The Hidden Value of Narrative Comments for Assessment: A Quantitative Reliability Analysis of Qualitative Data

Shiphra Ginsburg, MD, MEd, PhD, Cees P.M. van der Vleuten, PhD, and Kevin W. Eva, PhD

Abstract

Purpose

In-training evaluation reports (ITERs) are ubiquitous in internal medicine (IM) residency. Written comments can provide a rich data source, yet are often overlooked. This study determined the reliability of using variable amounts of commentary to discriminate between residents.

Method

ITER comments from two cohorts of PGY-1s in IM at the University of Toronto (graduating 2010 and 2011; n = 46–48) were put into sets containing 15 to 16 residents. Parallel sets were created: one with comments from the full year and one with comments from only the first three assessments. Each set was rank-ordered by four internists external to the program between April 2014 and May 2015 (n = 24). Generalizability analyses and a decision study were performed.

Results

For the full year of comments, reliability coefficients averaged across four rankers were G = 0.85 and G = 0.91 for the two cohorts. For a single ranker, G = 0.60 and G = 0.73. Using only the first three assessments, reliabilities remained high at G = 0.66 and G = 0.60 for a single ranker. In a decision study, if two internists ranked the first three assessments, reliability would be G = 0.80 and G = 0.75 for the two cohorts.

Conclusions

Using written comments to discriminate between residents can be extremely reliable even after only several reports are collected. This suggests a way to identify residents early on who may require attention. These findings contribute evidence to support the validity argument for using qualitative data for assessment.

The assessment of competence in medical education is undergoing a significant transformation. Steps are being taken to prioritize outcome (or competency-based) models rather than time-based ones, which will necessitate a major shift in how we assess our trainees.[1] To ensure that residents meet predetermined milestones,[2] it is necessary to collect much more information on each trainee to support valid judgment and decision making.[3,4] To this end, increasing value is being placed on qualitative and subjective data[5] and on the need to aggregate data from multiple, low-stakes sources.[6] The format and variety of evaluations is expanding in step with these changes, but the profession generally still relies heavily on end-of-rotation assessment forms (herein called in-training evaluation reports, or ITERs) that contain numeric and narrative data. Most of the literature (and use) of forms of this type is based entirely on the numeric data, yet there may be great value in the narratives.[7–9] Researchers and educators have called for the medical education community to “expand our horizons” of assessment and go beyond numeric ratings to incorporate qualitative and other forms of data.[10]

Some research has been conducted on the utility and feasibility of using assessment comments to evaluate learners or practitioners, yielding mixed results. For example, several studies have found that comments are usually concordant with scores assigned, suggesting that reading thousands of comments (e.g., for physician revalidation[11] or residency[12]) may not be worth the trouble. On the other hand, areas of nonconcordance can illustrate weaknesses not otherwise picked up by the scores,[7,13] thereby helping to overcome the well-described phenomenon of “failure to fail,”[14,15] and comments provide learners with more guidance regarding how to improve.[16,17] Determining how to balance these competing issues of gaining additional information and maintaining feasibility is an important challenge for the health professions to address, not only in formal training environments but across the continuum of training and practice.

Recent work has found that comments can provide a highly reliable way to distinguish between internal medicine (IM) residents even in the absence of numeric scores.[7,18] Those analyses, however, used data from an entire year’s worth of ITER comments for each resident, and the comments were assessed by faculty or senior residents who worked in the same training program as the residents.[7,18,19] The high reliability observed, therefore, may be related to the volume of comments amassed (given that reliability can generally be expected to increase with the amount of data available in any assessment process) and to the faculty’s awareness of the culture of assessment within the particular training program studied. From a practical sense, waiting an entire year for evaluations to accumulate would severely limit the

S. Ginsburg is professor, Department of Medicine, and scientist, Wilson Centre for Research in Education, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada.

C.P.M. van der Vleuten is professor of education, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, the Netherlands.

K.W. Eva is associate director and senior scientist, Centre for Health Education Scholarship, and professor and director of educational research and scholarship, Faculty of Medicine, University of British Columbia, Vancouver, British Columbia, Canada.

