The American Journal of Surgery 217 (2019) 288–295

Contents lists available at ScienceDirect

The American Journal of Surgery

journal homepage: www.americanjournalofsurgery.com

What do quantitative ratings and qualitative comments tell us about general surgery residents' progress toward independent practice? Evidence from a 5-year longitudinal cohort

Ara Tekian [a, *], Martin Borhani [b], Sarette Tilton [c], Eric Abasolo [c], Yoon Soo Park [a]

[a] Department of Medical Education, University of Illinois at Chicago College of Medicine, Chicago, IL, USA
[b] Department of Surgery, University of Illinois at Chicago College of Medicine, Chicago, IL, USA
[c] University of Illinois at Chicago College of Pharmacy, Chicago, IL, USA

ARTICLE INFO

Article history:
Received 10 May 2018
Received in revised form
12 September 2018
Accepted 28 September 2018

Keywords:
Next accreditation system
Qualitative feedback
Learning trajectories
Surgery residency program

ABSTRACT

Background: This study examines the alignment of quantitative and qualitative assessment data in end-of-rotation evaluations using longitudinal cohorts of residents progressing throughout the five-year general surgery residency.

Methods: Rotation evaluation data were extracted for 171 residents who trained between July 2011 and July 2016. Data included 6069 rotation evaluations forms completed by 38 faculty members and 164 peer-residents. Qualitative comments mapped to general surgery milestones were coded for positive/negative feedback and relevance.

Results: Quantitative evaluation scores were significantly correlated with positive/negative feedback, $r = 0.52$ and relevance, $r = -0.20$, $p < .001$. Themes included feedback on leadership, teaching contribution, medical knowledge, work ethic, patient-care, and ability to work in a team-based setting. Faculty comments focused on technical and clinical abilities; comments from peers focused on professionalism and interpersonal relationships.

Conclusions: We found differences in themes emphasized as residents progressed. These findings underscore improving our understanding of how faculty synthesize assessment data.

Introduction

In graduate medical education, decisions to promote and remediate learners are often based on combining information from multiple assessments.[1] In this respect, faculty use assessment data consisting of quantitative ratings and qualitative comments to provide feedback to learners and to inform the clinical competency committee (CCC) on progress toward independent practice.[2–6]

End-of-rotation evaluations (also known as in-training evaluation report [ITER] in Canada) require raters (faculty, fellows, or peer residents) to assign quantitative competency ratings based on predetermined anchors and write qualitative narrative comments.[7] Prior studies have noted concerns over rotation evaluation scores as not being reliable assessments of learner performance, despite their prevalent use in making promotion decisions.[8–13] However, recent studies have shown that rotation evaluations scores can be reliable and demonstrate validity evidence, provided that evaluations from multiple raters are collected and aggregated over a sufficient period of time.[7,14–21] This study focuses on synthesizing both quantitative and qualitative data from rotation evaluations, in the context of identifying competency-based developmental levels.

Analysis of qualitative data, captured as narrative comments on rotation evaluation forms, have shown to provide useful information that allow better discrimination of learner performance. For example, studies have found that narrative comments can accompany quantitative scores; moreover, studies have also identified them as more reliable in reflecting performance than quantitative data.[22,23] In addition, qualitative comments have demonstrated predictive utility in identifying residents' future performances.[24] These implications reveal potential use of qualitative data in contributing to making decisions of learner progress. Yet, the relationship between qualitative and quantitative data in rotation evaluation forms has not been examined within the context of the Next Accreditation System (NAS) by the Accreditation Council for

* Corresponding author. Department of Medical Education College of Medicine, University of Illinois at Chicago 808 South Wood Street, 963 CMET (MC 591), Chicago, IL, 60612-7309, USA.
E-mail address: tekian@uic.edu (A. Tekian).

https://doi.org/10.1016/j.amjsurg.2018.09.031
0002-9610/

Graduate Medical Education (ACGME),[25,26] leaving Clinical Competency Committees and faculty responsible for making supervision decisions of learners unsure how to incorporate qualitative data and themes into their assessment decision process. The NAS promotes ongoing assessment of learners throughout their residency training and the use of multiple assessment data to inform promotion decisions.

Since July 2014, General Surgery postgraduate programs in the United States began tracking learners based on milestones and reporting them to ACGME every six months. As such, improving our understanding of qualitative narrative comments and incorporating them to inform promotion decisions and learner feedback have become increasingly important.

