Perspectives on Medical Education
Journal of the Netherlands Association of Medical Education
PME

An Analysis of Written and Numeric Scores in End-of-Rotation Forms from Three Residency Programs

ORIGINAL RESEARCH

LAUREN M. ANDERSON
KATHERINE M. WRIGHT
KATHLEEN ROWLAND
YOON SOO PARK
DEBORAH EDBERG
ARA TEKIAN

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

Introduction: End-of-Rotation Forms (EORFs) assess resident progress in graduate medical education and are a major component of Clinical Competency Committee (CCC) discussion. Single-institution studies suggest EORFs can detect deficiencies, but both grades and comments skew positive. In this study, we sought to determine whether the EORFs from three programs, including multiple specialties and institutions, produced useful information for residents, program directors, and CCCs.

Methods: Evaluations from three programs were included (Program 1, Institution A, Internal Medicine: n = 38; Program 2, Institution A, Anesthesia: n = 9; Program 3, Institution B, Anesthesia: n = 11). Two independent researchers coded each written comment for relevance (specificity and actionability) and orientation (praise or critical) using a standardized rubric. Numeric scores were analyzed using descriptive statistics.

Results: 4869 evaluations were collected from the programs. Of the 77,434 discrete numeric scores, 691 (0.89%) were considered "below expected level." 71.2% (2683/3767) of the total written comments were scored as irrelevant, while 3217 (85.4%) of total comments were scored positive and 550 (14.6%) were critical. When combined, 63.2% (n = 2379) of comments were scored positive and irrelevant while 6.5% (n = 246) were scored critical and relevant.

Discussion: <1% of comments indicated below average performance; >70% of comments scored irrelevant. Critical, relevant comments were least frequently observed, consistent across all 3 programs. The low rate of constructive feedback and the high rate of irrelevant comments are inadequate for a CCC to make informed decisions. The consistency of these findings across programs, specialties, and institutions suggests both local and systemic changes should be considered.

CORRESPONDING AUTHOR:
Lauren M. Anderson, PhD
1700 W Van Buren St, Suite 470, Chicago, IL, US
Lauren_anderson@rush.edu

TO CITE THIS ARTICLE:
Anderson LM, Rowland K, Edberg D, Wright KM, Park YS, Tekian A. An Analysis of Written and Numeric Scores in End-of-Rotation Forms from Three Residency Programs. *Perspectives on Medical Education*. 2023; 12(1): 497–506. DOI: https://doi.org/10.5334/pme.41

INTRODUCTION

Faculty assessment of resident performance during graduate medical education (GME) clinical rotations is expected as part of the supervisory role and remains a "pillar of any programmatic assessment system" (pg. 20) [1]. The Accreditation Council on Graduate Medical Education (ACMGE) requires continuous monitoring and assessment of resident performance for both formative (for learning) and summative (of learning) purposes. Performance must be documented, and resident data provided to the Clinical Competency Committee (CCC) [2] as well as to the Accreditation Data System. The CCC and program directors use information from these assessments to identify residents' clinical and academic trajectories using the ACGME milestones [3–4] Easily accessible assessment data are also key for residents to access as part of their own learning and feedback to improve their own performance.

Assessment data from End of Rotation Forms (EORFs) are a major component in both CCC discussions and feedback to trainees [3, 5–7]. EORFs also form the main basis for promotion and entrustment decisions of learners which programs and faculty rely. Studies evaluating the quality and effectiveness of EORFs have shown EORFs can identify struggling residents in some cases [1, 5, 8–11]. A 2016 study found EORF scores between problem residents, as identified by the CCC, and non-problem residents were significantly different across all Milestones [8]. Another study found that an EORF can identify problem residents by comparing the comments and scores of those who were put on academic warning or probation and those who were not [9].

Despite the wide prevalence of EORFs and their need in GME, previous studies have found EORFs are flawed at evaluating comprehensive resident performance. When providing numeric ratings, evaluators systematically fail to use the lower end of the scale, creating rating errors [12–15]. Rating errors such as range restriction (failure to use full scale), leniency or severity error (distributional errors, "*hawks and doves*"), correlational error (give similar ratings across regardless of actual performance) and halo error (inflation) are all commonly found in assessment [16]. A 2004 study evaluating a single surgery residency found 18% of residents in need of some kind of remediation never received a score of "good," "fair," or "poor" on their EORFs, receiving only scores of "very good" or "excellent" [10]. Like numeric scores, written feedback also has been found to be flawed. Written comments found on EORFs tend to be low-quality, meaning comments are irrelevant or lack specific examples and recommendations for improvement [9, 17–19].

