Competencies and Feedback on Internal Medicine Residents' End-of-Rotation Assessments Over Time: Qualitative and Quantitative Analyses

Ara Tekian, PhD, MHPE, Yoon Soo Park, PhD, Sarette Tilton, Patrick F. Prunty, Eric Abasolo, Fred Zar, MD, and David A. Cook, MD, MHPE

Abstract

Purpose
To examine how qualitative narrative comments and quantitative ratings from end-of-rotation assessments change for a cohort of residents from entry to graduation, and explore associations between comments and ratings.

Method
The authors obtained end-of-rotation quantitative ratings and narrative comments for 1 cohort of internal medicine residents at the University of Illinois at Chicago College of Medicine from July 2013–June 2016. They inductively identified themes in comments, coded orientation (praising/critical) and relevance (specificity and actionability) of feedback, examined associations between codes and ratings, and evaluated changes in themes and ratings across years.

Results
Data comprised 1,869 assessments (828 comments) on 33 residents. Five themes aligned with ACGME competencies (interpersonal and communication skills, professionalism, medical knowledge, patient care, and systems-based practice), and 3 did not (personal attributes, summative judgment, and comparison to training level). Work ethic was the most frequent subtheme. Comments emphasized medical knowledge more in year 1 and focused more on autonomy, leadership, and teaching in later years. Most comments (714/828 [86%]) contained high praise, and 412/828 (50%) were very relevant. Average ratings correlated positively with orientation ($\beta = 0.46$, $P < .001$) and negatively with relevance ($\beta = -0.09$, $P = .01$). Ratings increased significantly with each training year (year 1, mean [standard deviation]: 5.31 [0.59]; year 2: 5.58 [0.47]; year 3: 5.86 [0.43]; $P < .001$).

Conclusions
Narrative comments address resident attributes beyond the ACGME competencies and change as residents progress. Lower quantitative ratings are associated with more specific and actionable feedback.

Defensible decisions in health professions training require robust data about trainees' performance.[1] In contrast to time-based curricular models, where learners may advance to subsequent stages of training at fixed time intervals, competency-based medical education rests on the assumption that learners must achieve a defined standard of performance before promotion. This, in turn, requires frequent and accurate competency assessments that provide information that completely represents the relevant competencies. Educators increasingly recognize the need for both quantitative and qualitative assessment data to present a complete picture of the trainee.[2–6] However, the unique contributions of quantitative and qualitative data, and how to integrate data from each modality into promotion decisions, remain incompletely understood.

As part of the competency-based medical education paradigm, the Accreditation Council for Graduate Medical Education (ACGME) has mandated that trainees' developmental progress be tracked over time (i.e., that it be longitudinally assessed).[7,8] One approach to accomplishing this involves the definition and periodic monitoring of specialty-specific training "milestones."[2,9] As of July 2013, all ACGME-accredited internal medicine postgraduate programs in the United States have appraised the progress of postgraduate physicians (residents) relative to 22 subcompetencies and have reported these data to the ACGME every 6 months. Preliminary evaluations in various specialties indicate that milestone assessments can discriminate between residents in different stages of training (i.e., residents in different cohorts),[10–15] but we found no longitudinal studies examining milestone progression within a single cohort of residents. To examine how milestone levels change for the same learners over the course of their training, a cohort analysis from entry to graduation is necessary.

Resident assessment commonly involves workplace-based assessments, and among these, the most common may be the end-of-rotation assessment (known as the in-training evaluation report in Canada).[16–18] For this assessment, the rater (e.g., a supervising staff physician, fellow, or another resident) documents impressions acquired while working with the resident in a clinical setting over a period of time (usually a week or longer), typically using several items with numeric response options and at least 1 free-text

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

Correspondence should be addressed to Ara Tekian, Department of Medical Education, College of Medicine, University of Illinois at Chicago, 808 S. Wood St., 963 CMET (MC 591), Chicago, IL 60612-7309; telephone: (312) 996-8438; email: tekian@uic.edu.

Acad Med. 2019;94:1961–1969.
First published online June 4, 2019
doi: 10.1097/ACM.0000000000002821

Supplemental digital content for this article is available at http://links.lww.com/ACADMED/A691.

item allowing for narrative comments (i.e., qualitative assessments). A recent systematic review found 22 studies evaluating the validity of end-of-rotation assessment narrative comments.[18] These and subsequent studies have found that most narrative comments have a positive (praising) orientation[19–22]—that is, that positive comments are associated with superior performance on quantitative ratings or satisfactory/remediation classifications[20–24]—and that higher-quality comments (i.e., comments that are more actionable or longer) are associated with lower performance.[21,22] One study found that themes inductively identified from narrative comments map reasonably well to ACGME competencies,[22] but this and other studies have also identified other constructs reflected in narrative comments, including personality, motivation, and normative and summative judgments.[20,22,25–28] However, it remains unknown how the themes of narrative comments vary as residents progress through training. Additionally, we found no studies examining the relationship between the orientation and quality (or relevance) of comments and quantitative ratings or how both comments and ratings might vary over time.

In sum, the relationship between qualitative and quantitative end-of-rotation assessments, and how these vary over training, remains incompletely understood. Such relationships have pragmatic implications for the unique contributions of qualitative and quantitative data, how these 2 types of data can be meaningfully integrated into promotion decisions, and the development of trainees over time. In the present study, we sought to answer the following questions:

1. How do qualitative themes (i.e., from narrative comments) and quantitative ratings from internal medicine residents' end-of-rotation assessments change over the course of residency training?
2. How are features of narrative feedback (orientation [praising or critical] and relevance [specificity and actionability]) associated with quantitative ratings, and how do these associations change over the course of training and across competency domains?

