We recruited PGY2 residents from our programme (n = 74) as participants by generating a randomised list and sending personalised invitations in batches of eight to 10 until we obtained 12 who consented (which occurred after 36 invitations had been sent). We chose PGY2s to ensure that participants each had a full year of experience of receiving ITERs in our system. This sample size was chosen because it enabled each PGY1's comments to be rank-ordered by four participants.

Each participant was given his or her package of ITER comments for 15–16 PGY1s, fully blinded and anonymised. They were asked to sort them into categories defined by previous research[11] (A = outstanding, excellent, exemplary; B = solid, safe, may need some fine tuning; C = borderline, bare minimum, remediable; D = unsafe, unacceptable, multiple deficits), after which they were asked to rank-order the comments within each category from best to worst. To answer the second and third research questions, participants were then interviewed by a research assistant (RA) who had experience with the protocol and who was unknown to the residents. Participants were asked to talk through how they made their rank-order decisions, how they interpreted the ITER comments, and to provide thoughts about the ITERs and assessment in general. Participants took a mean of 46.5 minutes (range: 32–65 minutes) to perform the task and to be interviewed. The interviews resulted in a total of 159 pages of transcripts for analysis.

All phases of this study were approved by the research ethics board at the Faculty of Medicine, University of Toronto.

Analysis

Rank-order data were analysed for reliability using urGENOVA (University of Iowa), with the data and study design being entered using G_String_IV Version 6.2 (University of Iowa), a software application that allows for straightforward coding of nested and unbalanced studies. We used an all-random, one-facet design with judges (our resident participants) nested within PGY1 resident being ranked. Spearman correlations were calculated using IBM SPSS Statistics for Windows Version 23 (IBM Corp., Armonk, NY, USA) between the rank-orders generated by resident judges in this study and rank-orders generated in the previous study conducted in faculty staff using the same dataset.[12]

The interviews were transcribed verbatim and analysed using a constructivist grounded theory approach. We began, of course, with knowledge of our previously developed framework, but it was only partially applicable to these new interview transcripts given the differences in participants' roles in assessment (i.e. faculty staff versus residents). The analysis was therefore conducted in an open, inductive way that allowed us to evolve a new framework for understanding residents' perceptions while being mindful of findings from similar research.[24] Sensitising concepts from the previously developed framework[13] included issues around subjectivity and fairness of assessments. The principal investigator (PI) and the RA who conducted the interviews read and discussed all transcripts as they were completed. The PI undertook primary open coding and developed a code book with definitions and examples, while keeping detailed memos. During this process, a second RA read and coded a subset of the transcripts based on these initial codes. The PI and RA met frequently to discuss the application and understanding of the codes using constant comparison and refining, merging or deleting codes as necessary until they reached consensus. The PI and RA then reviewed the key themes in depth to identify further nuances and sub-themes and to ensure theoretical sufficiency.[25] The final themes were discussed and debated with the entire research team. We returned to the transcripts repeatedly to resolve disagreements in interpretation and to respond to questions from team members regarding the themes. NVIVO Version 10 (QSR International Pty Ltd, Melbourne, Vic., Australia) was used to assist in the organising, analysis and coding of the data.

RESULTS

Reliability of rankings and correlations

The mean number of PGY2 resident rankers (represented as judges) per PGY1 comment set was

3.97. The G coefficient illustrating the extent to which the average of all rankers used the narrative comments to consistently differentiate between residents was 0.84. For a single ranker, inter-judge reliability was $G = 0.56$. This means that if the comments assigned to a PGY1 resident were to be assessed by a single ranker, the correlation between that ranker and another single ranker would be expected to be near $r = 0.56$. If a randomly selected set of four rankers were to be compared with another set of four rankers, this correlation would be expected to rise to 0.84. The variance components underlying these results are shown in Table 1.