This study examines the alignment of quantitative and qualitative assessment data using longitudinal cohorts of residents progressing throughout the five-year general surgery residency. We studied the longitudinal progression of learners using both quantitative ratings and qualitative narrative comments. The following questions were examined:

1. Do qualitative comments align with quantitative scores for trainees from entry to graduation?
2. Do faculty provide rich qualitative comments that are useful to provide relevant and actionable feedback to trainees?

Methods

Participants

Residents. Retrospective data from 171 residents, who were part of the General Surgery residency program at the University of Illinois at Chicago (UIC) College of Medicine were used (five years: July 1, 2011–June 30, 2016).

Raters. Rotation evaluation forms were rated by General Surgery faculty ($n = 38$) and peer residents ($n = 164$, including categorical General Surgery residents and rotating residents from other specialties).

Assessment data

End-Of-Rotation Evaluation. End-of-rotation evaluation forms from the General Surgery residency program at the University of Illinois at Chicago (UIC) College of Medicine were used. The end-of-rotation evaluation form has 30 items (13 items in faculty form, 17 items in peer form). Each item was mapped to one or more of the 16 General Surgery subcompetencies which are derived from the six ACGME Core Competencies (3 subcompetencies in Patient Care [PC], 2 subcompetencies in Medical Knowledge [MK], 2 subcompetencies in Systems-Based Practice [SBP], 3 subcompetencies in Practice-Based Learning and Improvement [PBLI], 3 subcompetencies in Professionalism [PROF], and 3 subcompetencies in Interpersonal Communication Skills [ICS]).[20] Each item (mapped to a General Surgery subcompetency) was rated on a 9-point scale, corresponding to milestone anchors ("Level 1" = 1, "Level 2" = 3, "Level 3" = 5, "Level 4" = 7; "Level 5" = 9 points). Even-numbered points corresponded to scores between milestone levels. Data were collected using the New Innovations (NI) database and extracted for analysis.

Qualitative analysis

Two types of coding were conducted: (1) extracting themes and the subthemes and (2) coding comments for positive/negative feedback and relevance of feedback. Coding was completed by two trained researchers. A sample of 50 comments were initially coded extracting themes and subthemes; discrepancies were discussed until agreement was reached. Interrater agreement kappa for the initial coding was 0.85. A third reviewer further refined and confirmed the themes and the subthemes. In the second type of coding, comments were coded on a 4-point scale: (1) positive/negative feedback, and (2) relevance of feedback (see Table 1). The basis for this coding was rooted in prior work by Hatala et al.[12] and Ginsburg et al.[13] who previously examined the alignment between qualitative and quantitative assessment data, focusing on positive/negative feedback. We added a second dimension on relevance to further examine whether the coding was actionable and diagnostic, to add meaning to the quality of feedback provided. Interrater agreement kappa for positive/negative and relevance were 0.79 and 0.72, respectively. Disagreement was resolved through consensus.

Coding positive/negative and relevance of feedback. Qualitative comments were coded to measure their association with quantitative evaluation scores assigned by raters. We wanted to examine whether a resident with high evaluation scores also had positive comments, and conversely whether low evaluation scores corresponded with negative comments. In addition, we wanted to examine the relationship between evaluation scores and the relevance of feedback provided in the qualitative comments. For example, a comment such as "excellent job" does not provide meaningful or actionable information; on the other hand, a comment such as "[resident] developed a positive rapport with his patients which was evidenced on my rounds" provides more relevant information about the learner. In the first dimension (positive/negative feedback), coding for 1 to 4 anchors corresponded to "Highly Critical", "Critical", "Modest Praise", and "High Praise", respectively. In the second dimension (relevance of feedback), coding for 1 to 4 anchors corresponded to "Highly Irrelevant", "Irrelevant", "Relevant" and "Highly Relevant" (see Table 1).

Themes. Comments were also analyzed for themes following a descriptive study using content analysis and generating categories.[27] Two reviewers extracted themes from the comments, by identifying themes and sub-themes, stratifying them by PGY level. As member check, results were compared and re-categorized into different themes and sub-themes, until agreement was reached by the two reviewers and confirmed by a surgeon-educator. Eight themes were identified and remained consistent throughout the five years (Table 2). An additional 46 sub-themes were also identified. Saturation was reached after analyzing 400 comments. The two reviewers were able to independently identify each unique idea represented.