As a secondary aim, we collected validity evidence on the use of end-of-rotation assessments to provide meaningful formative feedback and to inform promotion decisions for individual residents. Previously reported validity evidence for quantitative end-of-rotation assessments shows generally acceptable internal structure (i.e., high internal consistency reliability and variable interrater agreement) and relations with other variables (i.e., higher scores for more advanced trainees).[8,10–15] The concept of validation has recently been extended to qualitative (narrative) assessments,[6] and a systematic review of the validity of end-of-rotation narrative comments has been published.[18] To support the proposed uses for quantitative ratings, we would expect to find high reproducibility (internal structure), that ratings using developmental (or milestone-oriented) anchors increase over the course of training (relations with other variables), and that ratings correlate positively with quantitative coding of narrative comments (relations with other variables). To support the proposed uses for narrative (qualitative) assessments, we would expect the codes found to derive from a wide variety of assessors and rotations (content evidence), that inductively identified themes would align with desired competencies (content evidence), that change across training years would reflect evolving resident roles and competencies (relations with other variables), and that favorable (praising) narratives would correlate positively with quantitative ratings (relations with other variables).

Method

Overview

We collected quantitative and qualitative data from the end-of-rotation assessments of internal medicine residents over their 3 years of training. We identified themes represented in narrative comments and examined how these changed over the course of training. We also rated the orientation and relevance of the feedback represented in narrative comments and examined the association between these ratings of the feedback and the quantitative end-of-rotation ratings. Finally, we evaluated the reproducibility and change over time of quantitative ratings to gather validity evidence supporting internal structure and relations to other variables.

The University of Illinois at Chicago Institutional Review Board approved this study.

Data collection and assessment form

We obtained data from all end-of-rotation assessments for 1 cohort of internal medicine residents at the University of Illinois at Chicago College of Medicine over the course of their 3-year residency, July 2013–June 2016. Residents completed 13 four-week rotations per year (39 rotations total), including rotations in general medicine, medical specialties, and electives. Supervising physicians, fellows, and other residents rated each resident after every rotation, and faculty leaders met individually with each resident every 6 months to review their progress.

The end-of-rotation assessment form had 22 items, with language corresponding to the 6 ACGME core competencies (and their 22 subcompetencies) of patient care (5 items), medical knowledge (2 items), systems-based practice (4 items), practice-based learning and improvement (4 items), professionalism (4 items), and interpersonal and communication skills (3 items). Each item was rated on a 7-point scale that corresponded to the milestone language, with anchors of 1 = "critical deficiencies" and 7 = "ready for unsupervised practice" and 2 to 5 item-specific developmental milestones listed for response levels 1, 3, 5, and 7 (e.g., 4 milestones were listed for patient care item 1 response level 5, including "Consistently acquires accurate and relevant histories from patients"). Each item also contained a free-text box labeled "Comment." During postgraduate year (PGY) 2, quantitative ratings for systems-based practice, interpersonal and communication skills, and practice-based learning and improvement were omitted from the assessment form because these were measured in other training situations. Previous research has provided evidence supporting the internal structure (acceptable reliability and 6-factor domain structure), relations with other variables (rising scores across years and correlations with other measures), and response process (variation in scores across competency domains) for this instrument's scores.[8] Forms

were completed online using the New Innovations platform (New Innovations, Uniontown, Ohio). Faculty members with incomplete forms were sent weekly email reminders.

Data analysis

Quantitative coding of narrative comments. To facilitate correlation with quantitative data, we coded all narrative comments for 2 dimensions related to the nature of feedback: (1) orientation (i.e., whether the comment was praising or critical) and (2) relevance (i.e., the specificity of comments and whether actionable suggestions were included). For orientation, we used a 4-point bipolar scale with anchors of "very critical," "critical," "modest praise," and "high praise," and for relevance, we used a 4-point bipolar scale with anchors of "very irrelevant," "irrelevant," "relevant," and "very relevant." Two authors (P.F.P., S.T.), trained by a qualitative methods researcher, developed a coding rubric (see Supplemental Digital Appendix 1 at http://links.lww.com/ACADMED/A691) using a random sample of 25 comments. These 2 analysts then independently analyzed all narrative comments with support provided as needed from a clinical medical educator (F.Z.) and a nonclinical medical educator (A.T.). Multiple subthemes commonly emerged from a single comment, but each subtheme was counted only once per comment (i.e., once per assessment form). For example, a comment such as "hard working, thorough, and compassionate physician" would be coded once each for the subthemes of "compassion" and "work ethic." We did not have internal medicine clinicians code the data to avoid bias from reviewing data of residents who were known by the analysts. Interrater agreement was substantial for orientation and moderate for relevance (weighted kappa of 0.79 and 0.46, respectively). The analysts discussed and came to a consensus on all codes. To examine the possibility that some comments included both positive and negative elements, the full author group reanalyzed all comments searching for such mixed orientations. We found only 67/828 (8%) comments with mixed orientation, and after discussion, we satisfactorily classified all of these comments as reflecting a single orientation.

