MEDICAL EDUCATION ONLINE
2025, VOL. 30, 2471434
https://doi.org/10.1080/10872981.2025.2471434

RESEARCH ARTICLE

OPEN ACCESS

Narrative comments in internal medicine clerkship evaluations: room to grow

Christine Crumbley[a], Karen Szauter[b,c], Bernard Karnath[c], Lindsay Sonstein[c], L. Maria Belalcazar[c] and Sidra Qureshi[c]

[a]Department of Family & Community Medicine, Baylor College of Medicine, Houston, TX, USA; [b]Educational Affairs, John Sealy School of Medicine, The University of Texas Medical Branch, Galveston, TX, USA; [c]Department of Internal Medicine, University of Texas Medical Branch, Galveston, TX, USA

ABSTRACT

The use of narrative comments in medical education poses a unique challenge: comments are intended to provide formative feedback to learners while also being used for summative grades. Given student and internal medicine (IM) grading committee concerns about narrative comment quality, we offered an interactive IM Grand Rounds (GR) session aimed at improving comment quality. We undertook this study to determine the quality of comments submitted by faculty and post-graduate trainees on students' IM Clerkship clinical assessments, and to explore the potential impact of our IM-GR. Archived comments from clerkship cohorts prior to and immediately following IM-GR were reviewed. Clinical clerkship assessment comments include three sections: Medical Student Performance Assessment (MSPE), Areas of Strength, and Areas for Improvement. We adapted a previously published comment assessment tool and identified the performance domain(s) discussed, inclusion of specific examples of student performance, evidence that the comment was based on direct observations, and, when applicable, the inclusion of actionable recommendations. Scoring was based on the number of domains represented and whether an example within that domain was provided (maximum score = 10). Analysis included descriptive statistics, t-test, and Pearson correlation coefficients. We scored 697 comments. Overall, section ratings were MSPE 2.51 (SD 1.52, range 0–9), Areas of Strength 1.53 (SD 1.09, range 0–6), and Areas for Improvement 1.27 (SD 1.06, range 0–8). Significant differences were noted after Grand Rounds only in the MSPE mean scores. Within domains, trends toward increased use of specific examples in the post-GR narratives were noted. Assessment of both the breadth and depth of the included comments revealed low-quality narratives offered by our faculty and resident instructors. A focused session on best practices in writing narratives offered minimal change in the overall narrative quality, although we did notice a trend toward the inclusion of explanative examples.

ARTICLE HISTORY
Received 29 October 2024
Revised 30 December 2024
Accepted 18 February 2025

KEYWORDS
Narrative comments; internal medicine; clerkship clinical assessment; medical student; educational assessment

Introduction

Narrative comments serve as a qualitative assessment, allowing instructors to provide contextualized written statements about trainees and their performance during specific tasks or clinical encounters [1,2]. Narrative comments can document observations, benchmark clinical performance, and integrate and synthesize information about trainees' performance [3–5]. They are frequently used to provide feedback, to conduct assessments, and to make decisions about promotion to the next stage of training. Narrative comments can correlate with numeric ratings [6–8] and serve as indicators of progress and future performance [6,7,9–11], while also having the potential to capture elements of professional development not reflected in numeric rating scales [12]. Program directors, faculty, and even senior residents can reliably rank order residents using only narrative comments [6,10,13–15]. Additionally, narrative comments may identify struggling learners earlier than numeric ratings [15,16], particularly for deficits in technical skills, applied knowledge, and professionalism [17], and serve an important role in identifying learners in need of remediation [11].

Trainees appreciate narrative comments to guide their professional growth [6,11,18–21]. Specific characteristics important to perceived comment quality have been identified. High-quality narrative comments are offered in a timely manner and include specificity with examples of behaviors or skills that evaluators would like to reinforce or to correct, ideally based on direct observation [2,22,23]. These behaviors and skills should be compared to an expected standard to allow trainees to identify their strengths and areas for growth. When learning needs are identified by instructors, they should be

CONTACT Karen Szauter kszauter@utmb.edu Educational Affairs, John Sealy School of Medicine, The University of Texas Medical Branch Institution, 301 University Blvd, Galveston, TX 77555-0413, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/10872981.2025.2471434

prioritized and include actionable recommendations for improvement, ideally with suggested resources for study. High-quality narratives are written in complete sentences using first-person language and appropriate punctuation and avoid the use of negative politeness strategies, such as conventional indirectness, hedging, and impersonalization [24,25].

Despite the established value of narratives, the literature has shown that narrative comments are frequently absent or of low quality on trainees' evaluations. Studies have shown that narrative comments are missing from 7% to 34% of clerkship evaluations [26–29] and from 21% to 70% of residency evaluations [1,20,21,30–33]. Low-quality comments include remarks about personality, personal characteristics and attributes, and other nonbehavioral or nonspecific content [20,21,27,29,34,35].