It is expected that data from assessments can be used by residents to inform their own improvement, outside of the CCC deliberations. A study by Patel (2015) gathered faculty and resident perspectives on the use of EORFs found that faculty both believed that this form of assessment could be "*critical in moving residents forward in their learning*" but also the EORF is a tool to guide residents in their development. While faculty acknowledge the dual uses of these tools the residents feel that that they receive the form too late and therefore has little impact [20]. Another study found that raters took an average of 89 days to complete an EORF and scores with greater delays led to decreased score variability [8].

Despite their flaws, EORFs continue to be widely used [1, 12, 21–22]. The ACGME acknowledges the use of faculty assessment forms such as EORFs are "routine and necessary in virtually all GME programs" [1] (pg.18). In response to previously noted problems with EORFs, the ACGME recommends focusing "time and energy into faculty development rather than creation of 'new and better forms" [1] (pg.17). The ACGME requires EORF data [2] or similar rotation summary assessment data, although it also recommends against an "over-reliance" on it. Moreover, EORFs tend to use assessment items and language stemming from competency-based frameworks (e.g., ACGME Milestones) across different institutions, prompting a need for more comprehensive multisite studies, exploring similarities or differences in the use and inferences generated from EORFs at different institutions. A similar need is motivated also for the use of EORFs across specialties.

This study integrates these needs stemming from the field to better understand EORFs across multiple programs, in multiple specialties, and in multiple institutions, exploring their use in summative and formative assessment. By evaluating data from several programs in a cross-sectional manner, we intended to examine the overall landscape of EORF use in programs to attempt to determine whether the problems with EORFs were program-specific or more likely to reflect systemic issues. We sought to describe whether recent changes in recommendations for increased faculty development and changes in use of EORFs- whether adopted or not- were resulting in data likely to be useful for CCCs.

We build on previous work that focused on one program at one institution [11, 17, 19, 23–26] or studies solely focused on the written comments component [6, 27] by examining multiple specialties and types of institutions to explore whether the EORF is an effective method of assessing resident performance, given its prevalence in GME assessment.

METHODS

STUDY SETTING AND PARTICIPANTS

We examined assessment data from three residency programs. Program 1 represents an internal medicine (IM) program (resident n = 38; faculty n = 184) from a public, safety-net, urban hospital. Program 2 represents an anesthesia program (resident n = 9; faculty n = 58) from the same public hospital. Program 3 represents an anesthesia program (resident n = 11; faculty n = 57) at large academic suburban medical center. Faculty from each program included both evaluators from inside and outside the department that housed the program. External faculty from Program 3 were not uniquely identified.

Data were collected retrospectively. All End-of-Rotation Forms for a single class from the beginning of residency through graduation were included in the analysis. Our inclusion criteria included residents who completed the program. At Program 3, a four-year program, three residents started the program in year two and were included in the analysis.

DATA COLLECTION

All EORFs (numeric scores and written comments) completed by faculty from each program were extracted from the respective residency management systems. Only assessments classified as an EORF were included in the study; non-EORFs, such as an SCO (structured clinical observation) or a similar "on-demand" work-based assessment (WBA), were excluded. Peer evaluations and EORFs completed by evaluators other than faculty were excluded from this study. We redacted identifying information, including names or other information found in comments, prior to analysis. A total of 4,869 forms were collected from the three programs.

FACULTY DEVELOPMENT

Two of the programs conducted training on how to assess residents; one offering quarterly faculty development sessions focused on evaluation and feedback. No program offered any development on writing comments.

DATA ANALYSIS*: NUMERIC SCORES*

Program 1 used a three-point rating scale of "below level," "at level," and "above level" with additional option for "not observed" for all the EORFs, the number of numeric items ranged from 10–23 (plus comment boxes) depending on the rotation, Similarly, Program 3 used the same 3-point rating scale with a "N/A" option. The forms ranged in length from 5–17 numeric questions plus comment boxes. Program 2 used a yes, no or "not observed" scale; the number of numeric questions ranged from 14–39 items plus comments.

Descriptive statistics were generated using IBM SPSS Statistics for Windows, Version 26 (Armonk, NY: IBM Corp) for all numeric responses. Data were summarized for each year of training and rotation.

DATA ANALYSIS: *WRITTEN COMMENTS*

Written comments were extracted as part of the EORF and separated from the numeric scores. Both comments connected to individual questions and overall comments were included in the analysis. A total of 3,767 written comments were collected from the three programs.

Comments were analyzed for relevance and orientation using a previously developed quantitative coding rubric for the nature of feedback in narrative comments [23–24]. Using the coding rubric as a guide, each comment was assigned two scores, one for relevance and one for orientation using a 4-point coding schema.