Qualitative data analysis. Two authors (P.F.P., S.T.) inductively identified themes reflected in the narrative comments by analyzing all comments using the constant comparative analysis method.[29,30] Each analyst first independently and inductively identified themes reflected in the narratives. They then iteratively and collaboratively compared and refined these themes and subthemes until they had developed a parsimonious rubric that fully represented the narrative comments (thematic saturation). The analysts used 481 comments to inductively develop this rubric and then proceeded to review and classify all comments. As a member check, the themes were confirmed by the residency program director and associate program directors. We counted the frequency of each theme and subtheme and contrasted these counts across training years.

Quantitative data analysis. To examine descriptive statistics, we calculated the average quantitative rating on all assessment form items for each competency and PGY. In addition, to incorporate the nested data structure, we used mixed-effects regression, accounting for clustering of learners, to examine the longitudinal trends across training years for average overall and competency-specific ratings and to contrast ratings across competencies and years. We used a similar approach to evaluate the association between quantitative ratings and the feedback codes derived from the narrative comments. We conducted a [rater (r): person (p)] × [subcompetency (s): competency (c)] generalizability study to estimate variance components in quantitative ratings and used these to calculate reliability (Φ coefficient index of dependability).[17,31] P values less than .05 were considered statistically significant. Data compilation and analyses were conducted using Stata 14 (StataCorp, College Station, Texas).

Results

We obtained 1,869 end-of-rotation assessments on 33 internal medicine residents over 3 years (July 2013–June 2016); 1 resident withdrew during the second year, so the cohort size became 32 in PGY2 and PGY3. The assessments were completed by 408 different raters (134 faculty, 90 fellows, and 184 peer trainees [categorical internal medicine residents and rotating residents from other specialties]). Of these 1,869 assessments, 828 (44%) had narrative comments. The number of assessments differed by training year, with a mean of 27.4 (standard deviation [SD] = 2.9), 8.4 (SD = 2.3), and 21.8 (SD = 5.1) assessments per learner and a mean of 11.9 (SD = 2.9), 5.7 (SD = 1.6), and 9.3 (SD = 2.8) assessments containing narrative comments in PGY1, PGY2, and PGY3, respectively. In any given year, about half of the residents did not receive any narrative comments on any assessment.

The dependability (Φ coefficient) for 15 assessments, across all rating form items, was 0.71, 0.62, and 0.70 for PGY1, PGY2, and PGY3, respectively. For PGY2, 21 assessments would be needed to reach a Φ coefficient > 0.70.

Themes embodied in narrative comments and how these changed over time

Eight general themes emerged from the narrative comments. Five of these themes aligned with ACGME competencies (interpersonal and communication skills, professionalism, medical knowledge, patient care, and systems-based practice), and 3 did not (personal attributes, summative judgment, and comparison to level of training). Within these themes, we identified 43 unique specific classification codes (or subthemes; see Table 1). The competency of practice-based learning and improvement did not emerge from our inductive analysis.

Across all 3 years, comments related to personal attributes (n = 126) and interpersonal and communication skills (n = 112) had the highest frequency, while systems-based practice (n = 6) had the lowest. Narrative comments from PGY1 contained the highest frequency of themes (n = 232) and unique subthemes (n = 34), while comments from PGY2 contained the lowest frequency of themes (n = 100), and those from PGY3 had the fewest subthemes (n = 18; Table 1).

Residents' work ethic. The most prevalent theme in the qualitative analysis, present across all training years (usually as praise), was the subtheme of work ethic.

> Very hardworking trainee, she stays late to follow up on patient labs and is very good about getting her notes done and is always present for [morning] report. (Faculty assessment of PGY1 resident)

> He showed dedication to learning about [infectious disease] while on this rotation through his clinical participation, reading during less busy times and participating in our "lectures" that I provided. (Faculty assessment of PGY1 resident)

Table 1

Themes[a] Derived From Narrative Comments About Internal Medicine Residents at the University of Illinois at Chicago College of Medicine, 2013–2016

	PGY1		PGY2		PGY3		Theme frequency across all years, count
Theme	Subthemes	Theme frequency, count	Subthemes	Theme frequency, count	Subthemes	Theme frequency, count	
ACGME competency-related themes							
Interpersonal and communication skills	1. Communication skills (general) 2. Presentation skills 3. Rapport with patients or caregivers 4. Writing skills	26	1. Function as a role model 2. Delegation of tasks 3. Leadership skills 4. Rapport with patients or caregivers 5. Teaching skills	43	1. Function as a role model 2. Leadership skills 3. Teaching skills	43	112
Professionalism	1. Compassion 2. Ethical judgment 3. Level of professionalism 4. Punctuality and attendance 5. Self-awareness for improvement 6. Team player	33	1. Compassion 2. Level of professionalism 3. Team player	13	1. Approachability 2. Compassion 3. Level of professionalism 4. Team player	27	73
Medical knowledge	1. Analytical skills 2. Clinical knowledge 3. Fundamental knowledge 4. Implementation of guidelines	29	1. Clinical knowledge 2. Fundamental knowledge	11	1. Fundamental knowledge	19	59
Patient care	1. Clinical judgment 2. Complex management of clinical cases 3. Gathering patient information 4. Individualizing care 5. Managing patient plans 6. Patient care (general)	28	1. Clinical judgment 2. Decision-making skills 3. Managing patient plans	6	1. Clinical judgment 2. Patient care (highly specific functions)	9	43
Systems-based practice	1. Interdisciplinary collaboration	5	1. Interdisciplinary collaboration	1	—	—	6
Non-ACGME competency-related themes							
Personal attributes	1. Ability to work independently 2. Efficiency 3. Level of enthusiasm 4. Level of self-confidence 5. Likeability (attitude, demeanor) 6. Motivation to learn 7. Sense of responsibility 8. Work ethic	62	1. Efficiency 2. Likeability (attitude, demeanor) 3. Motivation to learn 4. Willingness to initiate action 5. Work ethic	18	1. Efficiency 2. Likeability (attitude, demeanor) 3. Motivation to learn 4. Sense of responsibility 5. Willingness to initiate action 6. Work ethic	46	126
Summative judgment	1. Ready to be promoted 2. Room for growth	28	—	—	1. Ready for unsupervised practice	4	32
Comparison to level of training	1. Below expectations 2. Exceeds expectations 3. Meets expectations	21	1. Exceeds expectations 2. Meets expectations	8	1. Meets expectations	1	30
Total	34 subthemes	232	21 subthemes	100	18 subthemes	149	481