Because the PGY1s forming the dataset in this study had been previously rank-ordered by faculty attendings,[12] we were able to calculate correlations between faculty and resident rankings of the same PGY1s, based on an average of four judges in each group. This correlation was extremely high at $r = 0.90$ ($p < 0.001$), indicating that faculty and residents ranked the same residents in largely the same order.

Analysis of interviews

During the interviews residents articulated the strategies they used to help them decide on how to rank a given PGY1. These strategies included looking for consistency, ascertaining the domains commented on, and the specificity and quantity of commentary, and examining contextual factors such as which attending physician was believed to have written the comment (given that author identity was not known to participants), on which rotation the comment had been delivered (which was sometimes embedded in the comment) and time of year (which was apparent from the chronological order of presentation). As these themes echo what has been previously reported by faculty attendings,[13] they will not be discussed in detail; rather, we will concentrate on the unique themes identified that reflect residents' own perspectives. These themes focus on the need for interpretation in order to construct meaning, the use of vague language and why it occurs, the relative value of scores versus comments, staff-dependent variability in assessment and residents' perceptions of the purpose of ITERs. It should be noted that residents' comments relate to both the ITERs read during the study task and to their own personal experiences with ITER comments.

Table 1 Reliability of the rank-ordering of postgraduate year 1 residents (PGY1) by PGY2 resident judges based on comments, with variance components.

Source of variance	PGY1 resident	PGY2 judge nested within PGY1 resident
Estimated variance component	0.054	0.042
Percentage of total variance	56.4	43.6
Reliability for a single judge	0.56	
Reliability based on average of four judges	0.84	

Interpretation to construct meaning

Our resident participants 'read between the lines' and took pains to explain how they interpreted what they read:

> So if they used "excellent" specifically. . ., I felt like it was very obvious what they were trying to say but there was a little bit of reading between the lines if they said "good" versus "very good". . . (R2)

> So many of them were generic: "excellent resident", "good enthusiasm", "happy to work with them", "pleasure to work with"; I mean that's easy enough to interpret that things are good but it doesn't really give me an idea of how the resident is doing. They're very nice and positive and I would say thank you if someone told me that, but it doesn't help me get better, so that's the only reason I have trouble interpreting it *per se*. I know what they mean but I don't know if they're not saying things. I don't know if they're avoiding saying things they want to say because they don't want to hurt the resident's feelings or what not. (R7)

Residents did not take language at face value; rather, they made interpretations and inferences. These inferences were influenced by their own personal experiences with ITERs and comments, as well as by rumours they heard from others in the IM programme:

> I know in my personal experience with these ITERs, there are certain rotations and certain staff that are known to rank lower, and that you

have to work harder to get that "exceeds expectations". (R12)

Vague language and why it occurs

Despite the high inter-ranker reliability observed, residents struggled with vague and non-specific language, as R7 explained above, and in the following:

> Most of the evaluations I found very vague in their wording as well. "Mature, sound clinical judgement" was one that came out a lot. Um, "hardworking and enthusiastic, good knowledge base", they say that about everyone; it doesn't really say anything. Another common thing people would say, um, "above expectations for this level of training". I would say three quarters of the pile was above expectations and so it makes me wonder what is "expectations"? Are the expectations too low because everyone is exceeding them? (R12)

Residents offered many potential reasons to explain why language might be vague. These rationales, grouped into several sub-themes, are described below.

Time and effort

One commonly offered rationale referred to the apparent time and effort required to write 'good' comments. Residents repeatedly used words such as 'busy' and 'time-consuming', and noted that it 'takes a dedicated effort' to provide more detailed comments. However, some also felt that this should not be regarded as an excuse:

> It's part of their job description, I think, to [[pause]] do this. And it's time-consuming; I'm not saying it's difficult – it's easy. (R1)

Emotions and discomfort

Residents also recognised that it could be emotionally difficult for their staff to give them meaningful feedback:

> Um, and it's also hard to sometimes write about weaknesses or stuff people can work on because it sort of instils conflict. (R3)