Furthermore, low-quality comments lack critical feedback or actionable recommendations for improvement. In studies involving medical students, less than a third of comments identified areas for improvement [27,28]. Narrative comments that are negative, generic, or otherwise ambiguous can be associated with lower clerkship grades [26] and being placed on academic probation [11]. However, it is possible that instructors are reluctant to document to corrective feedback on a permanent record [36,37]. As an example, Hemmer and colleagues (2000) noted that professionalism lapses were discussed at in-person evaluation sessions but not documented on evaluation checklists or in narrative comments [38].

Following discussions with students and our Internal Medicine (IM) grading committee, we had similar concerns about narrative comment quality submitted by instructors on our clerkship. Although episodic feedback to individual faculty or residents had been provided, the overall quality of our narrative comments appeared suboptimal. This prompted the presentation of an interactive Internal Medicine Grand Rounds (IM-GR). We offered a literature-based discussion of the importance of high-quality comments, including tools and tips for writing high-quality comments. Attendees were provided sample comments with a scoring rubric and asked to offer feedback on how the comment could be improved. This interactive portion of the presentation was intended to allow attendees to immediately apply the tools and tips offered. At the end of the presentation, pocket cards with guidelines on writing high-quality narrative comments were distributed. Cards were also available for pick up from the IM Medical Education Office after the presentation for remote attendees, and many people stopped by to obtain one. Evaluation of the IM-GR was overall positive. We organized this project following Grand Rounds to assess the potential impact of providing training and resources. The purpose of this study was twofold: (1) to objectively define the overall quality of narrative comments in our IM clerkship and (2) to determine whether our Grand Rounds presentation impacted comment quality.

Methods

Data sources

This study was approved by our Institutional Review Board and School of Medicine Education Research Committee. Archived clinical assessment forms from the IM clerkship were retrieved for this project. The web-based form includes numeric ratings for ten performance areas, broadly falling into four categories. Three free-text boxes for narrative comments appear at the end of the form: (1) the Medical Student Performance Evaluation (MSPE), (2) Areas of Strength, and (3) Areas for Improvement. Comments are required; the form cannot be submitted if a section is left blank. Instructors are told that the MSPE comments will be used verbatim for residency application materials, but that the latter two are for feedback to the students and the clerkship grading committee. For purposes of this study, comments were exported from the clinical assessment forms, and the three sections were grouped with a single unique identifier. Our clerkship clinical assessment form is available in the Supplemental Materials. [SUPPLEMENT A]

During the clerkship, students rotate on both general IM and subspecialty rotations. Clinical assessments are provided by instructors who work with the student for a minimum of one week (5 working days). Comments from three cohorts on the 8-week clerkship (5 June 2023, through 19 November 2023) were included. The IM-GR presentation was held on 28 September 2023, which aligned with the first week of the third cohort's rotation.

Comments were deidentified in preparation for study. Student name, instructor name, instructor status (faculty vs GME trainee), timing of the rotation, and duration of instructor exposure to the student were maintained in a separate secure file by the PI (KS). Our study aim did not explicitly include assessing comments for evidence of bias; therefore, pronouns were not removed from the comments. Comments, including the content of each of the three sections, were entered into an Excel sheet identified only by the study code, and stored in a password-protected file. The PI managed the master file, sharing only the comments assigned for review to individual study team members. For the selected clerkship cohorts whose comments were analyzed, student clerkship grades had already been assigned and finalized.

Scoring narrative comments

For purposes of this study, we developed a narrative comment scoring rubric modeled after the Narrative Evaluation Quality Instrument (NEQI) [39]. We chose to adapt the NEQI for our rubric after an extended review of the literature and review of other existing tools. Our primary interest was identifying the performance domain(s) covered by each comment, and whether an example of behavior was included. The NEQI identifies eight unique performance domains, including Overall Performance, Clinical Skills, Clinical Reasoning Skills, Prepares for and Participates in Patient Care Activities, Fund of Knowledge, Written and/or Oral Skills, Initiative, and Professionalism (interpersonal skills with patients/staff). As our aim was to focus on specific performance domains included in our comments, we removed the Overall Performance domain. We created a singular Skills domain, combining the Clinical Skills and the Written and/or Oral Skills from the NEQI, as we felt that communication skills and clinically oriented skills, such as physical exam, could be coded in a broader Skills domain. We also created an Attitudes and Behavior domain, incorporating concepts from the Prepares for and Participates in Patient Care Activities, Initiative, and Professionalism domains on the NEQI. For purposes of our work, we felt that descriptors of these characteristics represented professional behaviors expected of medical trainees and could be categorized more broadly into a single domain. Prior to use, the study team reviewed the scoring rubric in detail for clarity. For comments that could not be mapped to a specific performance domain, such as non-responses (e.g., ‘none,’ ‘n/a,’) or a simple complimentary statement without additional information (e.g., ‘good job,’) we created a Generic category. Our full scoring tool is available in Supplemental materials. [SUPPLEMENT B].

We scored each comment based on domains using the following scale: 0 = domain not mentioned, 1 = domain mentioned without specific example, and 2 = domain mentioned with specific example (s). Table 1 provides examples of how a sentence within a comment would be scored. If the entire comment included only Generic language, it was scored 0. Each of the three comment sections were scored independently. The domain ratings for each section were summed for a maximum of 10 points (range of possible scores = 0–10).