Correspondence should be addressed to Shiphra Ginsburg, 433–600 University Ave., Toronto, ON M5G1X5, Canada; telephone: (416) 586-8671; e-mail: shiphra.ginsburg@utoronto.ca.

Acad Med. XXXX;XX:00–00.
First published online
doi: 10.1097/ACM.0000000000001669

usefulness of narrative assessment for early intervention. Being dependent on raters who are intimately familiar with context would similarly limit the capacity to use such assessments for a variety of purposes.

The overall goal of this study was to contribute to the validity argument regarding the use of narrative data in assessment. Under current models of validity, reliability is considered to be an important aspect of validity, helpful for building an argument regarding whether or not assessment scores are fit for a given purpose with a given population.[20,21] Our study focuses on the reliability with which residents can be rank-ordered given variable amounts of commentary about their performance. For this purpose, we used ITER comments from two cohorts of postgraduate year 1 (PGY-1) trainees in an IM program to determine the comparability of reliabilities achieved if the narrative data consist of only comments received early in the year relative to including a full year's sample of comments. To address the question of whether faculty have to be "insiders" to make sense of residents' assessment comments, we recruited faculty who did not work in the same program as the residents but, rather, were drawn from different institutions and universities at a national level.

Method

Setting

After receiving Research Ethics Board approval from the University of Toronto's Office of Research Ethics, we collated ITERs from two cohorts of IM residents in PGY-1 at the Faculty of Medicine, University of Toronto. Cohort 1 graduated in 2011, and cohort 2 graduated in 2010; the total number of residents in each year was 55 or 56. Each resident receives one ITER at the end of each one-month clinical rotation, over 93% of which contain written comments. Our ITERs contain 18 items, each rated on a scale from 1 to 5 followed by an overall rating and a single free-text box in which to enter comments. The instructions to faculty state: "Provide a general impression of the trainee's development during this rotation, including general competence, motivation, and consultant skills. Please emphasize strengths and areas that require improvement." See Box 1 for an example of comments from the first part of the year for one resident.

We included residents who had received 8 or more ITERs containing comments over the course of one year, 3 of which had to come from the first 4 months of training. We chose 8 based on studies showing acceptable reliability of ITER scores aggregated across this number of ITERs.[7,22,23] We randomly selected 48 residents from each cohort who met these criteria so that we could create sets of comments as follows. Each resident was included 4 times, so that they appeared in 4 different sets and could be ranked by 4 different participants; this resulted in 192 documents, with each document containing a year's worth of comments from a given resident. On the basis of previous research,[7] we determined that faculty could read and rank-order comments from 16 residents in a reasonable time frame; thus, 12 sets of 16 residents' comments were compiled. To avoid potentially confounding the data by inadvertently grouping higher- or lower-performing residents together, we ensured that no 2 sets were identical. It should be noted that the older cohort did not have 48 residents who met these criteria because there were fewer residents that year who had ITERs with enough comments over the full year and in the first 4 months. Therefore, we had to include 3 residents with fewer than 8 ITERs (2 residents had 7 ITERs; 1 resident had 6 ITERs).

Using the 12 sets of documents described above, which contained the residents' entire year's worth of comments, we then created a parallel set of documents that were identical except that they contained only the comments from the first 3 comment-containing ITERs of the year for each resident. Thus, we created 24 unique sets of documents per cohort (12 sets × full year or partial year).

Participants

We recruited 24 IM faculty from institutions across Canada between April 2014 and May 2015 by accessing publicly available directories on academic department of medicine Web sites and e-mailing study invitations. We also displayed recruitment notices at national medical education conferences and meetings, sent open invitations via Twitter, and encouraged word-of-mouth referrals. Potential participants had to have at least two years' experience teaching and evaluating residents on IM clinical teaching units. This level of experience was chosen to align with previous studies and to ensure that

Box 1

Examples of Three Rotations' ITER Comments for One PGY-1 Resident, From a Canadian Study of the Reliability of Commentary for Residents' Assessment, April 2014–May 2015

First three months of comments, Resident #1234

XXXXXXX is a very pleasant, conscientious, hard-working, and reliable housestaff whose knowledge in general internal medicine is excellent. XXXXXXX's clinical assessment was thorough and complete. For example, by taking a careful history and following up with appropriate testing XXXXXXX picked up a case of non-STEMI that was missed by the referring service. During this rotation, XXXXXXX's consultative skills have matured nicely and with further experience I expect that XXXXXXX will do very well. Finally, XXXXXXX related well to the health care team and XXXXXXX's patients. Overall, an excellent performance in a very busy and demanding service.