Some residents felt that attendings might err on the side of 'mak[ing] their residents feel good when they should be commenting on how to make them better doctors' (R7)

Not memorable

Many residents suspected that staff might not always remember their residents in great detail, especially if evaluations are delayed, thus resulting in generic comments:

> So I could imagine trying to fill these things out for people who were, again, pretty solid residents but nothing really stood out and then trying to do that maybe a couple of weeks later would be kind of difficult. And so, yeah, you probably would go back to … pretty generic statements. (R10)

Apart from the time lag, they also noted that the person filling out the ITER comments may not actually know the resident well because he or she had been given limited exposure:

> A lot of the time you get vague, unhelpful comments. By people who have known you for a day, or have never met you before. (R12)

Permanent record

Residents perceived that faculty attendings were concerned about the permanency of a written record. In many instances this was described as manifesting in a reluctance to record feedback in writing because this might affect the resident in the future:

> I think that even, for example, if staff had some feedback to give you… if they're a staff that really cares about your career, I don't think they would include it there, but rather tell you verbally… (R1)

> I think most negative comments would come face-to-face, which I think allows them to be expressed but wouldn't necessarily hurt any future applications. (R3)

Many residents spoke about the differences between verbal and written feedback, generally favouring face-to-face interactions. For example, one resident agreed that verbal feedback is better:

> You can ask questions, you can seek more feedback… if they say you need to work on something you can talk to them about [it] and sort of get the back-and-forth, where the ITERs [are] just written. (R12)

Others felt that it might be 'easier to say stuff to people in person' (R11) and that it would be less subject to misinterpretation:

> I think it's perhaps, the person writing it doesn't want things to be misinterpreted, that if there is a culture that – that's just not what's usually put on ITERS, that they wouldn't want for it to be taken the wrong way. (R11)

The notion that constructive critique is 'just not what's usually put on ITERs' emphasises residents' perceptions that putting something critical on a permanent record was largely 'not done'. Together, these sub-themes capture residents' perceptions of why vague language occurs.

Scores versus comments

Participants were asked about the relative value or importance of the numeric scores versus comments on the ITER. Nearly all of them expressed opinions favouring comments over scores. For example, comments were seen to be more useful, more meaningful and simply 'better'. They felt comments were more trustworthy than scores and more useful for providing feedback than numbers. Some felt they were easier to interpret and would be better at generating reflection. Opinions on scores varied: some felt they were more reliable and objective, whereas others felt they were just as – or even *more* – subjective than comments. Of interest, the terms 'subjective' and 'subjectivity' came up only a handful of times in the interviews and applied to both scores and comments. Numeric scores were thought to be useful in giving a quick sense of where a resident is and how he or she compares with others. They could be seen as being more useful for external purposes, such as in comparing across residents and flagging weak residents.

In many cases residents talked about complementary purposes, such as in using the comments for evaluation that is not already included in the scores or using them to flesh out high or low scores:

> I think you can't separate the two. It's artificial to do that. I think the written comments will basically provide context to the numerical rankings and vice versa, like, you need them both. (R5)

Each can provide context for the interpretation of the other. That said, several residents felt that scores should be replaced by comments as they regarded scores as useless and arbitrary:

> Maybe if it was more comment-based I think it would be helpful... but right now I just – I don't think it relays what it's supposed to relay... Just the difficulty interpreting a 3 versus a 4 versus a 5. (R2)

Staff-dependent variability

The next major theme encompassed instances in which residents spoke of variations between attendings, using the familiar phrase 'it's staff-dependent'. In particular, they noted differences in writing style, such as the degree of vagueness or specificity or use of adjectives and 'flowery language':

> I think sometimes it's like, depending on the evaluator, some people are just more descriptive. (R8)

Another noted that if the same resident were to be evaluated by two people:

> ...one person will put "great team player, hard-working" and another one will write three paragraphs about how fantastic they were. And it could just be one evaluator's, um, much less wordy... (R12)