Additionally, we added a section to identify whether learner performance level was compared to a standard. Guidelines for high-quality narratives support inclusion of comparison to an expected level of performance [23]; comparison to peers is not considered appropriate. We designated whether no comparison made, a comparison made to an expected standard (e.g., ‘at the level of a third-year medical student,’ ‘at the level of an intern,’ or ‘for this student’s level of training’) or whether a comparison to peers was included (e.g., ‘above their peers’ or ‘better than other students I’ve worked with’). We did not include a numeric rating for these in the score of the comment. For the MSPE comments, we added documented mention of direct observation (e.g., ‘I witnessed,’ ‘I observed,’ or similar) that explicitly stated that the instructor had directly observed the medical student’s performance. Finally, in the Areas for Improvement section, we noted whether the instructor included a specific actionable recommendation.

Our study team consisted of clinician educators in the Department of Internal Medicine and one fourth-year medical student (CC). The study team met to discuss the scoring rubric and rate sample comments. Following that session, a set of 10 comments were scored independently and the team reconvened to discuss differences. A second set of 10 comments were scored to further calibrate the team. During the study, each comment was reviewed and scored by two team members. The MSPE comments were divided among the entire study team. The Areas of Strength and Areas for Improvement sections were scored by KS and CC,

Table 1. Examples of narrative comments: sample comments for each domain demonstrating how these were rated for the study.

Domain If not included, score = 0.	Mentioned (no example) Score = 1	Specific example Score = 2
Knowledge	Strong knowledge base	Presentation on rounds demonstrated excellent foundational knowledge – presented PPD induration sizes relative to household objects showing a keen ability apply knowledge and translate for easy interpretation by others
Skills	Good rapport with patients and families	Gathered relevant details during medical interview and physical examination with consistent, correct identification of abnormal physical findings
Reasoning	Assessment and plan well organized for patient	Presented prioritized differential diagnosis – justification for choices based on the synthesis of key features from the patient interview, exam, and available laboratory data.
Attitudes and Behaviors	Great work ethic	Student consistently took on a caseload of 5 patients and managed all responsibilities with dedication, close attention to detail, and compassion.
Team Connection	Interacted well with team	Found an inefficiency in the electronic medical record and, after reviewing with team, took the initiative to contact IT. The update allowed the team to function more efficiently and collaboratively.

as these comments were generally shorter and less complex than MSPE comments. The independent rater scores for each comment were compared, and differences were resolved through discussion. A single final score was assigned to each section (MSPE, Areas of Strength, and Areas for Improvement) for each comment.

Analysis

Initial analysis involved descriptive statistics. To address the first part of our study question, we included all comments in our dataset. We calculated an overall score for each comment section. We examined the frequency with which individual domains were mentioned for each section and within a domain whether it was simply mentioned or included an example. We also examined the frequency of statements within the MSPE section that supported direct observation and the frequency with which an actionable recommendation was provided within the Areas for Improvement. Comments that included Generic remarks with no usable information (e.g., ‘great job’ or ‘N/A’) were scored as zero and were included in the analysis. Pearson correlation coefficients were calculated to assess whether the quality score of the MSPE comment aligned with the quality score of the Areas of Strength or Areas for Improvement sections for each comment.

The second part of our study question addressed the potential impact of IM-GR on comment quality. We sorted the comments into pre- and post-GR groups. We compared the comment scores, assessed which domains were included, and also assessed the frequency with which examples supported the comments in each domain. We further analyzed comments based on the level of the instructor (faculty or GME trainee) and the duration of the interaction between the instructor and learner (<7 days, 7–14 days, >14 days). T-test and ANOVA were used to analyze for differences.

Results

Overall analysis of narrative comment quality

We analyzed comments from each of the three sections of 697 student clinical clerkship assessments. Overall mean scores for each section (maximum score = 10) were MSPE: 2.51 (SD: 1.52, range 0–9); Areas of Strength: 1.53 (SD:1.09; range 0–6); and Areas for Improvement: 1.26 (SD 1.06, range 0–8). The mean number of domains included by section (maximum domain number = 5) were MSPE: 2.06 (SD 1.11, range 0–5); Areas of Strengths: 1.27 (SD 0.82; range 0–4); and Areas for Improvement: 1.00 (SD 0.76; range 0–4). The primary domains mentioned in each section differed (Table 2); MSPE and Areas of Strength comments highlighted the Attitudes and Behaviors domain, whereas Areas for Improvement comments focused on the Knowledge domain.

Inclusion of a statement verifying direct observation of student performance was rare, mentioned in only 23 (3.3%) of the MSPE comments. Assessors infrequently compared performance to an expected standard: MSPE (110, 15.78%), Areas of Strength (29, 4.16%), Areas for Improvement (9, 1.29%). Comparison to peers was unusual, noted in less than 5% of the MSPE comments. In the Areas for Improvement section, provided primarily for feedback to the learners, there were 84 comments (12.1%) indicating there were no areas in need of improvement. Overall, only 23 comments (3.3%) provided an actionable recommendation for the learner, such as a specific suggestion on how to enhance skills or a recommended resource to guide learning. Pearson correlation coefficients showed no relationship between the comment quality score in the MSPE section and the two other sections of the comments, suggesting that comment quality was not consistent across the different comment sections for a given assessor.