Dr. XXXXXXX was a pleasure to have on the nephrology service. XXXXXXX did well in the diagnosis and management of both acute and chronic kidney disease. XXXXXXX contributed well to rounds and teaching sessions.

—Great work ethic

—Seemed to end up with many admissions whenever on call and therefore had very large patient load—handled this very well

—Complete notes

—Thoughtful approach to patient issues

—Well done

Abbreviations: ITER indicates in-training evaluation report; PGY-1, postgraduate year 1.

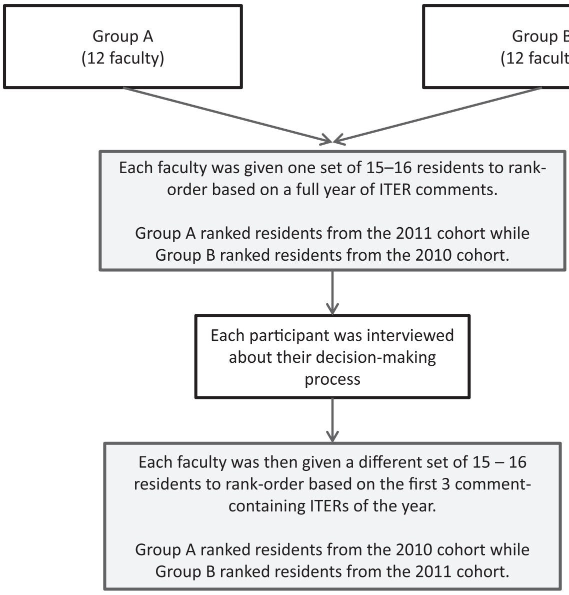
Group A
(12 faculty)
Group B
(12 faculty)
Each faculty was given one set of 15–16 residents to rank-order based on a full year of ITER comments.
Group A ranked residents from the 2011 cohort while Group B ranked residents from the 2010 cohort.
Each participant was interviewed about their decision-making process
Each faculty was then given a different set of 15 – 16 residents to rank-order based on the first 3 comment-containing ITERs of the year.
Group A ranked residents from the 2010 cohort while Group B ranked residents from the 2011 cohort.

Figure 1 Representation of study design, from a Canadian study of the reliability of using written comments for residents' assessment, April 2014–May 2015. Faculty were attending physicians from internal medicine from programs external to the University of Toronto. PGY-1 residents were internal medicine residents who graduated from the Faculty of Medicine, University of Toronto, in 2010 and 2011.
Abbreviations: ITER indicates in-training evaluation report, PGY-1, postgraduate year 1.

participants had reasonable familiarity with ITERs. There were no specific exclusion criteria. The study design is shown in Figure 1. Each consenting faculty member was sent a package containing two sets of data: One set contained the entire year's worth of comments for 15 to 16 residents in one of the two cohorts; the other set contained the first three comment boxes of the year from 15 to 16 different residents from the other cohort. Participants were offered a $100(Can) gift card for their time.

Protocol

A trained research assistant (RA) conducted a face-to-face meeting over Skype with each participant. Beginning with the set of full-year comments, participants were asked to read all 15 to 16 documents and sort them into categories derived during prior research (A = outstanding, excellent, exemplary; B = solid, safe, may need some fine tuning; C = borderline, bare minimum, remediable; D = unsafe, unacceptable, multiple deficits).[24] Afterwards, they were asked to rank-order the residents within each category, resulting in a final ranking of 1 to 15/16. After this task, they were interviewed by the RA to explore their decision-making process. They then repeated the task using a second set of documents, which contained a different set of residents' comments from the residents' first three assessments of the year. Time required for the full-year and part-year tasks was approximately 45 and 15–20 minutes, respectively, as recorded by the RA. In total, each resident's comments were expected to be ranked by four faculty within each condition (i.e., full-year vs. early comments).