Quantitative Data Analysis. Rotation evaluation scores were examined using descriptive statistics. Unit of analysis was the learner by year. Mean scores were created for the six core competencies, by taking the average across subcompetencies. Evaluation scores were correlated with coded values of positive/negative feedback and relevance of feedback from qualitative comments, to examine the alignment between qualitative and quantitative scores in rotation evaluation forms. Generalizability theory was used to examine variance components in rotation evaluation scores by PGY level and to estimate reliability (Table 3). A [rater (r): person (p)] x [subcompetency (s): competency (c)] design was used, following variance components structure from previous analysis.[13,22]

Data compilation and analyses were conducted using Stata 14 (Stata Corp, College Station, TX). The UIC institutional review board approved this study.

Results

Descriptive statistics

Evaluation records and qualitative comments. A total of 6069

Table 1
Coding rubric for narrative comments: Positive/negative and relevance of feedback.

Scale	Dimension 1: Positive/Negative Feedback		Dimension 2: Relevance of Feedback	
	Anchor	Sample Comment	Anchor	Sample Comment
4	High Praise	*Very strong overall performance. [Resident] is hard-working and anxious to improve his skills. To his credit, he seeks feedback and criticism. He does thorough evaluations. His clinical acumen should improve as he studies and gains experience. Surgical skills are typical of an intern; I am confident that he will continue to improve.*	Highly Relevant	*[Resident] is clearly one of the best junior level house staff in the past few years. He is very dependable and organized in patient care at the same time not missing any single opportunity to learn something. Manually talented as well. Performed a few cases amazingly well for his level of training. Excited about seeing further progress.*
3	Modest Praise	*[Resident] is smart and hard-working. He seemed a little overwhelmed on the trauma service early in this rotation, but seemed more comfortable as time went by and his confidence level increased.*	Relevant	*Good job as trauma senior. Operative skills OK. Did a good job of leading the service with excellent attention to detail.*
2	Critical	*Needs to become more engaged in the care of patient's. Very efficient resident, but start to recognize patient issues and make some simple decisions regarding care.*	Irrelevant	*Very good job while on the vascular service. Ran the service well.*
1	Highly Critical	*[Resident's] clinical performance is not up to the same level as his peers. He has difficulty reliably executing a care plan and requires excessive supervision. More concerning, he has very little insight into his deficiencies. He seems to have no desire to act upon criticism and modify his approach. This tendency should be closely monitored in his PGY3 year.*	Highly Irrelevant	*Excellent resident.*

Note: Coding for all narrative comments were conducted by two reviewers. Discrepancies were resolved through discussion until consensus was achieved for all comments.

rotation evaluation records were extracted from the NI database; among them, only 3129 of evaluation records (52%) had qualitative comments. On average, residents received 5 evaluations (SD = 2) from faculty and 19 evaluations (SD = 11) from peers per year. There were differences in the number of evaluation records by PGY-level.

Evaluation ratings by year. Overall, rotation evaluation scores increased significantly during the five years, $p < .001$. In particular, items measuring PC, MK, and SBP had greater rates of improvement across years (Table 3).

Quantitative coding of qualitative comments. Majority of narrative comments were very positive (71% from faculty and 84% from peer evaluations). However, for relevance, 56% of faculty and 69% of peer evaluations had relevant or very relevant comments. Between training years, there were no differences in quality of comments from peers, $p = .487$. However for faculty qualitative comments, there were significant difference in the quality of comments by training year, $p < .001$. Across training years, there were 28%, 31%, 41%, 58%, and 33% of comments coded as highly relevant from PGY1 to PGY5 (see Table 4). Between positive/negative and relevance coding, there was negative association, $r = -0.50$, $p < .001$.

Alignment between rotation evaluation scores and qualitative comments. The reliability of rotation evaluations ratings was good, Φ-coefficient = 0.72. Quantitative evaluation scores were significantly correlated with positive/negative feedback indicating alignment between quantitative and qualitative feedback, $r = 0.52$. However, when residents received higher quantitative ratings, the relevance of comment was significantly lower, $r = -0.20$, $p < .001$. These findings indicate alignment between quantitative and qualitative comments. However, when qualitative comments were more negative, there was greater quality of diagnostic and actionable feedback for residents.

