Novel In-Training Evaluation Report in an Internal Medicine Residency Program: Improving the Quality of the Narrative Assessment

Journal of Medical Education and Curricular Development
Volume 10: 1–6
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23821205231206058
Sage

Marc Gutierrez[1], Kelsey Wilson[1], Brant Bickford[1], Joseph Yuhas[1], Ronald Markert[2] and Kathryn M Burtson[3]

[1]Internal Medicine Program, Affiliated with Wright Patterson AFB and Wright State University, Wright-Patterson AFB, OH, USA. [2]Department of Internal Medicine and Neurology, Affiliated with Wright State University, Dayton, OH, USA. [3]Internal Medicine Program, Affiliated with Wright Patterson AFB, Boonshoft School of Medicine and Wright State University, Wright-Patterson AFB, OH 45433, USA.

ABSTRACT

OBJECTIVE: To determine whether incorporating our novel in-training evaluation report (ITER), which prompts each resident to list at least three self-identified learning goals, improved the quality of narrative assessments as measured by the Narrative Evaluation Quality Instrument (NEQI).

METHODS: A total of 1468 narrative assessments from a single institution from 2017 to 2021 were deidentified, compiled, and sorted into the pre-intervention form arm and post-intervention form arm. Due to limitations in our residency management suite, incorporating learning goals required switching from an electronic form to a hand-deliver form. Comments were graded by two research personnel utilizing the NEQI's scale of 0–12, with 12 representing the maximum quality for a comment. The outcome of the study was the mean difference in NEQI score between the electronic pre-intervention period and paper post-intervention period.

RESULTS: The mean NEQI score for the pre-intervention period was 2.43 ± 3.34, and the mean NEQI score for the post-intervention period was 3.31 ± 1.71, with a mean difference of 0.88 ($p < 0.001$). In the pre-intervention period, 46% of evaluations were submitted without a narrative assessment (scored as a zero) while 1% of post-intervention period evaluations had no narrative assessment. Internal consistency reliability, as measured by Ebel's intraclass correlation coefficient (ICC), showed high agreement between the two raters (ICC = 0.92).

CONCLUSIONS: Our findings suggest that implementing a timely, hand-delivered paper ITER that incorporates resident learning goals can lead to overall higher-quality narrative assessments.

KEYWORDS: narrative assessment, clinical competence, in-training evaluation report

RECEIVED: May 9, 2022. **ACCEPTED:** September 20, 2023

TYPE: Original Research

FUNDING: The authors received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Marc Gutierrez, Internal Medicine Program, Affiliated with Wright Patterson AFB and Wright State University, 4881 Sugar Maple Drive, Wright-Patterson AFB, OH 45433, USA.
Emails: marcrgutierrez.13@gmail.com; marc.r.gutierrez.mil@health.mil