Analysis of narrative comment quality relative to Grand Rounds presentation

To address the second part of our study question, the potential impact of the Grand Rounds presentation

Table 2. Frequency of inclusion of specific domains by comment by section. Percent based on total number of comments (n = 697).

n (% of 697)	Knowledge	Skills	Reasoning	Attitudes and Behaviors	Team Connection	Generic only (no domain identified)
MSPE	189 (27.1)	324 (46.5)	140 (20.1)	563 (80.8)	218 (31.3)	42 (6.0) Non-evaluative: 7 Compliment: 35
Strengths	106 (15.2)	185 (26.5)	43 (6.17)	441 (63.3)	105 (15.1)	114 (16.4) Non-evaluative: 107 Compliment: 7
Improvement	313 (44.9)	160 (23.0)	153 (21.9)	67 (9.6)	6 (1.0)	175 (25.15) Non-evaluative: 159 Compliment: 16

For ‘generic’ comments.
Non-evaluative = no comment offered, or marked ‘N/A’ or ‘none’.
Compliment only (could not be mapped to a domain) = ‘great job’, ‘excellent’.

on the quality of the narratives, we sorted the data by the timing of the comments. Of the 697 comments, 484 (69.4%) were written prior to and 213 (30.6%) after the Grand Rounds presentation.

Table 3 shows the mean score for each comment section. Results by individual domains, and whether a specific example of performance or behaviors for the domain was provided, are also shown. The only significant change was in the overall score of the MSPE comments before vs after Grand Rounds, whereas no significant difference was seen in other analyses comparing the pre- and post-Grand Rounds data. However, trends toward improvement, specifically with the inclusion of examples, are seen.

While the mean quality scores changed minimally after the Grand Rounds, the distribution of the scores shifted (Figure 1), and the median score for the MSPE comments increased from 2 to 3 in the post-Grand Rounds period.

Additional analyses

The literature suggests that many factors may impact the quality of the narrative comments provided by instructors [40–45]. We further analyzed our data by

Table 3. Comments by domain for the pre- and post-Grand Rounds periods. Percent of total based on 484 pre and 213 post comments. Within the domain, percent of examples provided is based on the number of comments with that group. *t-test: significant different at $p < 0.05$.

	MSPE		Areas of Strength		Areas for Improvement	
	Pre ($n = 484$)	Post ($n = 213$)	Pre ($n = 484$)	Post ($n = 213$)	Pre ($n = 484$)	Post ($n = 213$)
Score mean (SD)	2.43 (1.48)	2.68 (1.60)*	1.49 (0.95)	1.62 (1.35)	1.23 (0.99)	1.31 (1.20)
# of domains mean (SD)	2.03 (1.10)	2.13 (1.13)	1.26 (0.76)	1.27 (0.94)	1.01 (0.75)	0.98 (0.77)
	Pre ($n = 484$)	Post ($n =$ 213)	Pre ($n = 484$)	Post ($n = 213$)	Pre ($n = 484$)	Post ($n = 213$)
Knowledge						
Total (% of group)	130 (26.9)	59 (27.6)	74 (15.3)	32 (15.0)	221 (45.7)	92 (43.2)
With example (% of comments in this domain)	17 (13.1)	12 (20.3)	7 (9.5)	7 (21.9)	31 (14.0)	14 (15.2)
Skills						
Total (% of group)	216 (44.6)	108 (50.7)	124 (25.6)	61 (28.6)	108 (22.3)	52 (24.4)
With example (% of comments in this domain)	50 (23.1)	21 (19.4)	15 (12.0)	17 (27.9)	42 (38.9)	36 (69.2)
Reasoning						
Total (% of group)	91 (18.8)	49 (23.0)	31 (6.4)	12 (5.6)	105 (21.7)	48 (22.5)
With example (% of comments in this domain)	18 (19.8)	18 (36.7)	12 (38.7)	6 (50.0)	23 (21.9)	13 (27.1)
Attitude and Behaviors						
Total (% of group)	394 (81.4)	169 (79.3)	311 (64.3)	130 (61.0)	53 (11.0)	14 (6.6)
With example (% of comments in this domain)	94 (23.9)	60 (35.5)	72 (23.2)	39 (30.0)	10 (18.9)	4 (28.6)
Team Connection						
Total (% of group)	150 (30.99)	68 (31.9)	69 (14.3)	36 (16.9)	4 (0.83)	2 (0.9)
With example (% of comments in this domain)	15 (10.0)	7 (10.3)	7 (10.1)	5 (13.9)	0 (0.00)	2 (100)
No domain included (Generic statements, e.g., ‘great job,’ ‘n/a,’ etc.)	25 (5.17)	17 (7.98)	67 (13.8)	47 (22.0)	120 (24.8)	55 (25.8)