Abbreviations: PGY indicates postgraduate year; ACGME, Accreditation Council for Graduate Medical Education.
[a]These themes were coded from 828 comments from end-of-rotation assessments for 1 cohort of 33 internal medicine residents, over their 3-year course of training. One resident withdrew during the second year, so the cohort size became 32 in PGY2 and PGY3.

Increased focus on autonomy and leadership as residents progress. Raters became more concerned with autonomy as residents advanced through their PGYs, assessing them on progressively more independent activities. This shift was reflected in several themes including interpersonal and communication skills, personal attributes, and summative judgment.

One commonly mentioned attribute beyond autonomy was being a "team player," which was characterized by how well the resident integrated within the team as a supporting member, contributing to the group according to their expected role. For PGY1 residents, this referred to their dependability and tendency to ease the burden on more senior residents and attending physicians: "[Resident name] is very thorough and dependable. He gave very good presentations and was open to feedback. [Name] is definitely a team player" (faculty assessment of PGY1 resident). In later years, the concept of team player shifted to one of supporting others' autonomy and providing assistance without imposition: "A really good senior because [Name] made a conscious effort to let me make my own plan . . . very supportive and a really good team player" (PGY1 assessment of PGY2 resident).

The subtheme of leadership as an observed competency was not present during PGY1, but it appeared during PGY2 and PGY3: "Brilliant physician, she was our fearless leader. We had a difficult month, but [Name] always remained composed and was able to think and act quickly" (PGY1 assessment of PGY3 resident). However, even for younger residents, the idea of leading a team was expressed as an aspirational goal. Raters also often commented on competencies beyond direct patient care, such as empathy, communication skills, and teamwork, especially as residents demonstrated mastery of patient care tasks:

> He interacts well with patients and demonstrates empathy and an ability to cater care to individual needs. He should set goals to polish presentation skills and develop the organizational tools that will be required to run a full team as a PGY2. (Faculty assessment of PGY1 resident)

The competency of professionalism also reflected this evolution from an intern to a team leader. PGY1 and PGY2 residents often had their professionalism appraised as a general aptitude: "Very professional at patient bedside and with all support staff in in-person and phone interactions" (PGY3 assessment of PGY1 resident). This learner exhibited professionalism, but it is unclear what he or she was actually doing to demonstrate that professionalism. However, in PGY3, specific subthemes within professionalism began to emerge. One such subtheme was "approachability" or how the resident makes him- or herself available to the team:

> Also willing to answer questions when asked. Very approachable. If there were time he might consider volunteering to teach (5-minute review of something) because he clearly has the knowledge, instead of waiting to be asked questions. (PGY1 assessment of PGY3 resident)

Approachability as a subtheme of professionalism reflects the increased emphasis on leading a team for advanced residents. Raters seemed to first be concerned with how a resident can follow and later made judgments on his or her ability to lead.

Evolution of medical knowledge over time. Medical knowledge was a prevalent theme across all years, but the tenor of comments changed substantially over the course of training. During PGY1, raters often expressed dissatisfaction with a resident's knowledge. This became less frequent during PGY2, and in PGY3, knowledge was hardly mentioned except in praise. While, in general, knowledge was praised more often than criticized during PGY1, comments such as the following were somewhat common: "Should continue to read to expand knowledge base" (faculty assessment of PGY1 resident). This contrasts with the infrequent and typically rather light criticism of knowledge (e.g., noting a resident was at rather than above level) during PGY2:

> His medical knowledge base and clinical decision making was at the appropriate level for his training, and he will do well as he progresses to his third year. (Faculty assessment of PGY2 resident)

Emphasis on resident's teaching skills in PGY2 and PGY3. Teaching first emerged as a subtheme in PGY2 and became more prominent in PGY3, with comments showing that residents are relied on not only for their leadership skills but also for how they prioritize teaching their peers:

> [Name] also took a lot of time teaching about patients, and good tips for "surviving" residency. (Faculty assessment of PGY2 resident)

> [Name] spent time every day teaching the interns and students, and was actively involved in the management of patients and education of the team. (Faculty assessment of PGY3 resident)

Association of narrative feedback with quantitative ratings and changes in these associations over time

We coded narrative comments for the quality of feedback contained therein, namely, for their orientation (praising or critical) and relevance (specificity and actionability). Across all years, 714/828 (86%) comments contained high praise and 412/828 (50%) comments were very relevant. Only 12/828 (1%) comments were both very critical and very relevant, whereas 337/828 (41%) comments both contained high praise and were very relevant (Table 2).