Residents seem to accept that these differences were to be expected in relation to written comments in general:

> You have to understand the person that's evaluating, right? And I think if you have an understanding of what they usually [write] in terms of their written comments, I think you can get a better sense of what that person is trying to convey. (R6)

Consider as well the following:

> Obviously there's going to be variations from, uh, evaluators or physicians or supervisors... and some of them may, just by their own nature, be a little bit more verbose or more generic, which actually may end up affecting the overall evaluation. Like, for example, if you have a supervisor who tends to be a bit more generic, they may actually understate how good the resident is. Um, so it really is a lot of "luck of the draw". (R5)

The use of the word 'obviously' indicates that, from the resident's point of view, this is something that everyone knows, whereas the end phrase 'luck of the draw' further conveys the sense that this is just the way it is. This also applied to numeric scores as many residents noted differences in how different evaluators score:

> So some staff like to give everyone 3s, some staff like to give everyone 5s, um, so it's difficult to interpret and, um, the residents themselves have been cautioned in interpreting these, so, not to get too bummed out if you get all 3s from a specific staff when you're normally getting 4s and 5s, um, because it's. . . I think it is very staff-dependent. (R2)

Purpose of ITERs

We asked residents specifically about what they thought the purpose of the ITER to be as a way to elucidate their responses to the utility of comments. They thought ITERs had several potential purposes, including to help residents, to guide the programme and to meet accreditation standards:

> I mean it is obviously for evaluation and tracking progress but I think it's also great in terms of feedback and that's why I think people rely on the comments more than the numbers themselves. (R6)

Many residents felt the ITERs were – or could be – very helpful, depending on how they were conducted:

> So I think ITERs in general have the potential to be fantastic, and I mean, I had very similar things in medical school at U of T, and overall I find them very useful. (R12)

In relation to the written comments, particularly, residents also felt that specific comments were best, not just because they helped residents know where to improve, but also because they felt more personal and implied that the attending 'cares more about that person' (R10). Although vague comments could at times be frustrating to decipher, they were not always seen negatively. Some residents felt that they might still have a purpose:

> When there is useful feedback, constructive criticism, that's also really useful as well, and the rest – it's kind of nice, I'd say, just to hear you're doing a good job. (R10)

Interestingly, the opportunity to read comments delivered over an entire year made many residents realise how useful ITERs might be in aggregate:

> I think that, um, on any one rotation it means nothing. I think. . . it would be good to actually see [comments] for the entire year all together. (R11)

Similarly, reflecting on his or her own experience, one resident said:

> Personally I find that when I get the sort of longer-term ITERs, the 6-month ones, I can actually see a trend of how I am doing. I find those more helpful. (R12)

According to our participants, whereas a single rotation's comment or score might feel like an outlier, seeing a compilation of comments can give a good sense of how a resident has been doing over time. This allowed residents to see the trajectory of performance, including areas in which the resident seemed to improve and those in which performance may have remained problematic:

> In this case it was maybe a couple of things in the beginning that weren't really commented on at the end so I assumed the person [had] improved. (R10)

DISCUSSION

There has been ongoing debate about the value of ITERs as assessment instruments[1] and about the value of workplace-based assessment more generally.[17] Prior research found that faculty attendings can reliably interpret comments collected from these forms of assessment. This provided necessary validity evidence for using ITER comments for assessment,[14] but whether or not residents would understand this apparently 'hidden code' in the often vague written language of faculty attendings remained unknown. If residents misinterpret or devalue the messages conveyed in the comments they receive, the educational value of the ITER process is potentially limited. We found, in response to our first research question, that IM residents were able to interpret comments with a high degree of reliability and in much the same way as faculty attendings,[12] suggesting that they are able to 'crack the code' extremely well. This is despite the challenges they expressed in interpreting vague

comments and making inferences and interpretations to construct meaning from the language.