Analysis

To analyze the effectiveness of generating judgments based on ITER comments, rank-order data from all 24 faculty participants were entered into Excel and verified for accuracy. We assessed the reliability of resident rankings using generalizability theory, with ranker nested within resident. G_string was used because it enables analysis when the study design is unbalanced (while most PGY-1s had four rankers, some had three because of inadvertent miscoding of one resident's data and because two packages were each missing comments for one resident). Finally, we calculated correlations between three-rotation and full-year data using SPSS statistical software, version 23 (IBM SPSS Inc., Armonk, New York).

Results

Our 24 participants were affiliated with 7 universities: 5 from the Cumming School of Medicine, University of Calgary; 4 each from the University of British Columbia, University of Ottawa, and Western University; 3 each from University of Alberta and McGill University; and 1 from McMaster University.

The 48 residents from the 2011 cohort were rank-ordered by an average of 3.97 faculty, and the 46 residents from the 2010 cohort were rank-ordered by an average of 3.94 faculty. For the full year of comments, reliability coefficients averaged across all rankers were G = 0.85 and G = 0.91 for the first and second cohort, respectively. The reliability coefficients examining the extent to which residents could be consistently differentiated on the basis of the rank-ordering provided by a single faculty member were G = 0.60 and 0.73 for the full year's worth of ITER comments. When rankings were done based on the document set that contained only the first three assessments for each resident, their reliability (for the 2011 and 2010 cohorts, respectively) was G = 0.89 and G = 0.85 when all four rankers were included and G = 0.66 and G = 0.60 using the rank-ordering provided by a single faculty member. A decision study outlining the influence of increasing the number of faculty rankers on the reliability of the rankings is illustrated in Table 1.

Spearman correlations between rankings based on the first three assessments and based on the full year were calculated for each cohort and were found to be $r = 0.66$ and 0.63, respectively, both significant with $P < .01$. These correlations are comparable to a similar set calculated on the ITER scores themselves, which were found to be $r = 0.76$ and 0.63, respectively, again both significant with $P < .01$. In all cases it should be acknowledged that the absolute value of these correlations may be spuriously inflated because the full-year documents included the first three assessments, thereby preventing us from examining the correlation between truly independent sets of rankings.

Discussion

Our findings reveal that using narrative comments alone as a means of assessing residents can be extremely reliable. This high reliability was maintained even when we considered only the first three

Table 1

Reliability of Ranking Two Cohorts of PGY-1 Residents From the University of Toronto (Graduating in 2011 and 2010) by Internal Medicine Attendings Recruited From Seven Programs Across Canada, From a Study of the Reliability of Written Comments for Residents' Assessment, April 2014–May 2015[a]

Rankings from comments	2011, full year		2011, first three rotations		2010, full year		2010, first three rotations	
	R	F:R	R	F:R	R	F:R	R	F:R
Source of variance								
Estimated variance component	0.057	0.039	0.064	0.032	0.071	0.026	0.057	0.039
Percentage of total variance	59.9	41.0	66.4	33.6	72.8	27.2	59.6	40.4
Rankings from comments	2011, full year		2011, first three rotations		2010, full year		2010, first three rotations	
Reliability								
Reliability for a single ranker	0.60		0.66		0.73		0.60	
Reliability based on average of two rankings	0.74		0.80		0.84		0.75	
Reliability based on average of three rankings	0.81		0.86		0.89		0.82	
Reliability based on average of four rankings	0.85		0.89		0.91		0.85	

Abbreviations: R indicates resident; F:R, faculty ranker nested within resident.

[a] Rankings were based on assessment comments from in-training evaluation reports from both one full year and from the first three assessments (rotations).

comment-containing ITERs of the year (see Table 1). In both cohorts studied, 85% to 91% of the variance in resident ranking was attributable to the resident (i.e., the "signal" in the measurement) when the average ranking across four rankers was considered. Further, residents' rankings from the first three ITERs were highly correlated with their rankings based on the full year of data (although it must be kept in mind that the full year's ranking included the ITER comments collected on the first three ITERs). Table 1 also illustrates that a reliability of 0.75 to 0.80 can be achieved with only two faculty members ranking residents based on three rotations' worth of comments. Such numbers are within the range of acceptability for even high-stakes assessments,[25] suggesting that a simple intervention—having two faculty read residents' evaluation comments early in the year—can be a very fruitful enterprise and may enable the identification of residents requiring additional educational supports at an early time point.[17] If feasibility is less of an issue, then further gains in reliability can be achieved by increasing the number of rankers, as illustrated in Table 1.