A highly relevant (rated 4) comment included specific items that could be used to improve or sustain practice, help the CCC to make competence decisions, or help implement a learning plan. A relevant (rated 3) comment was helpful but lacked specifics or action. Irrelevant (rated 2) and highly irrelevant (rated 1) comments lacked specific details or often were a list of adjectives (smart, confident, hard worker, and so forth) without context.

Orientation scores reflect whether a comment was praise-oriented or critical/growth needed. Mixed orientation comments (both positive and critical language) were discussed between researchers and placed into one category. A comment was judged positive (rated 3 or 4) if the language was encouraging, complimentary, or identified a behavior to reinforce or continue, with more positive language being adjudicated as high praise. Critical comments (rated 2) were those with negative language or those that identified a behavior to change, stop, or where growth was necessary. More severe language or serious concerns were judged as very critical (rated 1). Additional descriptions are provided in supplemental material.

A sample of thirty comments were independently scored by two researchers for inter-rater calibration of the rubric. The results of the initial sample were then discussed and a consensus for each score reached. Both researchers independently scored the remaining comments for all three programs. All inconsistencies were discussed until agreement reached. The four-point scale was collapsed into the binary categories of "critical or positive" and "relevant or irrelevant" for some of the analysis.

RESULTS

NUMERIC SCORE DISTRIBUTION

Summary. A total of 4869 evaluations were collected from the three programs, yielding 77,434 total discrete numeric scores and 3767 total comments. Of the 77,434 total numeric data points, 691 items (0.89%) were scored as "below expectations/expected level" or "no." (Table 1)

Program 1. 184 faculty completed 1855 evaluations on 38 residents, resulting in 33,569 individual numeric data points over the three-year IM residency. Of these, only 39 (0.1%) were "below expected" level.

Program 2. The faculty evaluation scale used was "yes/no" or N/A for each question. The 732 faculty evaluations collected 20,579 data points over the 4-year training period. Faculty gave 201 "no" scores (1%).

Program 3. The program completed 2282 evaluations over the training program from faculty members, generating 23,286 individual data points. 1.9% of total scores (n = 451) were below expected level while 75.3% (n = 17,546) were at expected level.

WRITTEN COMMENT RELEVANCE AND ORIENTATION

Summary. In total, 4869 evaluations produced 3767 comments from the three programs. Overall, 28.8% (n = 1084) of the total comments were adjugated as relevant, while 71.2% (n = 2683) of the total comments were scored as irrelevant. 3217 (85.4%) of total comments were scored positive and 550 (14.6%) were judged critical.

Program 1. Faculty generated 2306 total comments, of which 654 (28.4%) were relevant, whereas 1652 (71.6%) of comments were deemed irrelevant. Of the 2306 comments, 2127 (92.2%), were positive and 179 (7.8%) were critical, as denoted in Table 2.

Program 2. Of the 558 comments generated over the four-year training program, 144 (25.8%) were relevant. 440 comments (78.9%) were positive, and 118 (21.1%) comments were critical.

		TOTAL EVALS	TOTAL NUMERIC DATA POINTS	BELOW EXPECTED LEVEL	AT EXPECTED LEVEL	ABOVE EXPECTED LEVEL	DID NOT OBSERVE/N/A*
Program 1	**Total**	**1855**	**33,569**	**39 (0.1%)**	**23,374 (69.6%)**	**10,156 (30.3%)**	**637**
	Year 1	523	10,149	8 (0.1%)	7280 (71.7%)	2861 (28.2%)	61
	Year2	670	11,730	17 (0.1%)	7923 (67.5%)	3790 (32.4%)	184
	Year 3	662	11,690	14 (0.1%)	8171 (69.9%)	3505 (30%)	392
		TOTAL EVALS	TOTAL NUMERIC DATA POINTS	NO	YES		DID NOT OBSERVE/N/A*
Program 2	**Total**	**732**	**20,579**	**201 (1%)**	**20,378 (99%)**		**1,938**
	Year 1	141	4252	41 (0.2%)	4211 (99.8%)		401
	Year 2	336	10,190	138 (1.4%)	10,052 (98.6%)		1,068
	Year 3	161	3799	12 (0.3%)	3787 (99.7%)		294
	Year 4	94	2338	10 (0.4%)	2328 (99.6%)		175
		TOTAL EVALS	TOTAL NUMERIC DATA POINTS	BELOW EXPECTED LEVEL	AT EXPECTED LEVEL	ABOVE EXPECTED LEVEL	DID NOT OBSERVE*
Program 3	**Total**	**2282**	**23,286**	**451 (1.9%)**	**17,546 (75.3%)**	**5289 (22.6%)**	**632**
	Year 1	110	1290	34 (2.6%)	841 (65.2%)	415 (32.2%)	5
	Year 2	730	8712	123 (1.4%)	6992 (80.3%)	1597 (18.3%)	280
	Year 3	758	6731	102 (1.5%)	5199 (77.2%)	1430 (21.3%)	259
	Year 4	684	6553	192 (2.9%)	4514 (68.9%)	1847 (28.2%)	88

Table 1 Distribution of numeric scores from end of rotation forms by program.

