The Quality of Written Feedback by Attendings of Internal Medicine Residents

Jeffrey L Jackson, MD MPH[1,3], Cynthia Kay, MD MSc[1], Wilkins C Jackson, HS[2], and Michael Frank, MD[1]

[1]C, GIM Section, Zablocki VA Medical Center (VAMC), Medical College of Wisconsin, Milwaukee, WI, USA; [2]University of Wisconsin, Milwaukee, WI, USA; [3], Milwaukee, WI, USA.

BACKGROUND: Attending evaluations are commonly used to evaluate residents.

OBJECTIVES: Evaluate the quality of written feedback of internal medicine residents.

DESIGN: Retrospective.

PARTICIPANTS: Internal medicine residents and faculty at the Medical College of Wisconsin from 2004 to 2012.

MAIN MEASURES: From monthly evaluations of residents by attendings, a randomly selected sample of 500 written comments by attendings were qualitatively coded and rated as high-, moderate-, or low-quality feedback by two independent coders with good inter-rater reliability (kappa: 0.94). Small group exercises with residents and attendings also coded the utterances as high, moderate, or low quality and developed criteria for this categorization. In-service examination scores were correlated with written feedback.

KEY RESULTS: There were 228 internal medicine residents who had 6,603 evaluations by 334 attendings. Among 500 randomly selected written comments, there were 2,056 unique utterances: 29 % were coded as nonspecific statements, 20 % were comments about resident personality, 16 % about patient care, 14 % interpersonal communication, 7 % medical knowledge, 6 % professionalism, and 4 % each on practice-based learning and systems-based practice. Based on criteria developed by group exercises, the majority of written comments were rated as moderate quality (65 %); 22 % were rated as high quality and 13 % as low quality. Attendings who provided high-quality feedback rated residents significantly lower in all six of the Accreditation Council for Graduate Medical Education (ACGME) competencies ($p < 0.0005$ for all), and had a greater range of scores. Negative comments on medical knowledge were associated with lower in-service examination scores.

CONCLUSIONS: Most attending written evaluation was of moderate or low quality. Attendings who provided high-quality feedback appeared to be more discriminating, providing significantly lower ratings of residents in all six ACGME core competencies, and across a greater range. Attendings' negative written comments on medical knowledge correlated with lower in-service training scores.

An important obligation of program directors and attendings in medical education programs is to provide feedback to their learners.[1–3] Feedback is "specific information about the comparison between a trainee's observed performance and a standard, given with the intent to improve trainee's performance,"[4] and is an essential component for the growth of trainees.[2,5] Unfortunately, despite considerable information on the subject, the quality of oral and written feedback is often low.[3] Previous studies have shown that feedback tends to be nonspecific, is not provided in a timely manner, and does not provide learners with sufficient information to improve their performance.[6–9] Residents and attendings frequently disagree on the quality and quantity of feedback provided,[10–15] with the result that feedback is commonly cited as needing improvement.[16,17]

Several studies have examined feedback. Frye and colleagues found that feedback varied widely in its organization, level of interaction, and depth.[18] Kogan found that feedback was complex, that there was considerable variability in feedback techniques, and that many factors affected how staff felt about delivering feedback.[19] Delva found that feedback was affected by four factors: learning culture, relationships, purpose of feedback, and emotional responses to feedback.[20] Ende found that feedback was often implicit and inferential rather than explicit, and consequently was frequently misunderstood by residents.[21] Several papers have provided opinions on improving feedback quality.[2,4,11,22,23] For example, Skeff characterized high-quality feedback as specific, emphasizing behavior, frequent, selective, timely, balanced, tailored to the learning climate, interactive, labeled as feedback, and resulting in an action plan for improving performance.[24] However, few studies have directly observed and evaluated feedback quality; most rely on resident and attending surveys of their opinions about the quality of feedback delivered. No previous study has developed criteria for assessing written feedback quality. The objectives of our study were to 1)

describe the characteristics of written feedback, 2) correlate written feedback with ratings of residents by their attendings and with scores on the in-service training examination, 3) develop criteria for assessing feedback quality, and 4) use that schema to rate the quality of written feedback.