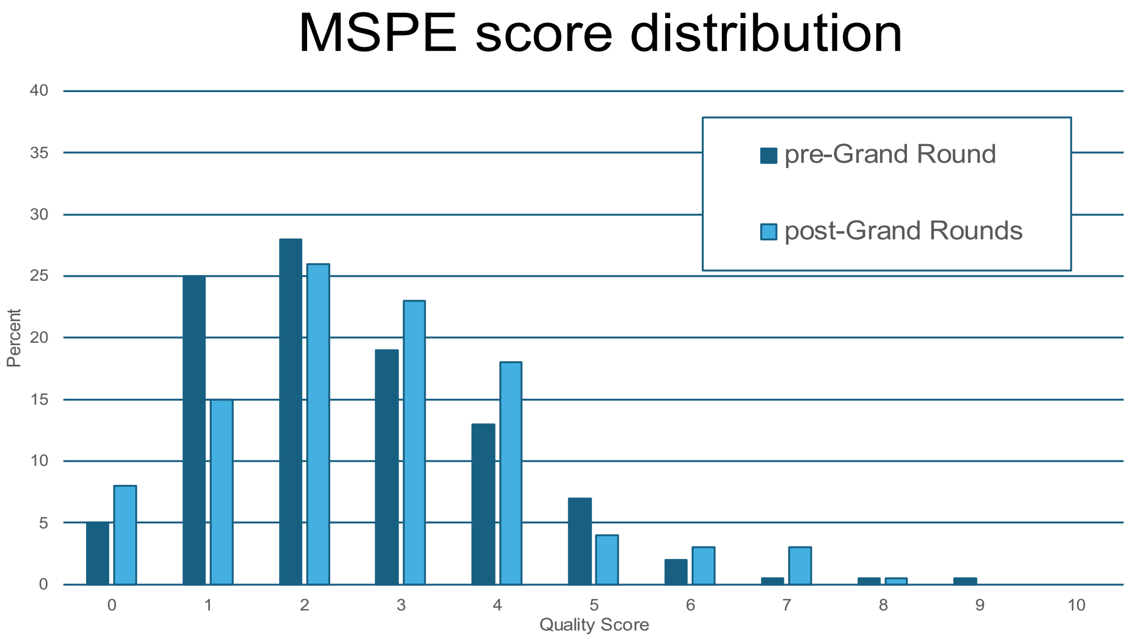

Figure 1. Distribution of MSPE section score before and after Grand Rounds.

Table 4. MSPE scores by instructor level of training and by duration of interaction.

	Pre-Grand Rounds		Post-Grand Round	
MSPE comments	n	Score (SD)	n	Score (SD)
Instructor level				
Resident/fellow Faculty	205	2.40 (1.45)	76	2.64 (1.53)
	279	2.45 (1.50)	137	2.70 (1.65)
Duration of exposure				
<7 days	102	2.11 (1.30)	54	2.46 (1.59)
7–14 days	226	2.37 (1.52)	85	2.69 (1.63)
>14 days	156	2.72 (1.48)	74	2.82 (1.59)

*Using t test for instructor level and ANOVA for duration, no statistical differences were found.

examining the level of the instructor (faculty vs. GME trainee), as well as the duration of the exposure between the instructor and learner. For these analyses, we focused only on the MSPE comments. Results are shown in Table 4.

Discussion

For this study, we set out to assess both the overall quality of the narrative comments in our Internal Medicine clerkship and to determine whether our Grand Rounds presentation impacted comment quality by comparing comments that were submitted for cohorts prior to and immediately following our presentation. For purposes of this work, we looked at the inclusion of key performance domains and the inclusion of examples of performance in defining quality. The initial review of our IM Clinical Assessment narratives confirmed our concerns that these comments overall were suboptimal. We found many comments to be limited to 1–2 domains, and specific examples were infrequent. After the Grand Rounds presentation, we saw early trends toward improvement, most notably more specific examples were available in the comments, suggesting that even a single training and resources on how to write high-quality narratives may be beneficial.

Our concern with low-quality narrative comments is the impact on both learners and their future training programs. The domain most addressed in our MSPE comments was Attitude and Behaviors, and these comments were consistently positive. While it is important for a future residency program director to be informed about characteristics including initiative, enthusiasm, and professionalism, focusing student descriptors on these without also describing the student's core knowledge, clinical skills, and clinical reasoning limits the value of this information. For students' professional growth, guidance on areas of strength and areas for improvement with tangible recommendations is essential. Many of our Areas for Improvement comments, mapped to the Knowledge domain, simply stated '*needs to read more*.' Our findings mirror those of many other studies on written narratives [20,29,35], supporting this as an area of healthcare education requiring ongoing attention, monitoring, and training. By offering specificity in comments, learners can continue to refine areas of strength and focus additional efforts in areas for growth.

Many reasons have been suggested for limitations in written narratives with time being a key consideration. Competing demands of patient care or research, regulatory and EMR documentation requirements for patient care, and other administrative tasks compete for the time and attention of instructors [40,41,46]. In addition, the number, length, and frequency of assessments to be completed can cause assessment fatigue [42,43]. We encourage faculty and residents to be deliberate in their observations of learners and to keep brief notes of skills and behaviors seen throughout their interactions. Such notes will help organize thoughts at the end of the rotation and may simplify offering quality comments.