*N/A/Did not observe are not included in the totals.

		TOTAL EVALS	TOTAL COMMENTS	% OF RELEVANT COMMENTS (NUMBER)	% OF IRRELEVANT COMMENTS (NUMBER)	% OF POSITIVE COMMENTS (NUMBER)	% OF CRITICAL COMMENTS (NUMBER)
Program 1	**Total**	**1855**	**2306**	**28.4% (654)**	**71.6% (1652)**	**92.2% (2127)**	**7.8% (179)**
	Year 1	523	775	26.6% (206)	73.4% (569)	93.2% (722)	6.8% (53)
	Year 2	670	843	32.9% (277)	67.1% (566)	91% (767)	9% (76)
	Year 3	662	688	24.9% (517)	75.1% (171)	92.7% (638)	7.3% (50)
		TOTAL EVALS	**TOTAL COMMENTS**	**% OF RELEVANT COMMENTS (NUMBER)**	**% OF IRRELEVANT COMMENTS (NUMBER)**	**% OF POSITIVE COMMENTS (NUMBER)**	**% OF CRITICAL COMMENTS (NUMBER)**
Program 2	**Total**	**732**	**558**	**25.8% (144)**	**74.2% (414)**	**78.9% (440)**	**21.1% (118)**
	Year 1	141	126	23.8% (30)	76.2% (96)	79.4% (100)	20.6% (26)
	Year 2	336	173	20.8% (36)	79.2% (137)	84.4% (146)	15.6% (27)
	Year 3	161	129	27.9% (36)	72.1% (93)	81.4% (105)	18.6% (24)
	Year 4	94	130	32.3% (42)	67.7% (88)	68.5% (89)	31.5% (41)
		TOTAL EVALS	**TOTAL COMMENTS**	**% OF RELEVANT COMMENTS (NUMBER)**	**% OF IRRELEVANT COMMENTS (NUMBER)**	**% OF POSITIVE COMMENTS (NUMBER)**	**% OF CRITICAL COMMENTS (NUMBER)**
Program 3	**Total**	**2282**	**903**	**31.7% (286)**	**68.3% (617)**	**72% (650)**	**28% (253)**
	Year 1	110	49	14.3% (7)	85.7% (42)	85.7% (42)	14.3% (7)
	Year 2	730	289	37% (107)	63% (182)	73% (211)	27% (78)
	Year 3	758	341	29% (99)	70% (242)	72.4% (247)	27.6% (94)
	Year 4	684	224	32.6% (73)	67.4% (151)	67% (150)	33% (74)

Table 2 End of rotation form written comment orientation and relevance by program.

Percentage of relevant comments include both comments scored as "relevant" and "highly relevant"; Percentage of irrelevant comments includes both comments scores as "irrelevant" or "highly irrelevant"; Percentage of positive comments includes comments scored as "high praise" and "moderate praise"; Percentage of critical comments includes comments scored as "critical" or "very critical".

Program 3. Of the 903 faculty written comments, 31.7% (n = 286) of comments were relevant, and 72% (n = 650) were positive, with 28% (n = 253) having a critical orientation.

COMBINING ORIENTATION AND RELEVANCE

Comment orientation and relevance were collapsed to binary categories, leaving 4 categories for results: critical/relevant; critical/irrelevant; positive/relevant; positive/irrelevant.

In total, 3767 comments were produced from the three programs. Overall, 63.2% (n = 2379) of the comments were positive and irrelevant while 6.5% (n = 246) were critical and relevant.

Program 1. Of 2306 written comments, 1530 (66.4%) were positive but irrelevant while 597 (25.8%) were positive and relevant. A total of 179 (7.8%) comments were deemed critical, of which 57 (2.5%) were critical and relevant as delineated in Table 3.

Program 2. Of 558 total comments, 339 (60.8%) were positive and irrelevant. 144 comments were scored as relevant (25.8%), of which 101 (18.1%) were positive and relevant and 43 (7.7%) were critical and relevant.

Program 3. Of 903 total comments, 510 (56.5%) were positive and irrelevant while 140 (15.5%) were positive and relevant and 146 (16.1%) were critical and relevant.