METHODS

Subjects for this retrospective analysis were Medical College of Wisconsin (MCW) internal medicine residents, across all training levels, who completed residency from 2004 to 2012. Residents were evaluated at least monthly by their attendings as they moved through various inpatient and ambulatory rotations and at least semiannually by their continuity clinic preceptors. These evaluations rated resident performance in six domains (patient care, medical knowledge, interpersonal communication, professionalism, practice-based learning and improvement, and systems-based practice),[25] and were rated on a scale from 1 through 9, anchored as 1 (unsatisfactory), 5 (satisfactory), and 9 (superior). Attendings provided an "overall" rating of residents on a scale from 1 through 9, and were also asked to provide written comments on their residents. Five hundred attending evaluations that included written feedback were randomly selected from among the 6,603 available evaluations. Randomization was achieved by assigning each attending evaluation a unique number and then randomly selecting, without replacement, 500 numbers between 1 and 6,603 for inclusion. Randomization and all calculations were performed using STATA software (v. 13.1; StataCorp LP. College Station, TX, USA)

Among these 500 resident evaluations, attending written comments were coded independently by two coders (JLJ, CK) with good inter-rater reliability (ICC: 0.85). Each statement that provided a single feedback item was coded as a unique utterance. For example a statement that the "resident was reliable and very well organized" would be coded as two utterances (reliable, well organized). Utterances were secondarily coded, when possible, into one of the six ACGME core competencies (patient care, medical knowledge, interpersonal communication, professionalism, practice-based learning and improvement, and systems-based practice). Statements that were generic, such as "this was a good resident," were coded as nonspecific. Statements about personality characteristics, such as "X was enthusiastic," were coded as personality characteristics. Secondary coding included whether the utterance was positive, negative, or neutral.

We led a series of small group exercises of medicine residents and medicine attendings. Attending written feedback statements were de-identified and placed on 4 × 6 index cards. The groups were asked to sort the statements into three categories of high-, moderate-, and low-quality feedback, and to discuss these decisions aloud, including the criteria used to determine the rating. One of the group members served as secretary, keeping track of the criteria on a flip chart. Field

notes were recorded by at least two observers (JLJ, CK, or WJ). In addition, the sessions were audiotaped, and de-identified transcripts were reviewed to confirm our notes and that all quality characteristics mentioned had been captured. The attendees were not provided a list of potential feedback characteristics, but were asked to discuss each attending utterance and specify how they would label the feedback. At the end of the exercise, participants formally developed criteria that they used to rate the feedback as high, moderate, or low quality. All discussion group members provided informed consent and received no compensation for participation.

In addition to coding the transcripts for content, informed by the criteria proposed by the small groups, our two coders then coded the transcripts as high-, moderate-, or low-quality feedback (Table 1). Feedback that met none of these criteria were rated as low quality. Moderate-quality feedback met at least one quality criteria. To be considered of high quality, feedback had to meet two or more of the above-mentioned quality domains.

In-service training examinations were conducted each year during the study period, and we had at least one in-service training examination score for all residents. There was very high correlation between service examination scores.[26] In cases where more than one was present, we used the average score. We examined the relationships between in-service scores and the quality of feedback and between the polarity (positive, negative, neutral) of feedback in the seven domains and in-service examination scores using analysis of variance. We used quadratic kappas and intraclass correlation coefficients to assess inter-rater reliability between the different group classifications of the quality of feedback as well as the coders. This study was approved by our institution's institutional review board.