Study limitations

Our study offers a snapshot of narrative quality from the IM core clerkship for third-year medical students. Working with data from a single institution, and from a single discipline, limits the generalizability of this work. Our assessment of the impact of the Grand Rounds presentation utilized data from the clerkship cohorts immediately prior to and following the intervention. The proximity of these cohorts makes the simple passage of time unlikely to account for any differences. Assessing comments from subsequent cohorts would offer additional insight into subsequent changes in comment quality and potential sustainability of any early trends.

Next steps

Our Clerkship Clinical Assessment form [Supplement A] includes descriptive anchors to guide the ratings of learners. Despite comments being entered after rating all ten sections, the specific language used to describe the behaviors and skills in each section was not seen in the written narratives. Although 'copy-paste' of these comments would not be ideal, the lack of specificity in the narratives suggests that instructors may be rating learners based on their own global assessment of the skill area (e.g., history taking) without considering the anchors provided. Additional work on the alignment between the numeric ratings and narratives will allow us to address this further.

As a continuation of this work, we have presented a second Grand Rounds on Narrative Comments. This session again included literature-based discussion and practical applications with debriefing. At the start of that session, we presented data from this initial work, in part to demonstrate the quality of comments overall and to suggest specific areas for improvement. We are preparing an online module for our faculty and housestaff which we intend to include as part of the annual training for all teaching faculty and trainees. We continue to monitor faculty and resident narratives in clerkship assessments and provide personalized feedback and recommendations when comments are poor. Ideally, offering periodic feedback to all faculty and trainees on the quality of comments will allow the skill of writing high-quality narratives to continue to develop and guide assessors to offer actionable recommendation to learners for growth [47]. This longitudinal ‘feedback on feedback’ approach has successfully improved narrative comment quality in other clerkships [48–50] and residency programs [51–54].

Acknowledgments

The authors recognize the contributions of Ni Zhang and Ella Tamayo for guidance in statistics and Kimberly Turner for her assistance with retrieval and formatting of information.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Ethical approval

This study was approved by the University of Texas Medical Branch Institutional Review Board on 6 February 2024, reference #:24–0019.

Previous presentations

Parts of this work have been presented at regional meetings as posters.

Texas Educators Academics Collaborative for Health Professions (TEACH-S) sub-analysis on the Areas for Improvements comments Galveston, TX, May 2024.

14th Annual Quality and Research Forum: a comparison of the sub-analyses for the Areas for Improvements and Areas of Strength in Galveston, TX, in June 2024.

ORCID

Karen Szauter http://orcid.org/0000-0002-2064-3535

References

[1] McGuire N, Acai A, Sonnadara RR. The McMaster narrative comment rating tool: development and initial validity evidence. Teach Learn Med. 2023;37(1):1–13. doi: 10.1080/10401334.2023.2276799

[2] Chakroun M, Dion VR, Ouellet K, et al. Quality of narratives in assessment: piloting a list of evidence-based quality indicators. Perspect Med Educ. 2023;12(1):XX–XX. doi: 10.5334/pme.925

[3] Cook DA, Kuper A, Hatala R, et al. When assessment data are words: validity evidence for qualitative educational assessments. Acad Med. 2016;91(10):1359–1369. doi: 10.1097/ACM.0000000000001175

[4] Hemmer PA, Dadekian GA, Terndrup C, et al. Regular formal evaluation sessions are effective as frame-of-reference training for faculty evaluators of clerkship medical students. J Gener Intern Med: JGIM. 2015;30(9):1313–1318. doi: 10.1007/s11606-015-3294-6

[5] Pangaro L. A new vocabulary and other innovations for improving descriptive in-training evaluations. Acad Med. 1999;74(11):1203–1207. doi: 10.1097/00001888-199911000-00012

[6] 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(10):1539–1544. doi: 10.1097/ACM.0b013e3182a36c3d

[7] Durning SJ, 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(6):448–452. doi: 10.7205/MILMED-D-09-00044

[8] Richards SH, Campbell JL, Walshaw E, et al. A multi-method analysis of free-text comments from the UK general medical council colleague questionnaires. Med Educ. 2009;43(8):757–766. doi: 10.1111/j.1365-2923.2009.03416.x

[9] Battistone MJ, 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(10 Suppl):S105–S107. doi: 10.1097/00001888-200110001-00035

[10] Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data. Acad Med. 2017;92(11):1617–1621. doi: 10.1097/ACM.0000000000001669

[11] Guerrasio J, Cumbler E, Trosterman A, et al. Determining need for remediation through postrotation

evaluations. J Grad Med Educ. 2012;4(1):47–51. doi: 10.4300/JGME-D-11-00145.1

[12] Ginsburg S, Watling CJ, Schumacher DJ, et al. Numbers encapsulate, words elaborate: toward the best use of comments for assessment and feedback on entrustment ratings. Acad Med. 2021;96(7S):S81–S86. doi: 10.1097/ACM.0000000000004089