DISCUSSION

The results of this study indicate the information obtained from End-of-Rotation Forms is not adequate for providing acceptable, constructive feedback to residents, program

	CRITICAL & RELEVANT	CRITICAL & IRRELEVANT	POSITIVE & RELEVANT	POSITIVE & IRRELEVANT
Program 1 N = 2306	2.5% n = 57	5.3% n = 122	25.8% n = 597	66.4% n = 1530
Program 2 N = 558	7.7% n = 43	13.4% n = 75	18.1% n = 101	60.8% n = 339
Program 3 N = 903	16.1% n = 146	11.9% n = 107	15.5% n = 140	56.5% n = 510
Combined Data N = 3767	6.5% n = 246	8.1% n = 304	22.2% n = 838	63.2% n = 2379

Table 3 End of rotation form written comment combined orientation and relevance by program.

directors, or program CCCs. Of the 77,434 discrete numeric scores less than 1% were considered "below expected level" while 63.2% written comments were irrelevant and positive. Just 6.5% of written comments were critical and relevant.

Refraining from using negative ratings has previously been described as "the path of least resistance [12]." The results of this study suggest evaluators do not use the lower part of the qualitative scale when providing summative feedback. . This study finds this problem is systemic, affecting 3 programs across all years of training, 2 specialties, and 2 institutions assessed in this study. Earlier studies have identified additional barriers including inadequate faculty development, guilt, concern for the resident's future, and institutional culture [28].

While residents value written comments [27] they are often not useful or actionable [17–18]. The majority of comments in this study were adjudicated as irrelevant (85%). This was true for both critical (55% of total critical comments) and positive (74% of total positive comments). This high proportion of irrelevant comments makes it harder for residents, program directors, and CCCs to find the relevant comments, and it further risks a reduction in the perceived validity of relevant comments. This study identified these assessment limitations of EORFs which should warrant further developments to the assessment methodology, given their wide use in GME.

Our study supports previous studies that found lack of negative data [10, 29] and adds additional information about relevance. Similar to findings by both Tekian (2019) and Raaum (2019) our study found that critical comments were more likely to be judged as relevant compared with positive comments, making them even more valuable for improving performance. However, there were far fewer of them overall, highlighting the need for further faculty development and further development of secure learning and teaching environments in GME.

Previous studies have enumerated flaws with EORFs, leading more recently towards more "on-demand" structured clinical observations (SCOs) which rely on specific and direct observations of learners with immediate feedback rather than at the end of the rotation. A recent study compared the effectiveness of written comments in a traditional EORF and a procedural performance assessment completed immediately after a case. They found that 58.3% of the procedural assessment comments and 10.7% of the EORF comments were considered effective [21]. Similar findings have been shown with written comments and overall expectations ratings when comparing EORFs to P-SCO [29]. Other studies have shown on-demand forms are similar to end-of-rotation forms in producing actionable feedback [30].

When taken with the results of this study, this suggests that one of the biggest gaps may be between the medical education specialists and the GME teachers and trainers. That is, a body of literature indicates the strength of SCO assessment methods. However, with continued barriers to implementation and pressure to produce an evaluation at the end of each rotation, the residents in this study continued to receive EORFs.

CONSEQUENCES

The CCC has clear expectations that assessment data will be useful [31], and CCC members report that written comments are important to identify potential concerns [5]. Without robust and accurate EORFs, the CCC does not have the information it needs to make appropriate recommendations about milestone progress and promotion. The low relevance of comments in this study makes written comments unlikely to be sufficient to help a CCC, a program director, or a resident use the comments to improve performance, identify strengths, or assess progress through milestones, especially if they must wade through a body of irrelevant comments to find them. Allowing a resident to continue training without identifying areas of concern can ultimately result in harm

to patients, especially if a resident is entrusted with a level of care for which they are not clinically prepared [32–33].

WHAT'S NEXT FOR ASSESSMENT

As noted, solutions may initially focus on local-level changes, including additional formal faculty development. Many programs may require development at every level to effectively implement on-demand or SCO assessments alongside EORFs. Because they receive so little of it, residents may not have skills to manage constructive feedback, and may require coaching to contextualize a score of "below expectations" or a critically oriented comment as an opportunity to improve their performance as a physician.

Focusing on local faculty development alone may be most effective if the underlying problem is limited to individual programs. However, because our study includes programs from different specialties, different institutions, and different types of institutions (publicly funded/safety net and academic medical center funded), our findings suggest a systemic rather than localized problem. Therefore, enduring solutions may additionally require higher level changes, such as training of program directors or DIOs, as well as additional support and input from the ACGME or sponsoring institutions.