RESULTS

There were 228 internal medicine residents, with a total of 6,603 evaluations by 334 attendings; 1,387 (21 %) had no written feedback. Among 500 randomly selected written comments, there were 2,056 unique utterances (mean 2.9, range 1–

Table 1 Characteristics of Feedback and Quality Ratings Identified During Group Discussion

Written feedback characteristics	Balanced
	Specific
	Generates action plan
	Behavioral/not personal
	Pays attention to learning climate
	Quantifiable
	Objective
	Based on goals
	Helpful/constructive
	Modifiable traits
Quality rating schema	Low quality: meets none of above
	Moderate quality: feedback has at least one of above elements
	High quality: feedback has at least two of the above elements

8). The 500 randomly selected comments were equally distributed among the 8 years comprising the sample time frame (p=0.87) as well as among interns and second- and third-year residents (p=0.63). The majority of evaluations were from inpatient rotations (n=1,826, 88 %) and consultation rotations (n=148, 7 %); a smaller number (n=82, 4 %) were from continuity experiences. Continuity written feedback had slightly more utterances than inpatient or other ambulatory rotations (5.1 vs. 3.9 vs. 4.1, p=0.002).

Characteristics of Written Feedback

Of unique utterances, the most common type was nonspecific (29 %, n=600); 20 % (n=415) of the comments were about resident personality, 16 % (n=324) about patient care, 14 % (n=292) interpersonal communication, 7 % (n=146) medical knowledge, 6 % (n=117) professionalism, and 4 % each on practice-based learning (n=89) and systems-based practice (n = 73) (Table 2). The majority of written feedback comments were positive (n=1,813, 88 %); 8 % (n=155) were negative, and 4 % (n=88) were neutral (Table 3). Nonspecific comments and comments on a resident's attitude or personality were less likely to be negative than the other domains (nonspecific, OR: 0.22, 95 % CI: 0.13–0.39; attitude/personality, OR: 0.53, 95 % CI: 0.34–0.82). Three ACGME competencies were more likely to include negative comments: medical knowledge (OR: 3.5, 95 % CI: 2.2–5.6), practice-based learning (OR: 2.5, 95 % CI: 1.3–4.8), and systems-based practice (OR: 4.6, 95 % CI: 2.5–8.3).

The distribution of utterance types differed significantly among inpatient, ambulatory, and continuity experiences (p=0.001). Ambulatory preceptors were similar to inpatient preceptors except that they were less likely to comment on resident communication skills (OR: 0.42, 95 %: 0.22–0.80; Table 4). Continuity preceptors were less likely to comment on the resident's personality characteristics (OR: 0.26, 95 % CI: 0.12–0.56), and were more likely to make negative comments (OR: 2.8, 95 % CI: 1.2–4.3) and to comment on the resident's systems-based practice (OR: 2.3, 95 % CI: 1.1–4.9) and professionalism (OR: 2.0, 95 % CI: 1.2–3.4).

Small Group Feedback Quality Measures

We conducted 10 small group sessions, with a total of 31 participants; 12 were faculty and 19 were medicine residents.

Table 2 Characteristics of Written Feedback Provided by Internal Medicine Attendings

ACGME domains	
Patient care	324 (16 %)
Medical knowledge	146 (7 %)
Interpersonal communication	292 (14 %)
Professionalism	117 (6 %)
Systems-based practice	73 (4 %)
Practice-based learning	89 (4 %)
Nonspecific	600 (29 %)
Resident personality	415 (20 %)

Table 3 Written Feedback Characterized as Positive, Negative, or Neutral, by ACGME Competencies

ACGME competency	Positive	Negative	Neutral
Patient care (n=324)	285 (88 %)	34 (10 %)	5 (2 %)
Medical knowledge (n=146)	111 (76 %)	26 (18 %)	9 (6 %)
Interpersonal communication (n=292)	258 (88 %)	28 (10 %)	6 (2 %)
Professionalism (n=117)	114 (97 %)	2 (2 %)	1 (1 %)
Systems-based practice (n=73)	57 (78 %)	16 (22 %)	0 (0 %)
Practice-based learning (n=89)	77 (87 %)	12 (13 %)	0 (0 %)
Nonspecific (n=600)	525 (88 %)	14 (2 %)	61 (10 %)
Resident personality (n=415)	386 (93 %)	23 (6 %)	6 (1 %)