[13] Ginsburg S, Regehr G, Lingard L, et al. Reading between the lines: faculty interpretations of narrative evaluation comments. Med Educ. 2015;49(3):296–306. doi: 10.1111/medu.12637

[14] Regehr G, Ginsburg S, Herold J, et al. Using "standardized narratives" to explore new ways to represent faculty opinions of resident performance. Acad Med. 2012;87(4):419–427. doi: 10.1097/ACM.0b013e31824858a9

[15] Cohen GS, Blumberg P, Ryan NC, et al. Do final grades reflect written qualitative evaluations of student performance? Teach Learn Med. 1993;5(1):10–15. doi: 10.1080/10401339309539580

[16] Ginsburg S, Gold W, Cavalcanti RB, et al. Competencies "plus": the nature of written comments on internal medicine residents' evaluation forms. Acad Med. 2011;86(10 Suppl):S30–S34. doi: 10.1097/ACM.0b013e31822a6d92

[17] Schwind CJ, Williams RG, Boehler ML, et al. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? Acad Med. 2004;79(5):453–457. doi: 10.1097/00001888-200405000-00016

[18] Duijin CCMA, Welink LS, Mandoki M, et al. Am I ready for it? Students' perceptions of meaningful feedback on entrustable professional activities. Perspect Med Educ. 2017;6(4):256–264. doi: 10.1007/S40037-017-0361-1

[19] Ferguson J, Wakeling J, Bowie P. Factors influencing the effectiveness of multisource feedback in improving the professional practice of medical doctors: a systematic review. BMC Med Educ. 2014;14(1):76–76. doi: 10.1186/1472-6920-14-76

[20] Jackson J, Kay C, Jackson WC, et al. The quality of written feedback by attendings of internal medicine residents. J Gen Intern Med. 2015;30(7):973–978. doi: 10.1007/s11606-015-3237-2

[21] Canavan C, Holtman MC, Richmond M, et al. The quality of written comments on professional behaviors in a developmental multisource feedback program. Acad Med. 2010;85(10):S106–S109. doi: 10.1097/ACM.0b013e3181ed4cdb

[22] Gulbas L, Guerin W, Ryder HF. Does what we write matter? Determining the features of high- and low-quality summative written comments of students on the internal medicine clerkship using pile-sort and consensus analysis: a mixed-methods study. BMC Med Educ. 2016;16(146):145–145. doi: 10.1186/s12909-016-0660-y

[23] Chakroun M, Dion VR, Ouellet K, et al. Narrative assessments in higher education: a scoping review to identify evidence-based quality indicators. Acad Med. 2022;97(11):1699–1706. doi: 10.1097/ACM.0000000000004755

[24] Branfield Day L, Rassos J, Billick M, et al. 'Next steps are …': an exploration of coaching and feedback language in EPA assessment comments. Med Teach. 2022;44(12):1368–1375. doi: 10.1080/0142159X.2022.2098098

[25] Ginsburg S, van der Vleuten C, Eva KW, et al. Hedging to save face: a linguistic analysis of written comments on in-training evaluation reports. Adv Health Sci Educ Theory Pract. 2016;21(1):175–188. doi: 10.1007/s10459-015-9622-0

[26] Plymale MA, Donnelly MB, Lawton J, et al. Faculty evaluation of surgery clerkship students: important components of written comments. Acad Med. 2002;77(10 Suppl):S45–S47. doi: 10.1097/00001888-200210001-00015

[27] Shaughness GBA, Georgoff PEMD, Sandhu GP, et al. Assessment of clinical feedback given to medical students via an electronic feedback system. J Surg Res. 2017;218:174–179. doi: 10.1016/j.jss.2017.05.055

[28] 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(1):123–123. doi: 10.1186/1472-6920-12-123

[29] Lye PS, Biernat KA, Bragg DS, et al. A pleasure to work with—an analysis of written comments on student evaluations. Ambul Pediatr: Off J Ambul Pediatr Assoc. 2001;1(3):128–131. doi: 10.1367/1539-4409(2001)001<0128:APTWWA>2.0.CO;2

[30] Tekian A, Park YS, Tilton S, et al. Competencies and feedback on internal medicine residents' end-of-rotation assessments over time: qualitative and quantitative analyses. Acad Med. 2019;94(12):1961–1969. doi: 10.1097/ACM.0000000000002821

[31] Gutierrez M, Wilson K, Bickford B, et al. Novel in-training evaluation report in an internal medicine residency program: improving the quality of the narrative assessment. J Med Educ Curric Dev. 2023;10:23821205231206058–23821205231206058. doi: 10.1177/23821205231206058

[32] Sebok-Syer SS, Klinger DA, Sherbino J, et al. Mixed messages or miscommunication? Investigating the relationship between assessors' workplace-based assessment scores and written comments. Acad Med. 2017;92(12):1774–1779. doi: 10.1097/ACM.0000000000001743