Ultimately, effective rotation forms may require graduate medical education to embrace honest, transparent, and constructive feedback. Previous literature has noted the challenges associated with creating a learning environment conducive to constructive feedback [27, 34].

STRENGTHS

This study includes programs from different specialties, different institutions, and different types of institutions (public and academic). This allows us to draw conclusions about more than just a single program. Likewise, the inclusion of 3 programs offers a large dataset, with more than 4800 total evaluations and 3700 written comments available for analysis. This data includes residents over the full cycle of their training, either 3 or 4 years, depending on the residency length.

LIMITATIONS

This study was not designed to allow for program-level or resident-level inferences about the residents' outcomes. We do not know what proportion of residents required remediation for individual rotations or competencies. Although we are aware of residents leaving a program without completing training, we do not know the reasons for this. For these reasons, future studies with more robust contextual data, as well as studies designed to isolate evaluator and program effects may inform or improve generalizability. We did not examine individual resident outcomes, such as milestone achievement or progression, in response to EORF or other information. Studies in the future may incorporate methodologies that consider cross-classification of data (different subsets of faculty rating different learners), an approach more nuanced than traditional clustering or hierarchical techniques, thereby allowing for the unbalanced nature of the data to be better reflected in the inferences. In the current design, we felt that preserving the data structure in its independent form provides a conservative overview of EORFs, for which clustering and other data adjustment techniques would further refine through more robust estimation of trends.

Our study examined three programs. Although results are generally consistent, we did find differences among these programs. A larger study might allow even better characterization of the problems described here. Where data were missing or where faculty chose "N/A" or "unable to answer," the data were not included in the analysis. We did not perform any analyses to assess the implications of not including these data in our analysis.

FUTURE RESEARCH

The goal of GME is to train residents who are competent to practice independently. Assessments, including EORFs, are a significant component of determining whether residents achieve this standard of independence. Future research in this area may include an examination of whether a certain type of evaluator (e.g., core faculty vs non-core faculty; intra-departmental vs extra-departmental) affects the likelihood of a useful evaluation, as well as an analysis of whether certain competencies (i.e., patient care or systems-based practice) are associated with particular numeric scores or comment orientation or relevance. Studying the contributions of different forms and evaluators could lead to better understanding the value in continuing to use rotation evaluations. Finally, future research may help understand other assessment methods, such as frequent, short observation forms that more accurately and honestly describe resident performance.

ADDITIONAL FILE

The additional file for this article can be found as follows:

- **Supplemental Material.** Relevance and orientation coding scheme for EORF comments using the coding for nature of feedback in narrative comments rubric. DOI: https://doi.org/10.5334/pme.41.s1

PREVIOUS PRESENTATIONS

This study has not previously been presented in a peer reviewed setting.

ETHICS AND CONSENT

This study was approved by the institutional review board of John H. Stroger of Cook County (Study 19–200, 12/4/2019) and Rush University Medical Center (22010703-IRB01, 2/14/2022). Waivers were obtained from other participating institutions.

ACKNOWLEDGEMENTS

The authors wish to thank the residency programs and the offices of Graduate Medical Education for their participation in this study.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Lauren M. Anderson, PhD orcid.org/0000-0003-4616-971X
Department of Family and Preventive Medicine, Rush University, Chicago, Illinois, US
Kathleen Rowland, MD, MS orcid.org/0000-0002-1383-0279
Department of Family and Preventive Medicine, Rush University, Chicago, Illinois, US
Deborah Edberg, MD orcid.org/0000-0001-9654-4763
Department of Family and Preventive Medicine, Rush University, Chicago, Illinois, US
Katherine M. Wright, PhD, MPH orcid.org/0000-0001-5967-8156
Department of Family & Community Medicine, Northwestern University Feinberg School of Medicine, Chicago, Illinois, US
Yoon Soo Park, PhD orcid.org/0000-0001-8583-4335
Department of Medical Education, University of Illinois Chicago, Chicago, Illinois, US
Ara Tekian, PhD, MHPE orcid.org/0000-0002-9252-1588
Department of Medical Education, University of Illinois Chicago, Chicago, Illinois, US