The small groups identified several characteristics of higher-quality written feedback, which included the following: quantifiable, specific, actionable, balanced, objective, based on goals, and behavioral/not personal (Table 1). The groups uniformly proposed that written feedback that included none of these characteristics should be rated as low quality, that feedback meeting at least one of these criteria was moderate, and that meeting more than one of these criteria was high-quality feedback. While all of the groups proposed the same criteria for judging feedback quality as low, moderate, or high, the inter-rater reliability among groups was low (quadratic kappa ranging from 0.22 to 0.28).

Feedback Quality

Two coders (JLJ, CK) independently applied these criteria, with good inter-rater reliability (quadratic kappa: 0.87). Based on the criteria, the majority of attendings' written comments were rated as moderate in quality (65 %, n=322); 22 % were rated as high quality (n=11,1) and 13 % low (n=65). None of the written feedback from continuity preceptors was rated as low quality, though rates of moderate- (61 %) and high-quality feedback (39 %) were similar to non-continuity rotations (p=0.36). There was a stepwise increase in the number of written comments as the feedback rating increased from low to moderate to high quality (average: 2.3 vs. 4.4 vs. 4.6, p <0.0001). Attendings who were rated as having high-quality written comments rated residents significantly lower and had greater

Table 4 Comparison of Written Feedback Between Continuity Primary Care and Non-continuity Preceptors

ACGME competency	Likelihood of comment from continuity preceptor (OR, 95 % CI)
Patient care	1.1 (0.71–2.1)
Medical knowledge	1.3 (0.69–2.6)
Interpersonal communication	0.74 (0.40–1.4)
Professionalism	2.0 (1.2–3.4)
Systems-based practice	2.3 (1.1–4.9)
Practice-based learning	0.61 (0.20–2.0)
Nonspecific (n=600)	1.2 (0.76–1.8)
Resident personality (n=415)	0.26 (0.12–0.56)

spread of ratings in all six of the ACGME competencies as well as on their overall performance (Table 5).

There was no relationship between in-service training examination scores and the quality (p=0.18) or polarity of feedback (positive, negative, neutral, p=0.32). However, residents who received negative attending comments regarding their knowledge had lower in-service training scores (53.6 vs. 57.5, p=0.009).

DISCUSSION

Attending written feedback was generally limited by several factors. First, 21 % of evaluations had no written comments at all. While the online evaluation system could require some kind of written comment, it is likely that attendings mandated to enter comments would not provide thoughtful or meaningful ones. Moreover, even when there were comments, only 22 % of evaluations were considered high quality. As might have been expected, the more comments that were provided, the more likely that the evaluation would meet criteria for meaningful feedback. While each evaluation had an average of four comments, the fact that only one-fifth had two or more meaningful comments (meeting criteria for high quality) suggests that most of the comments were not helpful.

Almost all comments were positive. Negative comments were mostly related to the medical knowledge, practice-based learning, and systems-based practice competencies. However, comments on practice-based learning and systems-based

Table 5 Relationship Between the Criteria-Based Quality of Written Feedback by Attending Physicians and the Mean and Spread of Their Numerical Ratings of Trainees

Domain	Quality of written feedback			p
	High	Moderate	Low	
Patient care				
Mean	6.6	7.6	7.6	< 0.00005
Range*	5	4	4	
Medical knowledge				
Mean	6.3	7.3	7.8	0.0001
Range	6	4	3	
Medical interviewing				
Mean	6.3	7.5	7.7	0.0004
Range	5	4	3	
Physical examination				
Mean	6.5	7.4	7.6	0.02
Range	4	4	3	
Communication				
Mean	6.6	7.9	7.9	<0.00005
Range	7	5	3	
Professionalism				
Mean	7.1	8.2	8.1	<0.00005
Range	5	5	2	
Systems-based practice				
Mean	6.5	7.6	7.5	<0.00005
Range	5	4	3	
Practice-based learning				
Mean	6.4	7.5	7.6	0.0001
Range	7	4	4	
Overall				
Mean	6.3	7.5	7.7	<0.00005
Range	5	4	3	

Range: maximum score minus minimum score

practice were rare (each only 4 % of the total) such that the benefit of these was quite limited. While it is difficult to correlate negative comments in these two competencies with outcomes, negative comments in the medical knowledge competency correlated with poorer scores on the in-training examination.