Our findings can support and extend those of previous researchers in several ways. The replicability of our quantitative and qualitative findings across different groups suggests that ITER comments are indeed meaningful and potentially useful sources of assessment data provided information is triangulated across multiple judgements.[12,13] Similar to other studies with learners, our participants also valued credibility and engagement, which they often expressed as 'exposure' to the attending physician filling out the assessment.[16–18,26] Residents' recurrent discussion of the importance and value they place on verbal, face-to-face feedback discussions reflects research highlighting the importance of a dialogue between the assessor and the learner.[27,28]

One unexpected finding in our study is that residents not only recognised that assessments can be quite 'staff-dependent', but that they seemed to accept this fairly unproblematically. This is surprising because such idiosyncrasy of assessment is often considered to be unfair, although recent work has begun to question whether a lack of consistency between assessors should be interpreted as error or rather as meaningful variability.[29] We therefore expected our resident participants to express more frustration over staff variability, or to take the opportunity to complain about unfairness, but the overall tone of their comments was more neutral. On some level, residents seemed to accept that this is the reality – attendings are different, and have different writing styles, personalities and expectations – and residents seem prepared to interpret commentary in context, both in the study and in real life. When they discussed staff variability, they did so in a matter-of-fact way: they would mention it and then move on, rather than persevering or complaining about it.

There are several potential interpretations of this finding. One possibility is that our participants' nonchalance stems from the nature of the task: they were not reviewing their own ITERs but, rather, those of depersonalised residents who had graduated years previously and thus the inconsistency held no personal meaning or consequence. By contrast, many of their comments about variability did relate to their own ITER experiences, which suggests that there is at least some element of acceptance of variability in practice. Another possibility is that the ability to examine an entire year's worth of comments allowed the resident to take a bird's eye view and to put outliers in context, by contrast with how they usually receive evaluations, which occurs for one rotation at a time. This interpretation is supported by comments made by some residents about how much more useful their own assessments are when seen at the 6-month or 1-year mark.

In this regard, our findings resonate with the medical education community's emerging appreciation for expert subjectivity and collective assessment. In a paper provocatively subtitled 'Learning to love the subjective and collective', Hodges[4] reminds us that 'subjective' should not be equated with 'biased' and indeed summarises from Surowiecki[30] the premise that 'many fallible judgements, summed together, create value'.[4] Gingerich *et al.*,[31] drawing on work from Yeates *et al.*,[32] suggest that variability between assessors may be a result of assessors having 'legitimate but different, and sometimes conflicting, interpretations of the same observations'.[31] Govaerts *et al.*[33] and van der Vleuten[34] have suggested that assessors should not be thought of as 'perfectly calibrated measurement instruments but [as] active agents constructing judgements'.[34] In this view, different assessors are not expected to make similar judgements and variability may actually be desirable. This emerging literature reflects exciting new ways of thinking about assessment, based on solid theoretical underpinnings, as well as empirical research.

It is important to acknowledge, however, that our participants' apparent acceptance (or at least tolerance) of attending physician variability does not mean residents necessarily *value* it. The question of the value of variability was not raised in the interviews, nor did it arise spontaneously, and hence we must be cautious in our interpretation. Residents' acceptance of some variability is an intriguing finding that can be explored further in future studies. This is important not just for educators and assessment researchers, but also because it more closely reflects the realities of practice: not all of our patients will see us in the same way and as professionals we need to learn to accept (and learn from) 'legitimate but different' opinions of ourselves held by others.