REFERENCES

1. **Holmboe ES, Lobst WI.** The Assessment Guidebook. Chicago: ACGME. Published 2020. https://www.acgme.org/globalassets/pdfs/milestones/guidebooks/assessmentguidebook.pdf.
2. **ACGME.** Common Program Requirements 2022. Retrieved 1 July 2022. https://www.acgme.org/what-we-do/accreditation/common-program-requirements/.
3. **Llyod RB, Park YS, Tekian A, Marvin R.** Understanding assessment systems for clinical competency committee decisions: evidence from multisite study of psychiatry residency training programs. *Academic Psychiatry*. 2019; 44(6): 734–740. DOI: https://doi.org/10.1007/s40596-019-01168-x
4. **Holmboe ES.** Realizing the promise of competency-based medical education. *Academic Medicine*. 2015; 90(4): 411–413. DOI: https://doi.org/10.1097/ACM.0000000000000515
5. **Schumacher DJ, King B, Barnes MM,** et al. Influence of clinical competency committee review process on summative resident assessment decisions. *Journal of Graduate Medical Education*. 2018; 10(4): 429–437. DOI: https://doi.org/10.4300/JGME-D-17-00762.1
6. **Ginsburg S, Regehr G, Lingard L, Eva KW.** Reading between the lines: faculty interpretations of narrative evaluation comments. *Medical Education*. 2015; 49(3): 296–306. DOI: https://doi.org/10.1111/medu.12637
7. **Regan L, Cope L, Omron R, Bright L, Bayram JD.** Do end-of-rotation and end-of-shift assessments inform clinical competency committees'(CCC) decisions? *Western Journal of Emergency Medicine*. 2018; 19(1): 121–127. DOI: https://doi.org/10.5811/westjem.2017.10.35290
8. **Park YS, Zar FA, Norcini JJ, Tekian A.** Competency evaluations in the next accreditation system: contributing to guidelines and implications. *Teaching and Learning in Medicine*. 2016; 28(2): 135–145. DOI: https://doi.org/10.1080/10401334.2016.1146607
9. **Guerrasio J, Cumbler E, Trosterman A, Wald H, Brandenburg S, Aagaard E.** Determining need for remediation through post rotation evaluations. *Journal of Graduate Medical Education*. 2012; 4(1): 47–51. DOI: https://doi.org/10.4300/JGME-D-11-00145.1
10. **Schwind CJ, Williams RG, Boehler ML, Dunnington GL.** Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? *Academic Medicine*. 2004; 79(5): 453–457. DOI: https://doi.org/10.1097/00001888-200405000-00016
11. **Ginsburg S, Gold W, Cavalcanti RB, Kurabi B, McDonald-Blumer H.** Competencies "plus": the nature of written comments on internal medicine residents' evaluation forms. *Academic Medicine*. 2011; 86(10): S30–S34. DOI: https://doi.org/10.1097/ACM.0b013e31822a6d92
12. **Watling CJ, Kenyon CF, Schulz V, Goldszmidt MA, Zibrowski E, Lingard L.** An exploration of faculty perspectives on the in-training evaluation of residents. *Academic Medicine*. 2010; 85(7): 1157–1162. DOI: https://doi.org/10.1097/ACM.0b013e3181e19722