Qualitative results

Themes extracted were divided into two categories: (1) ACGME-related and (2) non-ACGME related themes. Overall, there were 523 unique faculty evaluations with comments (70%), resulting in 1126 comments extracted for analysis (allowing for duplicates). ACGME-related themes included PC, MK, PROF, ICS, and SBP, with highest number of comments from PC and MK. Non-ACGME related themes were personal attributes and traits, summative judgements, and comparison to level of training. Table 2 summarizes the comments provided by faculty (see Table S1 in Supplementary Material for resident comments). The number of subthemes for the ACGME-related themes varied from one for SBP, to eight for professionalism in both tables. In addition, the number of subthemes were different at each PGY level. For example in Table 2, for professionalism, there was only one subtheme ("Team Player") during PGY3. Under the ACGME-related themes, the highest number of comments were for PC (178) and the lowest was for SBP (4). For non-ACGME related themes, the highest number of comments were for personal attributes and traits (484), and the lowest for comparison of training (64). The total number of comments per PGY level from year one to five were 415, 247, 95, 249, and 120 respectively, indicating a significant drop in the number of comments during PGY3 and an abrupt increase during PGY4. The percentage of evaluations without comments from PGY1 to PGY5 were 32%, 26%, 32%, 25%, and 30% respectively. A sample of quotes by the faculty are presented below highlighting positive (praise) and negative (concerns) areas.

Faculty comments

Operative Skills. Faculty frequently commented on residents'

Table 2
Competency-specific themes derived from narrative comments by faculty about surgery residents at the University of Illinois at Chicago, 2011–2016.

Theme	PGY1		PGY2		PGY3		PGY4		PGY5		All years
	Sub-theme	C	Sub-theme	C	Sub-theme	C	Sub-theme	C	Sub-theme	C	C
ACGME-related Themes											
Patient Care	1 Clinical Judgment 2 Complex Management 3 Decision-making 4 Managing Patient Plans 5 Operative Skills 6 Patient Care (General)	61	1 Clinical Judgment 2 Decision Making 3 Gathering Patient Information 4 Managing Patient Plans 5 Patient Care (General) 7 Patient Education	53	1 Clinical Judgment 2 Managing Patient Plans 3 Operative Skills 4 Patient Care (General)	12	1 Clinical Judgement 2 Decision Making 3 Managing Patient Plans 4 Operative Skills 5. Patient Care (General)	44	1 Clinical Judgment 2 Managing Patient Plans 3 Operative Skills 4 Patient Care (General)	8	178
Medical Knowledge	1 Clinical Knowledge 2 Fundamental Knowledge	72	1 Clinical Knowledge 2 Fundamental Knowledge	30	1 Fundamental Knowledge	8	1 Clinical Knowledge 2 Fundamental Knowledge	18	1 Clinical Knowledge 2 Fundamental Knowledge	11	139
Professionalism	1 Compassion 2 Ethical Judgment 3 Level of Professionalism 4 Punctuality 5 Respectfulness 6 Self-awareness for Improvements 7 Team Player 8 Trustworthiness	36	1 Compassion 2 Ethical Judgment 3 Level of Professionalism 4 Punctuality and Attendance 5 Respectfulness 6 Self-Awareness for Improvement 7 Team Player 8 Trustworthiness	29	1. Team Player	5	1 Compassion 2 Level of Professionalism 3 Respectfulness 4 Self-awareness for Improvement 5 Team Player 6 Trustworthiness	18	1 Level of Professionalism 2 Punctuality and Attendance 3 Self-awareness for Improvement 4 Team Player	9	97
Interpersonal/ Communication Skills	1 Clinical Judgment 2 Decision-making 3 Managing Patient Plans 4 Patient Care (General)	33	1 Communication Skills (General) 2 Delegation of Tasks 3 Leadership Skills 4 Rapport with Patients/ Caregivers 5 Teaching Skills 6 Writing Skills	22	1 Communication Skills 2 Delegation of Tasks 3 Leadership Skills 4 Rapport with Patients/Caregivers 5 Writing Skills	8	1 Communication Skills 2 Delegation of Tasks 3 Leadership Skills 4 Presentation Skills 5 Service as a Role Model 6 Writing Skills	15	1 Communication Skills 2 Leadership Skills 3 Rapport with Patients/ Caregivers 4 Teaching Skills	8	86
Systems-Based Practice	1 Difficulty with New/ Foreign System	4									4
Non-ACGME-related Themes	**PGY1**		**PGY2**		**PGY3**		**PGY4**		**PGY5**		
Personal Attributes and Traits	1 Ability to Incorporate Feedback/Able to Improve Upon Critique 2 Ability to Work Independently 2 Efficiency 3 Level of Enthusiasm 4 Level of Confidence 5 Maturity 6 Motivation to Learn 7 Personality, Attitude or Demeanor 8 Sense of Responsibility 9 Thorough/Attention to Detail 10 Willingness to Initiate Action	178	1 Ability to Work Independently 2 Efficiency 3 Level of Enthusiasm 4 Level of Self-Confidence 5 Maturity 6 Motivation to Learn 7 Personality, Attitude or Demeanor 8 Sense of Responsibility 9 Thorough/Attention to Detail 10 Willingness to Initiate Action 11 Work Ethic	93	1 Ability to Work Independently 2 Efficiency 3 Level of Enthusiasm 4 Level of Self-Confidence 5 Maturity 6 Motivation to Learn 7 Personality, Attitude or Demeanor 8 Sense of Responsibility 9 Willingness to Initiate Action 10 Work Ethic	54	1 Ability to incorporate feedback/able to improve upon critique 2 Efficiency 3 Level of Enthusiasm 4 Level of Confidence 5 Maturity 6 Motivation to Learn 7 Personality, Attitude or Demeanor 8 Sense of Responsibility 9 Thorough/Attention to Detail 10 Willingness to Initiate Action 11 Work Ethic	97	1 Efficiency 2 Level of Enthusiasm 3 Level of Self-Confidence 4 Maturity 5 Motivation to Learn 6 Personality, Attitude, or Demeanor 7 Sense of Responsibility 8 Thorough/Attention to Details 9 Willingness to Initiate Action 10 Work Ethic	62	**484**