We found statistically significant associations between quantitative competency ratings and the feedback codes derived from narrative comments (Table 3). For the association of feedback orientation with average ratings (across all competencies and all years), the standardized regression coefficient (which can be interpreted analogous to a correlation coefficient) was moderately positive ($\beta = 0.46$, $P < .001$). In other words, comments that reflected praise were accompanied by higher quantitative ratings. We found a weak but statistically significant negative association between relevance of feedback and average ratings ($\beta = -0.09$, $P = .01$), indicating that when quantitative ratings were lower, narrative comments became slightly more specific and actionable. When analyzed by training year and by individual competency, all associations with feedback orientation were of similar magnitude and statistical significance (Table 3). By contrast, for feedback relevance, none of the analyses by year or by competency had statistically significant correlations except for the analysis of PGY1 ratings.

Changes in quantitative assessment ratings by year

The average quantitative rating across all competencies increased slightly but significantly with each training year (PGY1: mean = 5.31 [SD = 0.59], PGY2: mean = 5.58 [SD = 0.47], PGY3:

Table 2

Distribution of the Orientation and Relevance[a] of Narrative Comments (n = 828) About Internal Medicine Residents at the University of Illinois at Chicago College of Medicine, 2013–2016

	Feedback relevance[b]				
Feedback orientation	Very irrelevant, no. (%)	Irrelevant, no. (%)	Relevant, no. (%)	Very relevant, no. (%)	Total, no. (%)
Very critical	0 (0)	1 (0.1)	4 (0.5)	12 (1)	17 (2)
Critical	0 (0)	0 (0)	1 (0.1)	21 (3)	22 (3)
Modest praise	13 (2)	5 (0.6)	15 (2)	42 (5)	75 (9)
High praise	92 (11)	92 (11)	193 (23)	337 (41)	714 (86)
Total	105 (13)	98 (12)	213 (26)	412 (50)	828 (100)

[a]These features (orientation [praising or critical] and relevance [specificity and actionability]) were coded from comments from end-of-rotation assessments for 1 cohort of 33 internal medicine residents, over their 3-year course of training. One resident withdrew during the second year, so the cohort size became 32 in postgraduate years 2 and 3.
[b]Percentages in these columns may add up to more than the percentage in the total column because of rounding.

mean = 5.86 [SD = 0.43]; $P < .001$ for each pairwise comparison). Ratings for each of the competencies likewise increased significantly from PGY1 to PGY3 ($P < .001$ or = .004), except for interpersonal and communication skills ($P = .24$; Table 4). We noticed a slight, but not statistically significant ($P = .08$), decrease in rating variance (SD) over time across all competencies.

Discussion

This study presents a longitudinal analysis of narrative (qualitative) and quantitative end-of-rotation assessment data for a single cohort of medicine residents from entry to graduation. Fewer than half of the assessments had narrative (qualitative) comments. Qualitative analysis of narrative comments found that residents' work ethic was the most commonly mentioned characteristic, that medical knowledge was emphasized more in PGY1 and much less in PGY2 or PGY3, and that abilities to function independently and lead a team were progressively emphasized in later years. Nearly all narrative comments were praising (rather than critical), and three-fourths were relevant (specific and actionable). The orientation of comments showed moderate positive correlations with quantitative ratings (i.e., there was greater praise for higher-rated residents), whereas relevance showed weak negative correlations with quantitative ratings (i.e., slightly more specific and actionable comments for lower-rated residents). Quantitative ratings increased slightly but significantly over the 3 years, except for interpersonal and communication skills.

Limitations and strengths

The data for this study came from a single institution, which could limit generalizability; however, the quantitative ratings followed a widely used milestones framework. The reliability for PGY2 ratings was relatively low, which we believe was due to fewer assessments and competencies being measured during PGY2. The qualitative data analysts were not clinicians; however, the kappas were at least moderate for coding of feedback, and the analysts came to a consensus on all codes. We recognize that counting the frequency of codes in qualitative analysis has limitations; however, in this study of trends over time, it served a necessary and useful function. Although a small proportion of comments reflected a mixed orientation, we ultimately classified all comments as positive or negative by consensus. This approach permits a more streamlined analysis but could mask nuanced comments. We analyzed faculty, fellow, and resident ratings together, which could blur important differences in themes, orientation, and/or relevance.

Strengths of the study include the finding of validity evidence, such as internal structure (reliability) and relations to other variables (changes in ratings by year), that largely aligned with expectations; the longitudinal analysis of a single cohort of residents over the

Table 3

Association of the Orientation and Relevance of Narrative Comments With Quantitative Assessment Ratings[a] for Internal Medicine Residents at the University of Illinois at Chicago College of Medicine, 2013–2016

	Feedback orientation[b]		Feedback relevance[c]	
Competency and year	Standardized β coefficient	*P* value	Standardized β coefficient	*P* value
Average (all competencies, all years)	0.46	< .001	−0.09	.01
All competencies, PGY1	0.45	< .001	−0.22	< .001
All competencies, PGY2[d]	0.52	< .001	−0.06	.45
All competencies, PGY3	0.38	< .001	0.11	.06
Patient care, all years	0.44	< .001	−0.07	.33
Medical knowledge, all years	0.41	< .001	−0.10	.12
Systems-based practice, all years[d]	0.44	< .001	−0.05	.46
Practice-based learning and improvement, all years[d]	0.38	< .001	−0.02	.90
Professionalism, all years	0.45	< .001	−0.08	.11
Interpersonal and communication skills, all years[d]	0.54	< .001	−0.06	.16