Our findings have immediate and practical implications. Given that residents saw value in examining aggregated ITER comments, it seems that a simple intervention might involve the provision of more regular opportunities for residents to view their

comments en masse. This should complement rather than replace the current practice of viewing assessments that come month by month, which allows for timely action on any deficiencies noted. Viewing of aggregate data is supported by the literature on programmatic assessment, which emphasises the use of multiple methods, assessment for learning, and qualitative information that relies on human judgement.[35,36] Further, the apparent value that residents place on written comments – even when they are vague – suggests, as others have argued,[3,26] that more commentary would be welcome. In addition, the importance of a face-to-face dialogue at the time of ITER assessments cannot be overstated. Having faculty attendings write brief assessments every 1–2 weeks would have the advantages of enhancing opportunities for feedback and allowing for more aggregated comments to be seen at each time-point, and might go a long way towards overcoming the sense of disengagement that results from frequent staff turnover. Although faculty members may resist the extra work, this reluctance can be mitigated by educating them about the great value to be obtained by residents; further, if the assessments are considered as formative, the task may not appear to be as threatening.

Interpretations of our findings should consider several limitations. Our participants were volunteers and their views may not reflect those of all IM residents. Likewise, we can comment on perceptions of utility only in a single, large IM programme. Further studies are required to assess transferability to other contexts. The RA who conducted the interviews was unknown to participants; however, SG is a faculty member and attending physician within the same department and this may have had an influence on who chose to accept or decline the invitation to participate, and may have affected the issues discussed during the interviews.

CONCLUSIONS

Our findings add to the growing evidence supporting the use of narrative comments and subjectivity in assessment.[2,4,14] Residents in this study understood the hidden code in written comments, rank-ordered trainees reliably from them, perceived value in aggregate comment data, and accepted staff-dependent variability as a reality. Future research might explore the issue of whether residents not only tolerate but also value differing opinions of their performance.

Contributors: SG conceived the study, acquired, analysed and interpreted the data, and drafted the work. CvdV, KE and LL made substantial contributions to the study design, the analysis and interpretation of data, and the critical revision of the paper. All authors approved the final manuscript for publication and have agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of it are appropriately investigated and resolved.
Acknowledgements: We gratefully acknowledge the contributions of our research assistant, Lisa St. Amant. We also thank our resident participants for generously contributing their time and expertise.
Funding: This study was funded by the National Board of Medical Examiners Stemmler Fund for Research in Education.
Conflicts of interest: None.
Ethical approval: This study was approved by the Research Ethics Board, University of Toronto.

REFERENCES

1 Schuwirth L, van der Vleuten C. Merging views on assessment. *Med Educ* 2004;**38** (12):1208–10.
2 Govaerts M, van der Vleuten CPM. Validity in work-based assessment: expanding our horizons. *Med Educ* 2013;**47** (12):1164–74.
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 Hodges BD. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach* 2013;**35** (7):564–8.
5 Littlefield JH, Darosa D, Paukert J, Williams RG, Klamen DL, Schoolfield JD. Improving resident performance assessment data: numeric precision and narrative specificity. *Acad Med* 2005;**80** (5):489–95.
6 van der Leeuw RM, Schipper MP, Heineman MJ. Lombarts KMJMH. Residents' narrative feedback on teaching performance of clinical teachers: analysis of the content and phrasing of suggestions for improvement. *Postgrad Med J* 2016;**92** (1085):145–51.
7 Cohen G, Blumberg P, Ryan N, Sullivan P. Do final grades reflect written qualitative evaluations of student performance? *Teach Learn Med* 1993;**5** (1):10–5.
8 Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ* 2008;**42** (8):816–22.
9 Lye PS, Biernat KA, Bragg DS, Simpson DE. A pleasure to work with: an analysis of written comments on student evaluations. *Ambul Pediatr* 2001;**1** (3):128–31.
10 Holmes AV, Peltier CB, Hanson JL, Lopreiato JO. Writing medical student and resident performance evaluations: beyond 'performed as expected'. *Pediatrics* 2014;**133** (5):766–8.