(*continued on next page*)

Table 2 (*continued*)

Non-ACGME-related Themes	PGY1		PGY2		PGY3		PGY4		PGY5		
	11 Work Ethic										
Summative Judgement	1 Ready for Unsupervised Practice	8	1 Ready for Unsupervised practice 2 Ready to be Promoted	5	1 Prediction of Success 2 Ready for Unsupervised Practice	5	1 Prediction of Success 2 Prediction of Success as Chief 3 Room for Growth 4 Shown Improvement	40	1 Prediction of Success 2 Prediction of Success as Chief 3 Room for Growth 4 Shown Improvement	16	74
Comparison to Level of Training	1 Below Expectations 2 Exceeds Expectations	23	1 Exceeds Expectations 2 Meets Expectations	15	1 Below Expectations 2 Exceeds Expectations	3	1 Below Expectations 2 Exceeds Expectations 3 Meets Expectations	17	1 Below Expectations 2 Exceeds Expectations 3 Meets Expectations	6	64
TOTAL	*N = 33 sub-themes*	***415***	*N = 38 sub-themes*	***247***	*N = 25 sub-themes*	***95***	*N = 37 sub-themes*	***249***	*N = 31 sub-themes*	***120***	***1126***

These themes were drawn from 743 evaluations by faculty (1126 comments) regarding 171 General Surgery residents. Abbreviations: PGYn = postgraduate year n, C = counts.

operative skills across all five years. As effective operative skills are essential in a surgical resident, faculty members often focused on the residents' deficiencies.

> *"… Unfortunately, [he] stands also at the bottom of his class in terms of surgical skills and manual dexterity as assessed by multiple attendings in both operative theater and at the surgical skills lab. He is trying very hard to overcome these deficiencies and we should give credit to him for this genuine and extraordinary effort he is putting in place to perform inside the operating room at the level of his peers and classmates."* (Faculty assessment of PGY2 resident)

Although faculty indicated shortcomings in the residents' operative skills, they expressed encouragement by giving residents credit for striving to improve, hoping they do well in the future, or stating the resident will continue to develop. This form of constructive criticism is prevalent when commenting on other themes as well.

Level of Self-Confidence. Level of self-confidence emerged across all years. Faculty commonly related the residents' level of confidence to their overall performance.

> *"... He seemed easily overwhelmed by the work load and was absent on occasion without explanation. He does not seem to function well while under stress. He tended to be inefficient in performing the typical work responsibilities of a surgical intern. He did seek assistance appropriately. We cannot gauge his surgical skills since he rarely if ever came to the OR."* (Faculty evaluation of PGY1 resident)

In this case, the faculty member noted that his level of self-confidence negatively impacted his work flow and time spent practicing his operative skills. Other narrative comments also mentioned residents' level of confidence affecting their efficiency by impeding multitasking.

> *"[Name] is a good man and tries hard to perform at the level of his Urology colleagues but, unfortunately, has some limitations. He gives the impression of being constantly overwhelmed and has serious difficulty to multi-tasking…"* (Faculty evaluation of PGY3 resident)

While some narrative comments related their level of self-confidence to other deficiencies, other comments found the residents were generally proficient apart from their self-confidence. The following two faculty members share similar observations regarding the same resident.