13. **Hanson JL, Rosenberg AA, Lane JL.** Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Frontiers in Psychology*. 2013; 4: 1–10. DOI: https://doi.org/10.3389/fpsyg.2013.00668
14. **Dudek NL, Marks MB, Regehr G.** Failure to fail: the perspectives of clinical supervisors. *Academic Medicine*. 2005; 80(10): S84–S87. DOI: https://doi.org/10.1097/00001888-200510001-00023
15. **Williams RG, Klamen DA, McGaghie WC.** Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*. 2003; 15(4): 270–292. DOI: https://doi.org/10.1207/S15328015TLM1504_11
16. **Pangaro LN, Durning SJ, Holmboe ES.** Evaluation frameworks, forms, and global rating scales. In: Holmboe ES, Durning SJ, Hawkins RE (eds.), *Practical guide to the evaluation of clinical competence*, 2nd edition. Philadelphia: Elsevier; 2018. pp. 37–58.
17. **Jackson JL, Kay C, Jackson WC, Frank M.** The quality of written feedback by attendings of internal medicine residents. *Journal of General Internal Medicine*. 2015; 30(7): 973–978. DOI: https://doi.org/10.1007/s11606-015-3237-2
18. **Canavan C, Holtman MC, Richmond M, Katsufrakis PJ.** The quality of written comments on professional behaviors in a developmental multisource feedback program. *Academic Medicine*. 2010; 85(10): S106–S109. DOI: https://doi.org/10.1097/ACM.0b013e3181ed4cdb
19. **Raaum SE, Lappe K, Colbert-Getz JM, Milne CK.** Milestone implementation's impact on narrative comments and perception of feedback for internal medicine residents: A mixed methods study. *Journal of General Internal Medicine*. 2019; 34(6): 929–935. DOI: https://doi.org/10.1007/s11606-019-04946-3
20. **Patel R, Drover A, Chafe R.** Pediatric faculty and residents' perspectives on in-training evaluation reports (ITERs). *Canadian Medical Education Journal*. 2015; 6(2): e41. DOI: https://doi.org/10.36834/cmej.36668
21. **Ahle SL, Eskender M, Schuller M,** et al. The quality of operative performance narrative feedback: a retrospective data comparison between end of rotation evaluations and workplace-based assessments. *Annals of Surgery*. 2022; 275(3): 617–20. DOI: https://doi.org/10.1097/SLA.0000000000003907
22. **Luckoski J, Jean D, Thelen A,** et al. How do programs measure resident performance? A multi-institutional inventory of general surgery assessments. *Journal of Surgical Education*. 2021; 78(6): e189–195. DOI: https://doi.org/10.1016/j.jsurg.2021.08.024
23. **Tekian A, Borhani M, Tilton S, Abasolo E, Park YS.** What do quantitative ratings and qualitative comments tell us about general surgery residents' progress toward independent practice? Evidence from a 5-year longitudinal cohort. *The American Journal of Surgery*. 2019; 217(2): 288–295. DOI: https://doi.org/10.1016/j.amjsurg.2018.09.031
24. **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. *Academic Medicine*. 2019; 94(12): 1961–1969. DOI: https://doi.org/10.1097/ACM.0000000000002821
25. **Sebok-Syer SS, Klinger DA, Sherbino J, Chan TM.** Mixed messages or miscommunication? investigating the relationship between assessors' workplace-based assessment scores and written comments. *Academic Medicine*. 2017; 92(12): 1774–1779. DOI: https://doi.org/10.1097/ACM.0000000000001743
26. **Arkin N, Lai C, Kiwakyou LM,** et al. What's in a word? qualitative and quantitative analysis of leadership language in anesthesiology resident feedback. *Journal of Graduate Medical Education*. 2019; 11(1): 44–52. DOI: https://doi.org/10.4300/JGME-D-18-00377.1
27. **Ginsburg S, Van der Vleuten CP, Eva KW, Lingard L.** Cracking the code: residents' interpretations of written assessment comments. *Medical Education*. 2017; 51(4): 401–410. DOI: https://doi.org/10.1111/medu.13158
28. **Yepes-Rios M, Dudek N, Duboyce R, Curtis J, Allard RJ, Varpio L.** The failure to fail underperforming trainees in health professions education: A BEME systematic review: BEME Guide No. 42. *Medical Teacher*. 2016; 38(11): 1092–1099. DOI: https://doi.org/10.1080/0142159X.2016.1215414
29. **Young JQ, Lieu S, O'Sullivan P, Tong L.** Development and initial testing of a structured clinical observation tool to assess pharmacotherapy competence. *Academic Psychiatry*. 2011; 35: 27–34. DOI: https://doi.org/10.1176/appi.ap.35.1.27
30. **Marcotte L, Egan R, Soleas E,** et al. Assessing the quality of feedback to general internal medicine residents in a competency-based environment. *Canadian Medical Education Journal*. 2019; 10(4): e32. DOI: https://doi.org/10.36834/cmej.57323
31. **Ekpenyong A, Baker E, Harris I,** et al. How do clinical competency committees use different sources of data to assess residents' performance on the internal medicine milestones? A mixed methods pilot study. *Medical Teacher*. 2017; 39(10): 1074–1083. DOI: https://doi.org/10.1080/0142159X.2017.1353070
32. **Hauer KE, Ciccone A, Henzel TR,** et al. Remediation of the deficiencies of physicians across the continuum from medical school to practice: a thematic review of the literature. *Academic Medicine*. 2009; 84(12): 1822–1832. DOI: https://doi.org/10.1097/ACM.0b013e3181bf3170
33. **Roberts NK, Williams RG.** The hidden costs of failing to fail residents. *Journal of Graduate Medical Education*. 2011; 3(2): 127. DOI: https://doi.org/10.4300/JGME-D-11-00084.1
34. **Ramani S, Könings KD, Mann KV, Pisarski EE, Van der Vleuten CP.** About politeness, face, and feedback: Exploring resident and faculty perceptions of how institutional feedback culture influences feedback practices. *Academic Medicine*. 2018; 93(9): 1348–1358. DOI: https://doi.org/10.1097/ACM.0000000000002193

TO CITE THIS ARTICLE:
Anderson LM, Rowland K, Edberg D, Wright KM, Park YS, Tekian A. An Analysis of Written and Numeric Scores in End-of-Rotation Forms from Three Residency Programs. *Perspectives on Medical Education.* 2023; 12(1): 497–506. DOI: https://doi.org/10.5334/pme.41

Submitted: 21 October 2022 **Accepted:** 24 October 2023 **Published:** 03 November 2023