Abbreviation: PGY indicates postgraduate year.
[a]Data were derived from end-of-rotation assessments for 1 cohort of 33 internal medicine residents, over their 3-year course of training; 1 resident withdrew during the second year, so the cohort size became 32 in PGY2 and PGY3. Competencies were rated on a 7-point scale with anchors of 1 = "critical deficiencies" and 7 = "ready for unsupervised practice." See footnotes below for information on the scales used to code feedback orientation and relevance. Both predictors (feedback orientation [praising or critical] and relevance [specificity and actionability]) and outcomes (competencies) were standardized to range from 0 to 1; the standardized β coefficient can be interpreted analogous to a correlation coefficient.
[b]Orientation codes: 1 indicates very critical; 2, critical; 3, modest praise; 4, high praise.
[c]Relevance codes: 1 indicates very irrelevant; 2, irrelevant; 3, relevant; 4, very relevant.
[d]During PGY2, quantitative ratings for systems-based practice, interpersonal and communication skills, and practice-based learning and improvement were omitted from the assessment form because these were measured in other training situations.

Table 4

Changes in Quantitative Assessment Ratings[a] by Year for Internal Medicine Residents at the University of Illinois at Chicago College of Medicine, 2013–2016

Competency	PGY1, mean (SD)	PGY2, mean (SD)	PGY3, mean (SD)	*P* value[b]
Average (all competencies)	5.31 (0.59)	5.58 (0.47)	5.86 (0.43)	< .001
Patient care	5.19 (0.59)	5.85 (0.51)	5.80 (0.42)	< .001
Medical knowledge	5.18 (0.59)	5.77 (0.49)	5.75 (0.39)	< .001
Systems-based practice[c]	5.38 (0.63)	—	5.89 (0.53)	.004
Practice-based learning and improvement[c]	4.63 (0.53)	—	6.25 (0.72)	< .001
Professionalism	5.56 (0.62)	5.13 (0.47)	5.92 (0.45)	< .001
Interpersonal and communication skills[c]	5.55 (0.73)	—	5.86 (0.82)	.24

Abbreviations: PGY indicates postgraduate year; SD, standard deviation.
[a]These ratings are from end-of-rotation assessments for 1 cohort of 33 internal medicine residents, over their 3-year course of training. One resident withdrew during the second year, so the cohort size became 32 in PGY2 and PGY3.
[b]*P* values represent analysis of variance (ANOVA) comparing mean ratings across years.
[c]This competency was not assessed on the end-of-rotation assessment form in PGY2.

entirety of their 3-year training; and the use of mixed quantitative and qualitative data and data analysis methods.

Integration with prior work

Our finding that comments are nearly always positive but are less often relevant mirrors previous work[19–22,31,32] and represents an area for potential improvement in assessment. This may reflect, in part, the tacit "code" used in crafting and interpreting narrative comments (i.e., being deliberately vague, couching criticism in a veil of praise, and leaving critical comments unsaid).[33–35] The emphasis on work ethic across all years and the deemphasis of medical knowledge in later years may also reflect efforts to avoid outright criticism. Efforts to improve end-of-rotation assessment quality—including workshops[36]; feedback to faculty[37,38]; organizational change[38]; and changes in the structure, timing, or wording of forms and prompts[27,39,40]—have met with variable success. Standardized ratings for the quality of narrative comments may be helpful in future investigations.[41]

Our inductive examination of the competencies reflected in narrative comments likewise adds to the literature regarding competency frameworks.[42–47] Five of the ACGME competencies were represented in these comments, and the absence of narrative comments related to practice-based learning and improvement is noteworthy. This could indicate that this competency is not routinely observed on internal medicine rotations or that raters are not attuned to observing and providing feedback on this competency. We also identified 3 non-ACGME competency themes in the narrative comments: (1) personal attributes (e.g., work ethic, enthusiasm, independence), (2) summative judgment, and (3) comparison to level of training; similar themes have been identified in previous work.[20,22,25–27] These are not competencies per se, yet they represent salient attributes that may merit consideration as part of a comprehensive assessment program. Other frameworks such as CanMEDS[43] and Tomorrow's Doctors[44] may provide additional insights into the competencies and roles expected of practicing physicians and how these might be effectively assessed.

Implications

As residents progressed through their training, the themes identified in the narrative comments also changed, reflecting changes in the expected roles and competencies of residents. These findings may suggest that measurement of resident competence over time should not simply quantify a given attribute (e.g., "how much" professionalism or system-based practice) but should also consider that the target competencies themselves may change over time as a resident matures. For example, the *nature* of the professionalism or systems-based practice expected of a resident may change over time rather than just the quantity or level of the competency.

Narrative comments complemented quantitative ratings. For example, comments seemed to contain more specific feedback for lower-scoring residents and emphasized different themes at different stages of training. Yet, we found substantial variability in the quality of narrative comments: Less than half of the assessments had narrative comments, and a quarter of these were rated as irrelevant. Moreover, nearly all comments were judged as praising. While praise is easy to give and comforting to receive, we suspect that more critical comments (if offered constructively) could help residents better identify gaps for improvement. Educators need to investigate and implement practical approaches, such as workshops and carefully worded prompts, to generate more and higher-quality narrative comments. We also need to better understand how to analyze comments once collected, synthesize this analysis with other qualitative and quantitative data, and—perhaps most important—effectively communicate this rich information with residents to facilitate meaningful change.