> *"[Name] is a very serious and hardworking resident. He has sound judgement and demonstrates the ability to apply newly learned principles. He needs to build self-confidence to match his level of performance."* (Faculty evaluation of PGY4 resident)

> *"I basically like [Name] a lot. He is quiet and effective. Does his work very well without fanfare. But the question is whether he really gets all the credit he deserves. He will have to become a leader as a Chief Resident and you cannot do that by being exceedingly quiet."* (Faculty evaluation of PGY4 resident)

In this comment, the faculty member noted that the resident's lack of self-confidence would cost him chief resident despite his other positive qualities.

Observation of Growth. As faculty follow residents throughout training, they note the residents who have shown improvement

Table 3
Quantitative ratings by competency and training year: Mean (SD).

Competency	PGY-1		PGY-2		PGY-3		PGY-4		PGY-5	
PC	7.65	(1.63)	7.95	(1.56)	7.88	(1.77)	7.73	(1.83)	8.25	(1.66)
MK	7.21	(1.57)	7.52	(1.45)	7.57	(1.78)	7.27	(1.67)	8.10	(1.64)
PBLI	7.41	(1.60)	7.47	(1.55)	7.57	(1.73)	7.47	(1.69)	8.00	(1.55)
ICS	7.99	(1.72)	7.97	(1.68)	8.19	(1.48)	8.02	(1.95)	8.10	(1.67)
PROF	7.84	(1.59)	7.90	(1.49)	8.17	(1.62)	7.84	(1.81)	8.36	(1.67)
SBP	7.59	(1.59)	7.75	(1.38)	8.09	(1.50)	7.77	(1.56)	8.30	(1.56)

Note: There were significant growth in scores for PC, MK, PBLI, PROF, and SBP, $p < .05$.

Table 4
Highly relevant comments by training year and rater group: %.

Rater Group	PGY-1	PGY-2	PGY-3	PGY-4	PGY-5	Overall	*p*-value
Faculty	28	31	41	58	33	35	<.001
Peer Residents	36	35	40	37	41	37	.487

Note: There were significant differences in the proportion of highly relevant comments for faculty, but not for peer-residents.

and those who strive to improve. Below is an example of two faculty members describing the same resident.

> *"[Name] makes great efforts to recognize and address his own deficiencies and actively pursues self-improvement. He is respectful, considerate and has outstanding interpersonal skills..."* (Faculty evaluation of PGY2 resident)

> *"[Name] continues to show substantial growth and improvement. He was one of the most engaged senior residents on the service this year…"* (Faculty evaluation of PGY2 resident)

At times, observations of growth are most highlighted in the final PGY5 year.

> *"Since the last rotation [Name] has really progressed well and is showing a steep learning curve. She is responsible and a hard worker. Her patient care has continued to improve with more experience. She is reliable and is a terrific member of the team. She is progressing well and I look forward to seeing her continue to improve and shine among her peers. She is outspoken but not to a fault and her eagerness is apparent. Keep up the good work!"* (Faculty evaluation of PGY5 resident)

When faculty measured their improvement, they relate it to not only technical skills, but to their interpersonal and communication skills and their personal and professional development.

Maturity. Maturity often emerged in the faculty's narrative comments. Often statements touching on maturity also touched on comparison level of training.

> *"[Name] did a great job on his trauma rotation. He is a mature and confident senior resident. He understands his role as a service chief fully. He handled a very difficult situation with an insubordinate junior EM resident with tact and maturity. Clearly, he is one of the best residents in his class."* (Faculty evaluation of PGY1 resident)

Residents' comments

The total number of narrative comments by residents (4,127) were almost fourfold, compared to the number of faculty comments (1,126). Under the ACGME-related themes, the highest number of comments were for ICS (611) and professionalism (563) in contrast to comments by the faculty, where the highest were for PC (178) and MK (139). For the non-ACGME related themes, the highest number of comments were for personal attributes and traits (1,898), which comprised 46% of the entire comments. Furthermore, the highest number of comments were provided during PGY1 level (39%, 1,628). The total number of comments per PGY level from year one to five were 1628, 811, 559, 619, and 510 respectively, indicating greater number of comments during PGY1 and PGY2. A sample of quotes by residents are presented below highlighting positive (praise) and negative (concerns) areas.