While our coders achieved very high reliability in coding utterances and applying the criteria to categorize written feedback quality as high, moderate, or low, our small groups had low inter-rater reliability. This is interesting given that all of the small groups came up with similar criteria for rating the quality of the feedback. Field notes indicate considerable a discrepancy between groups in determining when statements were sufficiently specific; some groups were more liberal and others stricter. A second area of disagreement was in categorizing statements as examples of providing actionable feedback.

Characteristics of higher-quality written feedback included being quantifiable, specific, actionable, balanced, objective, goal-based, and behavioral rather than personal. We found two characteristics in particular where faculty commonly fail when providing feedback: 29 % of comments were nonspecific, and another 20 % were based on the resident's personality rather than behavior-based. Addressing these two factors alone could significantly improve the quality in half of the feedback comments provided by faculty.

Several barriers to providing high-quality feedback have been identified in the literature. A common one is inadequate time to evaluate the resident. This could explain why there were no examples of low-quality feedback from continuity preceptors who are evaluating every 6 months based on a longer exposure period. Other barriers include concern about damaging the relationship with the resident and the tendency for negative feedback to elicit emotional responses.[3] A recent challenge is the "millennial generational issue," which suggests that the current generation of residents were raised in an environment in which their mentor feedback led them to feel that they were special, and they are consequently now poor at self-assessment[27] and lack the reflective skills to incorporate feedback.[28]

Some aspects of our work are similar to previous findings; studies have found that written comments are often sparse[29–31] and nonspecific,[8,32] and fail to distinguish among competence levels of residents.[33] In addition, resident evaluations commonly suffer from both grade inflation and range restriction.[34] Faculty who put the time and thought into providing more meaningful comments may also be more accurately assessing the performance level of the resident.

There are a few notable limitations to this study. First, it was at a single site involving a single specialty. While other studies have suggested that poor feedback is a common problem, generalizing our results to other specialties or sites should be done with caution. Secondly, we had in-training examination scores for all participants rather than the more important American Board of Internal Medicine (ABIM) scores, and

did not have other objective outcomes by the residents for comparison. However, we have previously shown that in-training exam scores correlate significantly with ABIM exam scores (reference the Acad Med paper).[26] Third, the inter-rater reliability among the groups for rating feedback was low. The groups were consistent in developing the characteristics comprising higher-quality feedback, but differed in their decisions whether specific statements met those criteria. Fortunately, our coders, trained to the same standard for determining when statements met criteria for higher-quality feedback (specific, balanced, actionable, etc.), had very good inter-rater reliability. Strengths of this study include the large number of evaluations that were analyzed, the use of discussion groups and standardized criteria for assessing quality, and the fact that the evaluations were completed before the study was planned, so that there is no Hawthorne effect of faculty filling out evaluations differently because they knew that they would be studied. A final limitation was that this study was based on the prior version of the ABIM/ACGME evaluation tool. We had previously shown that both the immediately two preceding versions of the medicine resident evaluation forms had poor validity and reliability.[35] Whether assessments based on the new ACGME Internal Medicine Milestones[36] will truly improve the evaluation process remains to be seen.