[33] Brucker K, Whitaker N, Morgan ZS, et al. Exploring gender bias in nursing evaluations of emergency medicine residents. Academic Emerg Med. 2019;26(11):1266–1272. doi: 10.1111/acem.13843

[34] Ringdahl EN, Delzell JE, Kruse RL. Evaluation of interns by senior residents and faculty: is there any difference? Med Educ. 2004;38(6):646–651. doi: 10.1111/j.1365-2929.2004.01832.x

[35] Zelenski AB, Tischendorf JS, Kessler M, et al. Beyond "read more": an intervention to improve faculty written feedback to learners. J Grad Med Educ. 2019;11(4):468–471. doi: 10.4300/JGME-D-19-00058.1

[36] Ginsburg S, van der Vleuten CPM, Eva KW, et al. Cracking the code: residents' interpretations of written assessment comments. Med Educ. 2017;51(4):401–410. doi: 10.1111/medu.13158

[37] Patel R, Drover A, Chafe R. Pediatric faculty and residents' perspectives on In-training evaluation reports (ITERs). Can Med Educ J. 2015;6(2):e41–e53. doi: 10.36834/cmej.36668

[38] Hemmer PA, Hawkins R, Jackson JL, et al. Assessing how well three evaluation methods detect deficiencies in medical students' professionalism in two settings of an internal medicine clerkship. Acad Med. 2000;75(2):167–173. doi: 10.1097/00001888-200002000-00016

[39] Kelly MS, Mooney CJ, Rosati JF, et al. Education research: the narrative evaluation quality instrument - development of a tool to assess the assessor. Neurology. 2020;94 (2):91–95. doi: 10.1212/WNL.0000000000008794

[40] Branch WT Jr, Kroenke K, Levison W. The clinician-educator—present and future roles. J Gen Intern Med. 1997;12(2):S1–S4. doi: 10.1046/j.1525-1497.12.s2.16.x

[41] Steinmann AFMD, Dy NMMD, Kane GCMD, et al. The modern teaching physician—responsibilities and challenges: an APDIM white paper. Am J Med. 2009;122 (7):692–697. doi: 10.1016/j.amjmed.2009.03.020

[42] McQueen SA, Petrisor B, Bhandari M, et al. Examining the barriers to meaningful assessment and feedback in medical training. Am J Surg. 2016;211(2):464–475. doi: 10.1016/j.amjsurg.2015.10.002

[43] Hauer KE, Nishimura H, Dubon D, et al. Competency assessment form to improve feedback. Clin Teach. 2018;15(6):472–477. doi: 10.1111/tct.12726

[44] Speer AJ, Solomon DJ, Fincher R-M. Grade inflation in internal medicine clerkships: results of a national survey. Teach Learn Med. 2000;12(3):112–116. doi: 10.1207/S15328015TLM1203_1

[45] Govaerts MJB, van der Vleuten CPM, Schuwirth LWT, et al. Broadening perspectives on clinical performance assessment: rethinking the nature of In-training assessment. Adv Health Sci Educ. 2007;12(2):239–260. doi: 10.1007/s10459-006-9043-1

[46] Reddy ST, Zegarek MH, Fromme HB, et al. Barriers and facilitators to effective feedback: a qualitative analysis of data from multispecialty resident focus groups. J Grad Med Educ. 2015;7(2):214–219. doi: 10.4300/JGME-D-14-00461.1

[47] Hall AM, Gray A, Ragsdale JW. Making narrative feedback meaningful. Clin Teach. 2024 Oct;21(5): e13766. doi: 10.1111/tct.13766

[48] Bartlett M, Crossley J, McKinley R. Improving the quality of written feedback using written feedback. Educ Primary Care. 2017;28(1):16–22. doi: 10.1080/14739879.2016.1217171

[49] Mooney CJ, Powell SJ, Dahl S, et al. Education research: a long-term faculty development initiative improves specificity and usefulness of narrative evaluations of clerkship students. Neurol Educ. 2022;1(1): e200003. doi: 10.1212/NE9.0000000000200003

[50] Mooney CJ, Pascoe JM, Blatt AE, et al. Predictors of faculty narrative evaluation quality in medical school clerkships. Med Educ. 2022;56(12):1223–1231. doi: 10.1111/medu.14911

[51] Dudek NL, Marks MB, Bandiera G, et al. Quality in-training evaluation reports—does feedback drive faculty performance? Acad Med. 2013;88(8):1129–1134. doi: 10.1097/ACM.0b013e318299394c

[52] Nichols D, Kulaga A, Ross S. Coaching the coaches: targeted faculty development for teaching. Med Educ. 2013 May;47(5):534–535. doi: 10.1111/medu.12187

[53] Littlefield JH, Darosa DA, Paukert J, et al. Improving resident performance assessment data: numeric precision and narrative specificity. Acad Med. 2005 May;80 (5):489–495. doi: 10.1097/00001888-200505000-00018

[54] Ross S, Hamza D, Zulla R, et al. Development of and preliminary validity evidence for the EFeCT feedback scoring tool. J Grad Med Educ. 2022;14(1):71–79. doi: 10.4300/JGME-D-21-00602.1