Teaching skills

When junior residents evaluate senior residents on teaching skills, they were able to further elaborate on this topic because they had personal experiences given the nature of the mentor/mentee relationship.

> ***"...****He has outstanding empathy and perception, able to realize fine differences between errors that occur from unrealistic expectations versus errors that occur from inexperience. In realizing this, [Name] is able to teach with tactful and judicious consideration, making it an excellent experience to work with and learn from him."* (PGY1 resident assessment of a PGY3 resident)

Team Player. Similar to teaching skills, observations of how one works in a team draws from interactions between collaborating residents.

> *"Not a team player, very rude to other residents and nurses."* (PGY1 resident assessment of a PGY2 resident)

> *"...A good team player who ensures all team members are informed and present for learning opportunities. Always willing to help."* (PGY1 resident assessment of a PGY2 resident)

Quotes highlighted ideal and non-ideal traits when working in a team.

Personality, Attitude or Demeanor. Themes around personality, attitude or demeanor emerged from daily interactions between residents.

> *"Poor quality resident. Rude, selfish, disrespectful and has peculiar personality and attitude. She uses offensive wording all the time when dealing with other residents and when talking about patients. This resident is not helpful to other residents at all, always defer work to other residents to complete … She was described as not helpful by almost all residents working with her and they don't like working with her given the hostile environment she's creating…"* (PGY2 resident assessment of a PGY4 resident)

Narrative comments showed both favorable and unfavorable characteristics regarding personality, attitude or demeanor. Overall, comments provided by residents compared to faculty were more personal, speaking to nuances of actually working closely with residents.

Discussion

This study presents a longitudinal analysis of qualitative and quantitative rotation evaluation data, following cohorts of surgery residents from entry to graduation. We also examine the corresponding milestone levels of residents based on end-of-rotation evaluation forms completed by supervising faculty and peer residents. Results reveal the value of qualitative comments, as they are aligned with the quantitative comments, but also provide meaningful and relevant information, beyond what is captured solely in the quantitative ratings. Moreover, we report the underuse of qualitative comments in rotation evaluation forms, as only half of rotation evaluation records contained any narrative comments, which may call for programmatic policy on methods to promote their use. Findings from this study are consistent with prior studies on the use of qualitative comments from ITERs.[13,28] However, we also expand their utility by demonstrating that qualitative comments do have a longitudinal component and can be used to track resident progress.

In this study, we coded qualitative comments into two dimensions: (1) positive/negative and (2) relevance. These dimensions provided information on whether qualitative comments were aligned with quantitative ratings and also provided information on the types of actionable and relevant feedback provided to residents. In this manner, this coding provides the first demonstration in General Surgery, relating to how qualitative and quantitative ratings can be synthesized in workplace-based assessments, following the implementation of NAS which provides guidelines for developmental milestones toward unsupervised practice in General Surgery. Prior studies have mostly examined quantitative ratings and how they have been used to inform CCC decisions or correlate with other assessment scores. As such, this study contributes to advancing our understanding of how qualitative comments can inform and even reinforce feedback from quantitative scores.

Beyond the association between quantitative and qualitative data, our study also shows the impact that qualitative comments have on possible improvement to resident performance, as measured using quantitative scores. During PGY4, there was significant increase in both the number and quality of qualitative comments provided to residents – this is particularly meaningful, as quantitative scores increased significantly during the latter two years of training. This may provide insights on identifying causal relationships between the quality of comments and subsequent performance. Additional research may be needed to confirm these findings.

Our study also showed significantly more PC and MK comments from faculty, whereas resident comments focused more heavily on ICS and PROF. This is consistent with prior studies[7,17] which noted greater variability in ICS and PROF competencies from peer-resident evaluations perhaps due to longer daily contact. Moreover, resident evaluations also provided meaningful information pertaining to personal and professional development. These findings underscore a need to promote better use of peer evaluations, which contain useful information.

This study was conducted at a single residency program. As such, findings may need further generalization through multisite studies and collaboration with other medical specialties. However, our data were based on large-scale retrospective analysis covering five years of records, which provide longitudinal trends as identified in this study. Moreover, additional efforts are needed by surgeons and clinician educators to identify strategies to link essential quantitative and qualitative assessment data that can best inform learner progress and remediation needs.

In summary, our findings call for better integration to synthesize and combine information conveyed in qualitative and quantitative data that can inform resident progress. Our results highlight alignment in qualitative and quantitative data; however, there was variation in the quality of comments provided by faculty and peer residents, covering different trends throughout their residency training. These findings can be used to inform better synthesis of assessment data for feedback, monitoring of resident progress, and ongoing faculty development, as residents train toward unsupervised practice.