Unlike previous work, a unique feature of this study is that the faculty participants were external to our training program and were not trained in assessing ITER comments, although they were experienced in IM assessment. Previous research found that faculty belonging to the same program as the residents whose ITERs were being assessed were adept at "reading between the lines" to decode assessment comments that could often appear to be vague and lacking in specificity.[18,26] The fact that external, untrained faculty appear capable of reading between the lines just as readily implies that there is a degree of universality to how IM faculty write and understand narratives about their residents. This further suggests that there is a shared understanding on the part of faculty of what performance should look like for PGY-1s in IM, at least within a single country. This knowledge can help in the attempt to set expectations and standards for PGY-1s in evolving competency-based curricula.[1,27]

Our findings have broad relevance to other assessments that collect words as data, such as "field notes" in family medicine[28] or evaluation of teacher competence.[29] Further, they might help to facilitate the educational advantages of assessment processes that are strived for during the continuing professional development stage of practice, a context in which scores are often not helpful because of the narrow range and positive skew that is commonly reported. Before concluding in these regards, our findings would require replication in different contexts, but the reality that our comments were easily collected, fairly brief, and involved no special training on the part of the attendings makes it easy to envision numerous potential applications.

Several limitations should be kept in mind when interpreting our findings. The replicability of our work in other programs may be limited as all of our assessment comments came from a single, albeit large IM program that might have a specific culture of assessment regarding the extent and nature of comments and because our participants were required to have two years' worth of experience with ITERs. This potential raises an alternative explanation of the mechanisms that enabled our participants to "read between the lines" in that perhaps reading multiple resident assessment comments from a given program can allow readers to learn what "typical" language use is within that program, thereby allowing them to calibrate their rankings accordingly. We think this explanation is less likely, as the marked differences in writing style and content noted between attending physicians argue against the notion of "typical" language use.[26] An additional limitation comes from the fact that we used an open recruitment strategy, which prevents us from stating a response rate to our call for participation and from making claims about the representativeness of our sample compared with all academic internists. However, the geographic representation (including faculty from programs of all sizes at seven universities) speaks to the generalizability of our findings. We can also make no claims as to the transferability of our findings beyond PGY-1. Finally, it should be noted that although we presented only comments to our research participants, the comments were taken from actual ITERs that required those who wrote them to assign numeric scores to residents as well. It is possible, therefore, that the presence of rating scales influenced the generation and consistency of narrative comments, so we cannot be certain that these results would generalize to contexts in which only commentary is requested of examiners.

Conclusions

The incorporation of narrative comments as a routine part of assessment in medical education is overdue.[30] Our study adds to the growing validity evidence for the utility of narratives[21] by demonstrating that they can be reliably used as a way to discriminate between residents after a small number of reports are collected. This is particularly useful knowledge if one hopes to intervene quickly to assist residents in difficulty because, relative to ratings, comments have been reported to offer residents more informative guidance regarding what to do to improve.[16,31] From a practical point of view, most IM programs could probably implement a system in which the first three sets of ITER comments are assessed by one or two attendings. From a program perspective, narrative comments can provide insight into common areas of weakness that can then be addressed at a curricular level.

Importantly, these findings add to a growing literature[7,21,32] that should help to dispel the common opinion that ITERs are "useless" for assessment in IM, which might further reinforce the importance of writing rich and meaningful comments.

Acknowledgments: The authors are grateful to Ms. Lisa St. Amant for conducting and transcribing the interviews.

Funding/Support: This study was funded by the National Board of Medical Examiners' Edward J. Stemmler Fund for Research in Education.

Other disclosures: None reported.

Ethical approval: This study was approved by the Research Ethics Board of the University of Toronto in February 2014 (protocol #26409).

References

1 Royal College of Physicians and Surgeons of Canada. Competence by design (CBD). http://www.royalcollege.ca/rcsite/competence-design-e. Accessed February 1, 2017.