The appraisal of 2 dimensions of comment quality—orientation and relevance—allowed for the insightful finding that praising comments were strongly associated with higher quantitative ratings, whereas the association between feedback relevance and quantitative ratings was negative and weak. Both of these findings make sense: Praise should be (and was) associated with high quantitative ratings, whereas both high- and low-performing residents need specific, actionable feedback (resulting in high relevance codes for both high and low quantitative ratings). Beyond orientation and relevance, other characteristics of high-quality narrative comments might include context descriptions and the presence of specific examples.[41]

Although we found statistically significant differences in mean quantitative ratings across years, the absolute magnitude of variation was small (5.31, 5.58, and 5.86 for PGY1, PGY2, and PGY3, respectively). If these truly represent criterion-referenced ratings, as intended by the milestones paradigm and the wording of the items, this finding would suggest rather high baseline competence for interns followed by very minimal progression over 3 years of training.[8,12] However, we suspect that such is not the case; rather, we suspect that faculty assessors are incorporating

at least some degree of normative referencing in their ratings.[48] This could be the subject of further research and/or faculty development training.[49]

In contrast with other competency ratings, we found that interpersonal and communication skills ratings did not increase significantly over time. This may indicate that interpersonal and communication skills do not change much over time (suggesting that PGY1 residents already possess good skills and/or that PGY3 residents are not improving), that the standard for this competency changes over time (i.e., there are higher expectations for senior residents), or that interpersonal and communication skills ratings are not reflective of the actual competency.

The validation results largely aligned with expectations. For the qualitative assessment, the capture of narrative comments from a large number and wide variety of raters, including peers, senior residents, and staff physicians, provided evidence of appropriate content, as most inductively identified themes corresponded with expected ACGME competencies, although practice-based learning and improvement did not emerge in this analysis. Regarding relations with other variables, the qualitative themes evolved as residents matured and praising comments were associated with higher quantitative scores. Similar data analysis and validation approaches might be used in future research on qualitative assessments.[18]

In future studies, additional evidence might be sought from strategic (purposeful) sampling to enrich important or underrepresented competencies (such as practice-based learning and improvement), analysis of raw comments for richness and diversity, and triangulation with data from other sources.[6] Furthermore, given the limited mentions of the concepts of trust and entrustment in the narrative comments, revisions to assessments to capture these themes may be needed to inform decisions about promoting learners. Our findings provisionally support using qualitative data from end-of-rotation assessments to provide residents with formative feedback.

Acknowledgments: The authors are grateful for the technical assistance provided by Kuan Xing.

Funding/Support: None reported.

Other disclosures: None reported.

Ethical approval: This study was approved by the University of Illinois at Chicago Institutional Review Board.

A. Tekian is professor and associate dean for international affairs, Department of Medical Education, University of Illinois at Chicago College of Medicine, Chicago, Illinois; ORCID: https://orcid.org/0000-0002-9252-1588.

Y.S. Park is associate professor, Department of Medical Education, University of Illinois at Chicago College of Medicine, Chicago, Illinois; ORCID: http://orcid.org/0000-0001-8583-4335.

S. Tilton is a PharmD candidate, University of Illinois at Chicago College of Pharmacy, Chicago, Illinois.

P.F. Prunty is a PharmD candidate, University of Illinois at Chicago College of Pharmacy, Chicago, Illinois.

E. Abasolo is a PharmD candidate, University of Illinois at Chicago College of Pharmacy, Chicago, Illinois.

F. Zar is professor and program director, Department of Medicine, University of Illinois at Chicago College of Medicine, Chicago, Illinois.

D.A. Cook is professor of medicine and medical education and associate director, Office of Applied Scholarship and Education Science, and consultant, Division of General Internal Medicine, Mayo Clinic College of Medicine, Rochester, Minnesota; ORCID: https://orcid.org/0000-0003-2383-4633.