Conflicts of interest

The authors do not have any conflict of interest.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Summary

This study examines the synthesis of qualitative and quantitative data from rotation evaluations collected from longitudinal cohorts of General Surgery residents. Results showed alignment in qualitative and quantitative data, with significant variation in themes and quality of comments as residents progressed toward unsupervised practice.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.amjsurg.2018.09.031.

References

1. Nasca TJ, Philibert I, Brigham T, Flynn TC. The next GME accreditation system. *N Engl J Med*. 2012;366:1051–1056.
2. Kuper A, Reeves S, Albert M, Hodges BD. Assessment: do we need to broaden our methodological horizons? *Med Educ*. 2007;41:1121–1123.
3. Ginsburg S, McIlroy J, Oulanova O, et al. Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Acad Med*. 2010;85:780–786.
4. Govaerts M, van der Vleuten CP. Validity in work-based assessment: expanding our horizons. *Med Educ*. 2013;47:1164–1174.
5. Andolsek K, Padmore J, Hauer KE. Clinical competency committees: a guidebook for programs (ACGME). https://www.acgme.org/Portals/0/ACGMEClinicalCompetencyCommitteeGuidebook.pdf; 2015. Accessed May 10, 2018.
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. Park YS, Riddle J, Tekian A. Validity evidence of resident competency ratings and the identification of problem residents. *Med Educ*. 2014;48:614–622.
8. Chaudhry S, Holmboe E, Beasley B. The state of evaluation in internal medicine residency. *J Gen Intern Med*. 2008;23:1010–1015.
9. Cleland J, Knight L, Rees C, et al. Is it me or is it them? Factors that influence the passing of underperforming students. *Med Educ*. 2008;42:800–809.
10. Ginsburg S, Gold W, Cavalcanti R, et al. Competencies "plus": the nature of written comments on internal medicine residents' evaluation forms. *Acad Med*. 2011;86(suppl 10):S30–S34.
11. Lurie S, Mooney C, Lyness J. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. *Acad Med*. 2009;84:301–309.
12. Hatala R, Sawatsky AP, Dudek N, et al. Using in-training evaluation report (ITER) qualitative comments to assess medical students and residents: a systematic review. *Acad Med*. 2017;92(6):868–879.
13. 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.
14. 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.
15. Beeson MS, Holmboe ES, Korte RC, et al. Initial validity analysis of the emergency medicine milestones. *Acad Emerg Med*. 2015;22:838–844.
16. 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.
17. Park Y, Zar F, Norcini J, Tekian A. Competency evaluations in the next accreditation system: contributing to guidelines and implications. *Teach Learn Med*. 2016;28:135–145.

18. 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. *J Am Med Assoc*. 2016;316:2253–2262.
19. 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.
20. Li ST, Tancredi DJ, Schwartz A, et al. Competent for unsupervised practice: use of Pediatric Residency training milestones to assess readiness. *Acad Med*. 2017;92:385–393.
21. Gardner AK, AbdelFattah K. Comparison of simulation-based assessments and faculty ratings for general surgery resident milestone evaluation: are they telling the same story? *Am J Surg*. 2017;214:547–553.
22. Battistone M, Pendleton B, Milne C, et al. Global descriptive evaluations are more responsive than global numeric ratings in detecting students' progress during the inpatient portion of an internal medicine clerkship. *Acad Med*. 2001;76(suppl 10):S105–S107.
23. Durning S, Hanson J, Gilliland W, et al. Using qualitative data from a program director's evaluation form as an outcome measurement for medical school. *Mil Med*. 2010;175:448–452.
24. Frohna A, Stern D. The nature of qualitative comments in evaluating professionalism. *Med Educ*. 2005;39:763–768.
25. Pusic MV, Boutis K, Hatala R, Cook DA. Learning curves in health professions education. *Acad Med*. 2015;90:1034–1042.
26. Pusic MV, Boutis K, Pecaric MR, et al. A primer on the statistical modelling of learning curves in health professions education. *Adv Health Sci Educ Theory Pract*; 2016. https://doi.org/10.1007/s10459-016-9709-2.
27. Charmaz K. Reconstructing theory in grounded theory studies. In: *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. London, UK: SAGE; 2006:123–151.
28. Ginsburg S, Regehr G, Lingard L, Eva K. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Educ*. 2015;49:296–306.