11 Regehr G, Ginsburg S, Herold J, Hatala R, Eva KW, Oulanova O. Using 'standardised narratives' to explore new ways to represent faculty opinions of resident performance. *Acad Med* 2012;**87** (4):419–27.
12 Ginsburg S, Eva KW, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. *Acad Med* 2013;**88** (10):1539–44.
13 Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ* 2015;**49** (3):296–306.
14 Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med* 2016; **91** (10):1359–69.
15 Watling CJ, Kenyon CF, Zibrowski EM, Schulz V, Goldszmidt MA, Singh I, Maddocks HL, Lingard L. Rules of engagement: residents' perceptions of the in-training evaluation process. *Acad Med* 2008;**83** (10 Suppl):97–100.
16 Dijksterhuis MGK, Schuwirth LW, Braat DDM, Teunissen PW, Scheele F. A qualitative study on trainees' and supervisors' perceptions of assessment for learning in postgraduate medical education. *Med Teach* 2013;**35** (8):e1396–402.
17 Mazotti L, O'Brien B, Tong L, Hauer KE. Perceptions of evaluation in longitudinal versus traditional clerkships. *Med Educ* 2011;**45** (5):464–70.
18 Bates J, Konkin J, Suddards C, Dobson S, Pratt D. Student perceptions of assessment and feedback in longitudinal integrated clerkships. *Med Educ* 2013;**47** (4):362–74.
19 Massie J, Ali JM. Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. *Adv Health Sci Educ Theory Pract* 2016;**21** (2):455–73.
20 Ginsburg S, van der Vleuten CPM, Lingard L. Hedging to save face: a linguistic analysis of ITER comments. *Adv Health Sci Educ Theory Pract* 2016;**21** (1):175–88.
21 Creswell JW, Klassen AC, Plano VL, Smith KC. *Best Practices for Mixed Methods Research in the Health Sciences*. Bethesda, MD: Office of Behavioural and Social Sciences Research, National Institutes of Health 2011.
22 Johnson RB, Onwuegbuzie AJ. Mixed methods research: a research paradigm whose time has come. *Educ Res* 2004;**33** (7):14–26.
23 Morgan DL. Paradigms lost and pragmatism regained: methodological implications of combining qualitative and quantitative methods. *J Mix Methods Res* 2007;**1** (1):48–76.
24 Charmaz K. Coding in grounded theory practice. In: Charmaz K, ed. *Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis*. London: Sage Publications 2009;42–71.
25 Dey I. *Grounding Grounded Theory : Guidelines for Qualitative Inquiry*. San Diego, CA: Academic Press 1999.
26 Watling CJ, Lingard L. Toward meaningful evaluation of medical trainees: the influence of participants' perceptions of the process. *Adv Health Sci Educ Theory Pract* 2012;**17** (2):183–94.
27 Sargeant J, Mann KV, Sinclair D, van der Vleuten CPM, Metsemakers J. Understanding the influence of emotions and reflection upon multi-source feedback acceptance and use. *Adv Health Sci Educ Theory Pract* 2008;**13** (3):275–88.
28 Sargeant J, Eva KW, Armson H, Chesluk B, Dornan T, Holmboe E, Lockyer JM, Loney E, Mann KV, van der Vleuten CPM. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ* 2011;**45** (6):636–47.
29 Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgements: rethinking the aetiology of rater errors. *Acad Med* 2011;**86** (10 Suppl):1–7.
30 Surowiecki J. *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. New York, NY: Doubleday 2004.
31 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** (11):1055–68.
32 Yeates P, O'Neill P, Mann KV, Eva KW. Seeing the same thing differently: mechanisms that contribute to assessor differences in directly-observed performance assessments. *Adv Health Sci Educ Theory Pract* 2013;**18** (3):325–41.
33 Govaerts MJB, van der Vleuten CPM, 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** (2):239–60.
34 van der Vleuten CPM. When I say … context specificity. *Med Educ* 2014;**48** (3):234–5.
35 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ* 2005;**39** (3):309–17.
36 Schuwirth LW, van der Vleuten CPM. Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 2011;**33** (6):478–85.

Received 13 January 2016; editorial comments to author 26 February 2016, 25 April 2016; accepted for publication 18 July 2016