Most clinical teaching is performed by clinicians who have no formal training in medical education, and this is likely why there has been a lag in the translation of the considerable theoretical and practical knowledge regarding feedback to medical education settings.[3] Fortunately, studies have found that faculty development can modestly improve the quality of written and oral feedback.[8,32,37] Several specific recommendations emerge from this study that can help guide faculty development in providing feedback. First, faculty should understand the value of providing written comments that are multiple in number and scope. Second, comments should be specific, focusing on elements of the resident's performance in the assessed competencies, and not just generalized comments on the resident overall. Third, comments should address behaviors in the resident's performance, and not personality or personal characteristics. The use of specific incidents as examples may help in this regard. Fourth, feedback should be balanced, providing both positive comments to reinforce good behaviors and constructive comments with action items and goals to address deficiencies. Formal mechanisms for providing feedback such as field notes have been shown to improve feedback quality.[38] Interventions to improve feedback optimally need to occur at the individual, collective, and institutional cultural levels.[39] Further research should evaluate the effectiveness of specific interventions to improve the quality of feedback to residents, with the ultimate outcome of improved resident performance.

Corresponding Author: *Jeffrey L Jackson, MD MPH; , 5000 W National Ave, Milwaukee, WI 53295, USA (e-mail: jjackson@mcw.edu).*

REFERENCES

1. **Eisenberg JM.** Evaluating internists' clinical competence. J Gen Intern Med. 1989;4:139–43.
2. **Ende J.** Feedback in clinical medical education. JAMA. 1983;250(6):777–81.
3. **Anderson PA.** Giving feedback on clinical skills: are we starving our young? J Grad Med Educ. 2012;4:154–8.
4. **van der Ridder JMM, Stokking KM, McGaghie WC, ten Cate OT.** What is feedback in clinical education? Medical Education. 2008;42(2):189–97.
5. **Kluger AN, DeNisi A.** The effects of feedback intervention on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychol Bull. 1996;119:254–84.
6. **Berbano EP, Browning R, Pangaro L, Jackson JL.** The impact of the Stanford Faculty Development Program on ambulatory teaching behavior. J Gen Intern Med. 2006;21:430–4.
7. **Jackson JL, O'Malley PG, Salerno SM, Kroenke K.** The teacher and learner interactive assessment system (TeLIAS): a new tool to assess teaching behaviors in the ambulatory setting. Teach Learn Med. 2002;14:249–56.
8. **Salerno SM, O'Malley PG, Pangaro LN, Wheeler GA, Moores LK, Jackson JL.** Faculty development seminars based on the one-minute preceptor improve feedback in the ambulatory setting. J Gen Intern Med. 2002;17:779–87.
9. **Salerno SM, Jackson JL, O'Malley PG.** Interactive faculty development seminars improve the quality of written feedback in ambulatory teaching. J Gen Intern Med. 2003;18:831–4.
10. **Sender-Liberman A, Liberman M, Steinert Y, McLeod P, Meterissian S.** Surgery residents and attending surgeons have different perspectives of feedback. Med Teach. 2005;27(5):470–2.
11. **Archer JC.** State of the science in health professional education: effective feedback. Medical Education. 2010;44(1):101–8.
12. **Jensen AR, Wright AS, Kim S, Horvath KD, Calhoun KE.** Educational feedback in the operating room: a gap between resident and faculty perceptions. Am J Surg. 2012;204:248–55.
13. **Bing-You RG, Towbridge RL.** Why medical educators may be failing at feedback. JAMA. 2009;302(12):1330–1.
14. **Gil DH, Heins M, Jones PB.** Perceptions of medical school faculty members and students on clinical clerkship feedback. J Med Educ. 1984;59:856–64.
15. **Delva D, Sargeant J, MacLeod T.** Feedback: a perennial problem. Med Teach. 2011;33:861–2.
16. **Bahar-Ozvaris S, Aslan D, Sahin-Hodoglugil N, Sayek I.** A faculty development program evaluation: from needs assessment to long-term effects, of the teaching skills improvement program. Teach Learn Med. 2004;16:368–75.
17. **Moss HA, Derman PB, Clement RC.** Medical student perspective: working toward specific and actionable clinical clerkship feedback. Med Teach. 2012;34:665–7.
18. **Frye AW, Hollingsworth MA, Wymer A, Hinds MA.** Dimensions of feedback in clinical teaching: a descriptive study. Acad Med. 1996;71:S79–81.
19. **Kogan JR, Conforti LN, Bernabeo EC, Durning SJ, Hauer KE, Holmboe ES.** Faculty staff perceptions of feedback to residents after direct observation of clinical skills. Med Educ. 2012;46:201–15.
20. **Delva D, Sargeant J, Miller S, et al.** Encouraging residents to seek feedback. Med Teach. 2013;35:e1625–31.
21. **Ende J, Pomerantz A, Erickson F.** Preceptors' strategies for correcting residents in an ambulatory care medicine setting: a qualitative analysis. Acad Med. 1995;70:224–9.
22. **Cantillon P, Sargeant J.** Giving feedback in clinical settings. BMJ. 2008;337:a1961.