References

1 Williams RG, Dunnington GL, Klamen DL. Forecasting residents' performance—Partly cloudy. Acad Med. 2005;80:415–422.
2 Holmboe ES. Competency-based medical education and the ghost of Kuhn: Reflections on the messy and meaningful work of transformation. Acad Med. 2018;93:350–353.
3 Kuper A, Reeves S, Albert M, Hodges BD. Assessment: Do we need to broaden our methodological horizons? Med Educ. 2007;41:1121–1123.
4 Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward authentic clinical evaluation: Pitfalls in the pursuit of competency. Acad Med. 2010;85:780–786.
5 Govaerts M, van der Vleuten CP. Validity in work-based assessment: Expanding our horizons. Med Educ. 2013;47:1164–1174.
6 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.
7 Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system—Rationale and benefits. N Engl J Med. 2012;366:1051–1056.
8 Park YS, Zar FA, Norcini JJ, Tekian A. Competency evaluations in the next accreditation system: Contributing to guidelines and implications. Teach Learn Med. 2016;28:135–145.
9 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.
10 Bartlett KW, Whicker SA, Bookman J, et al. Milestone-based assessments are superior to Likert-type assessments in illustrating trainee progression. J Grad Med Educ. 2015;7:75–80.
11 Beeson MS, Holmboe ES, Korte RC, et al. Initial validity analysis of the emergency medicine milestones. Acad Emerg Med. 2015;22:838–844.
12 Hauer KE, Clauser J, Lipner RS, et al. The internal medicine reporting milestones: Cross-sectional description of initial implementation in U.S. residency programs. Ann Intern Med. 2016;165:356–362.
13 Hauer KE, Vandergrift J, Hess B, et al. Correlations between ratings on the resident annual evaluation summary and the internal medicine milestones and association with ABIM certification examination scores among US internal medicine residents, 2013–2014. JAMA. 2016;316:2253–2262.
14 Goldman RH, Tuomala RE, Bengtson JM, Stagg AR. How effective are new milestones assessments at demonstrating resident growth? 1 year of data. J Surg Educ. 2017;74:68–73.
15 Li ST, Tancredi DJ, Schwartz A, et al; Association of Pediatric Program Directors (APPD) Longitudinal Educational Assessment Research Network (LEARN) Validity of Resident Self-Assessment Group. Competent for unsupervised practice: Use of pediatric residency training milestones to assess readiness. Acad Med. 2017;92:385–393.
16 Turnbull J, van Barneveld C. Assessment of clinical performance: In-training evaluation. In: Norman G, van der Vleuten C, Newble D, eds. International Handbook of Research in Medical Education. Dordrecht, the Netherlands: Kluwer Academic Publishers; 2002:793–810.
17 Chou S, Cole G, McLaughlin K, Lockyer J. CanMEDS evaluation in Canadian postgraduate training programmes: Tools used and programme director satisfaction. Med Educ. 2008;42:879–886.
18 Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA. Using in-training evaluation report (ITER) qualitative comments to assess medical students and residents: A systematic review. Acad Med. 2017;92:868–879.
19 Ringdahl EN, Delzell JE, Kruse RL. Evaluation of interns by senior residents and faculty: Is there any difference? Med Educ. 2004;38:646–651.
20 Frohna A, Stern D. The nature of qualitative comments in evaluating professionalism. Med Educ. 2005;39:763–768.
21 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.
22 Jackson JL, Kay C, Jackson WC, Frank M. The quality of written feedback by attendings of internal medicine residents. J Gen Intern Med. 2015;30:973–978.
23 Plymale MA, Donnelly MB, Lawton J, Pulito AR, Mentzer RM. Faculty evaluation of surgery clerkship students: Important components of written comments. Acad Med. 2002;77(suppl 10):S45–S47.
24 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.

25 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(suppl 10):S30–S34.
26 White JS, Sharma N. "Who writes what?" Using written comments in team-based assessment to better understand medical student performance: A mixed-methods study. BMC Med Educ. 2012;12:123.
27 Nagler J, Pina C, Weiner DL, Nagler A, Monuteaux MC, Bachur RG. Use of an automated case log to improve trainee evaluations on a pediatric emergency medicine rotation. Pediatr Emerg Care. 2013;29:314–318.
28 Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: Theory and application. Am J Med. 2006;119:166.e7–166.e16.
29 Charmaz K. Reconstructing theory in grounded theory studies. In: Constructing Grounded Theory: A Practical Guide Through Qualitative Analysis. London, UK: SAGE; 2006:123–151.
30 Varpio L, Ajjawi R, Monrouxe LV, O'Brien BC, Rees CE. Shedding the cobra effect: Problematising thematic emergence, triangulation, saturation and member checking. Med Educ. 2016;51:40–50.
31 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.
32 Vivekananda-Schmidt P, MacKillop L, Crossley J, Wade W. Do assessor comments on a multi-source feedback instrument provide learner-centred feedback? Med Educ. 2013;47:1080–1088.
33 Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: Faculty interpretations of narrative evaluation comments. Med Educ. 2015;49:296–306.
34 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.
35 Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: Residents' interpretations of written assessment comments. Med Educ. 2017;51:401–410.
36 Dudek NL, Marks MB, Wood TJ, et al. Quality evaluation reports: Can a faculty development program make a difference? Med Teach. 2012;34:e725–e731.
37 Dudek NL, Marks MB, Bandiera G, White J, Wood TJ. Quality in-training evaluation reports—Does feedback drive faculty performance? Acad Med. 2013;88:1129–1134.
38 Littlefield JH, Darosa DA, Paukert J, Williams RG, Klamen DL, Schoolfield JD. Improving resident performance assessment data: Numeric precision and narrative specificity. Acad Med. 2005;80:489–495.
39 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;16:427–434.
40 McOwen KS, Bellini LM, Shea JA. Including resident photographs on electronic evaluations: Is a picture worth a thousand words? Teach Learn Med. 2010;22:304–306.
41 Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. Med Educ. 2008;42:816–822.
42 Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the Accreditation Council for Graduate Medical Education: A systematic review. Acad Med. 2009;84:301–309.
43 Frank JR, Snell L, Sherbino J. CanMEDS 2015 Physician Competency Framework. Ottawa, ON, Canada: Royal College of Physicians and Surgeons of Canada; 2015.
44 General Medical Council. Outcomes for Graduates (Tomorrow's Doctors). Manchester, UK: General Medical Council; 2015.
45 Choe JH, Knight CL, Stiling R, Corning K, Lock K, Steinberg KP. Shortening the miles to the milestones: Connecting EPA-based evaluations to ACGME milestone reports for internal medicine residency programs. Acad Med. 2016;91:943–950.
46 Edgar L, Roberts S, Yaghmour NA, et al. Competency crosswalk: A multispecialty review of the Accreditation Council for Graduate Medical Education milestones across four competency domains. Acad Med. 2018;93:1035–1041.
47 Boyd VA, Whitehead CR, Thille P, Ginsburg S, Brydges R, Kuper A. Competency-based medical education: The discourse of infallibility. Med Educ. 2018;52:45–57.
48 Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. Med Educ. 2014;48:614–622.
49 Kogan JR, Conforti LN, Bernabeo E, Iobst W, Holmboe E. How faculty members experience workplace-based assessment rater training: A qualitative study. Med Educ. 2015;49:692–708.