2 Iobst WF, Sherbino J, Cate OT, et al. Competency-based medical education in postgraduate medical education. Med Teach. 2010;32:651–656.

3 Caverzagie KJ, Iobst WF, Aagaard EM, et al. The internal medicine reporting milestones and the Next Accreditation System. Ann Intern Med. 2013;158:557–559.

4 Williams RG, Dunnington GL, Mellinger JD, Klamen DL. Placing constraints on the use of the ACGME milestones: A commentary on the limitations of global performance ratings. Acad Med. 2015;90:404–407.

5 Hodges B. Assessment in the post-psychometric era: Learning to love the subjective and collective. Med Teach. 2013;35:564–568.

6 Schuwirth LW, van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011;33:478–485.

7 Ginsburg S, Eva K, 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;88:1539–1544.

8 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.

9 Overeem K, Lombarts MJ, Arah OA, Klazinga NS, Grol RP, Wollersheim HC. Three methods of multi-source feedback compared: A plea for narrative comments and coworkers' perspectives. Med Teach. 2010;32:141–147.

10 Govaerts M, van der Vleuten CP. Validity in work-based assessment: Expanding our horizons. Med Educ. 2013;47:1164–1174.

11 Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. Med Educ. 2009;43:757–766.

12 Durning SJ, Hanson J, Gilliland W, McManigle JM, Waechter D, Pangaro LN. Using qualitative data from a program director's evaluation form as an outcome measurement for medical school. Mil Med. 2010;175:448–452.

13 Frohna A, Stern D. The nature of qualitative comments in evaluating professionalism. Med Educ. 2005;39:763–768.

14 Cleland JA, Knight LV, Rees CE, Tracey S, Bond CM. Is it me or is it them? Factors that influence the passing of underperforming students. Med Educ. 2008;42:800–809.

15 Dudek NL, Marks MB, Regehr G. Failure to fail: The perspectives of clinical supervisors. Acad Med. 2005;80(10 suppl):S84–S87.

16 Watling CJ, Kenyon CF, Zibrowski EM, et al. Rules of engagement: Residents' perceptions of the in-training evaluation process. Acad Med. 2008;83(10 suppl):S97–S100.

17 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.

18 Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: Faculty interpretations of narrative evaluation comments. Med Educ. 2015;49:296–306.

19 Ginsburg S, van der Vleuten CPM, Eva KW, Lingard L. Cracking the code: Residents' interpretations of written assessment comments [published online ahead of print January 16, 2017]. Med Educ. doi: 10.1111/medu.13158.

20 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.

21 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.

22 Carline JD, Paauw DS, Thiede KW, Ramsey PG. Factors affecting the reliability of ratings of students' clinical skills in a medicine clerkship. J Gen Intern Med. 1992;7:506–510.

23 Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. Acad Med. 1998;73:1294–1298.

24 Regehr G, Ginsburg S, Herold J, Hatala R, Eva K, Oulanova O. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. Acad Med. 2012;87:419–427.

25 van der Vleuten CP, Schuwirth LW. Assessing professional competence: From methods to programmes. Med Educ. 2005;39:309–317.

26 Ginsburg S, van der Vleuten C, Eva KW, Lingard L. Hedging to save face: A linguistic analysis of written comments on in-training evaluation reports. Adv Health Sci Educ Theory Pract. 2016;21:175–188.

27 Carraccio CL, Englander R. From Flexner to competencies: Reflections on a decade and the journey ahead. Acad Med. 2013;88:1067–1073.

28 Donoff MG. Field notes: Assisting achievement and documenting competence. Can Fam Physician. 2009;55:1260–1262, e100.

29 Myers KA, Zibrowski EM, Lingard L. A mixed-methods analysis of residents' written comments regarding their clinical supervisors. Acad Med. 2011;86(10 suppl):S21–S24.

30 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.

31 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.

32 Hatala R, Sawatsky AP, Dudek NL, Ginsburg S, Cook DA. Using in-training evaluation report (ITER) qualitative comments to assess medical students and residents: A systematic review [published online ahead of print December 6, 2016]. Acad Med. doi: 10.1097/ACM.0000000000001506.