23. **Turnbull J, Gray J, MacFadyen J.** Improving in-training evaluation programs. J Gen Intern Med. 1998;13:317–23.

24. **Skeff KM, Stratos GA, Berman J, Bergen MR.** Improving clinical teaching. Evaluation of a national dissemination program. Arch Intern Med. 1992;152:1156–61.

25. ACGME Program requirements for graduate medical education in internal medicine. Accreditation council for graduate medical education . 7-1-2013. 12-22-2014.

26. **Kay C, Jackson JL, Frank M.** The relationship between internal medicine residency graduate performance on the ABIM certifying examination, yearly in-service training examinations, and the USMLE Step 1 Examination. Acad Med 2014.

27. **Davis DA, Mazmanian PE, Fordis M, Van HR, Thorpe KE, Perrier L.** Accuracy of physician self-assessment compared with observed measures of competence: a systematic review. JAMA. 2006;296:1094–1102.

28. **Mann K, Gordon J, MacLeod A.** Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ Theory Pract. 2009;14:595–621.

29. **Gray JD.** Global rating scales in residency education. Acad Med. 1996;71:S55–63.

30. **Haber RJ, Avins AL.** Do ratings on the American Board of Internal Medicine Resident Evaluation Form detect differences in clinical competence? J Gen Intern Med. 1994;9:140–5.

31. **Thompson WG, Lipkin M Jr, Gilbert DA, Guzzo RA, Roberson L.** Evaluating evaluation: assessment of the American Board of Internal Medicine Resident Evaluation Form. J Gen Intern Med. 1990;5:214–7.

32. **Berbano EP, Browning R, Pangaro L, Jackson JL.** The impact of the Stanford Faculty Development Program on ambulatory teaching behavior. J Gen Intern Med. 2006;21:430–4.

33. **Hawkins RE, Sumption KF, Gaglione MM, Holmboe ES.** The in-training examination in internal medicine: resident perceptions and lack of correlation between resident scores and faculty predictions of resident performance. Am J Med. 1999;106:206–10.

34. **Durning SJ, Pangaro LN, Lawrence LL, Waechter D, McManigle J, Jackson JL.** The feasibility, reliability, and validity of a program director's (supervisor's) evaluation form for medical school graduates. Acad Med. 2005;80:964–8.

35. **Durning SJ, Cation LJ, Jackson JL.** The reliability and validity of the American Board of Internal Medicine Monthly Evaluation Form. Acad Med. 2003;78:1175–82.

36. **Caverzagie KJ, Iobst WF, Aagaard EM, et al.** The internal medicine reporting milestones and the next accreditation system. Ann Intern Med. 2013;158:557–9.

37. **Holmboe ES, Fiebach NH, Galaty LA, Huot S.** Effectiveness of a focused educational intervention on resident evaluations from faculty a randomized controlled trial. J Gen Intern Med. 2001;16:427–34.

38. **Laughlin T, Brennan A, Brailovsky C.** Effect of field notes on confidence and perceived competence: survey of faculty and residents. Can Fam Physician. 2012;58:e352–6.

39. **Mann K, van der Vleuten C, Eva K, et al.** Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. Acad Med. 2011;86:1120–7.