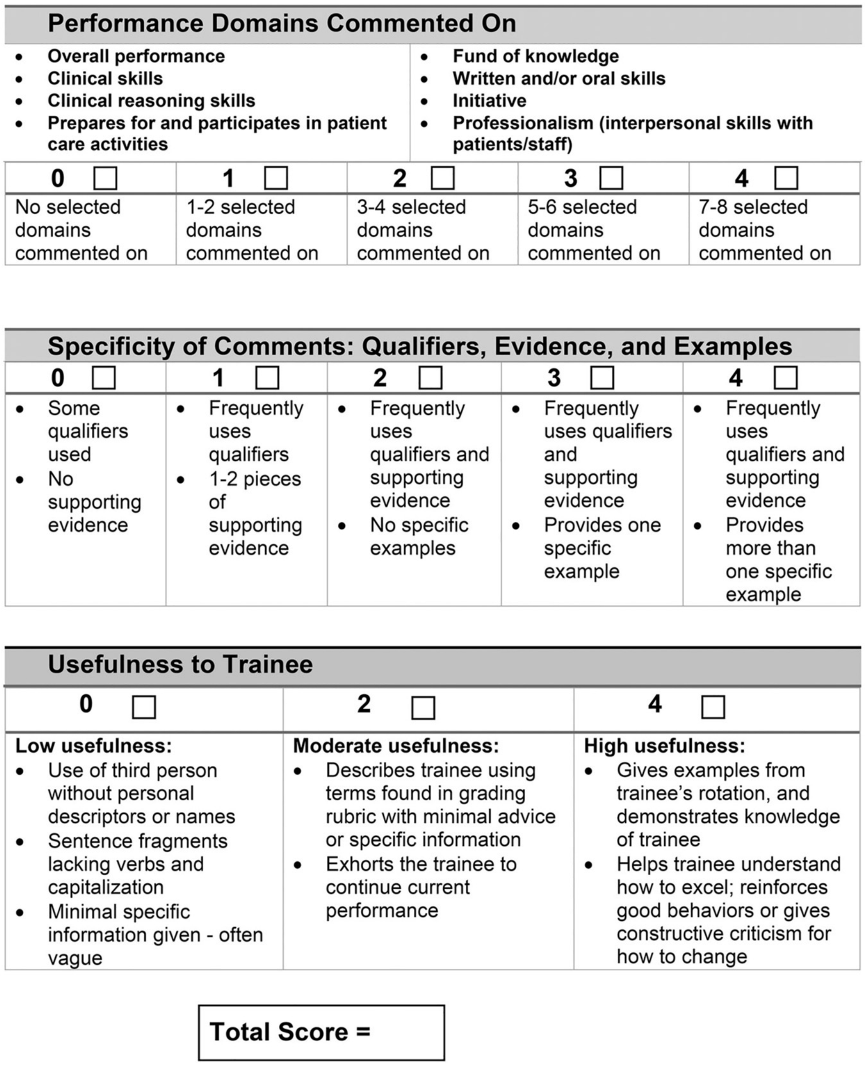
Introduction

The prevailing model of competency-based graduate medical education requires defensible, evidence-based evaluation of performance. One tool that supervising faculty can use to document resident performance in clinical rotations is an in-training evaluation report (ITER). ITERs usually consist of a series of numeric ratings of various competencies (ie, "demonstrates appropriate medical knowledge"), a global rating, and a section for a narrative assessment.[1]

Most published evidence regarding the validity of ITERs focuses on the numeric ratings of competencies. Literature demonstrates narrative-based evaluations can be a valid assessment with better reliability than numeric scores for grading and ranking purposes[1,2] and can address important resident attributes beyond standard Accreditation Council for Graduate Medical Education (ACGME) competencies.[3] There exists a practical need for a simple, manually-scored tool to assess the quality of narrative assessments on resident ITERs.

The Narrative Evaluation Quality Instrument (NEQI; Figure 1) rates narrative assessments on three dimensions—the number of performance domains commented on, the specificity of comments, and the usefulness to the trainee.[2] First introduced at the University of Rochester, the NEQI was reliable and valid when used in a neurology clerkship.[2] Importantly, the "usefulness to the trainee" domain draws on previous learner-focused research on the helpfulness of narrative assessments. Gulbas et al found that helpful comments shared characteristic qualities such as knowledge of the learner, specific examples of behavior to be reinforced or eliminated, and, importantly, correct grammar and punctuation.[5] The NEQI transforms "helpful" into a quantitative score and measures the helpfulness to the learner rather than relying on the subjective opinion of the scorer to determine what the learner would find helpful. The NEQI allows us to answer not only whether a narrative assessment was helpful, but also how helpful.

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).



Performance Domains Commented On

- **Overall performance**
- **Clinical skills**
- **Clinical reasoning skills**
- **Prepares for and participates in patient care activities**
- **Fund of knowledge**
- **Written and/or oral skills**
- **Initiative**
- **Professionalism (interpersonal skills with patients/staff)**

0 □	**1** □	**2** □	**3** □	**4** □
No selected domains commented on	1-2 selected domains commented on	3-4 selected domains commented on	5-6 selected domains commented on	7-8 selected domains commented on

Specificity of Comments: Qualifiers, Evidence, and Examples

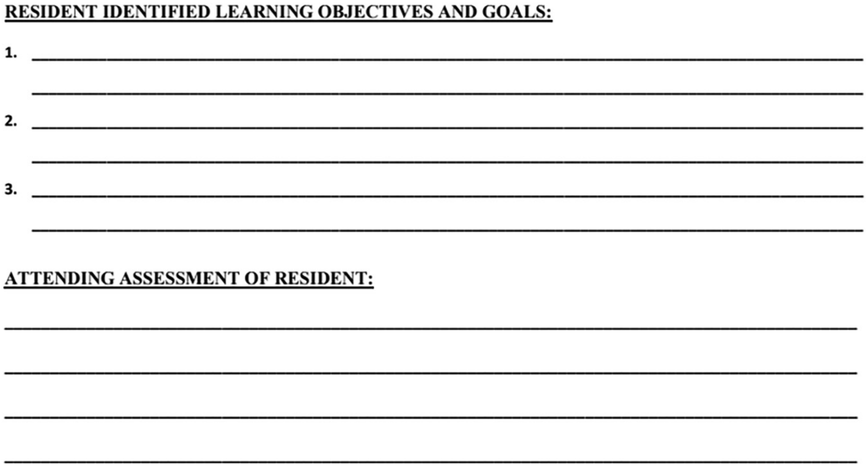
0 □	**1** □	**2** □	**3** □	**4** □
• Some qualifiers used • No supporting evidence	• Frequently uses qualifiers • 1-2 pieces of supporting evidence	• Frequently uses qualifiers and supporting evidence • No specific examples	• Frequently uses qualifiers and supporting evidence • Provides one specific example	• Frequently uses qualifiers and supporting evidence • Provides more than one specific example

Usefulness to Trainee

0 □	**2** □	**4** □
Low usefulness: • Use of third person without personal descriptors or names • Sentence fragments lacking verbs and capitalization • Minimal specific information given - often vague	**Moderate usefulness:** • Describes trainee using terms found in grading rubric with minimal advice or specific information • Exhorts the trainee to continue current performance	**High usefulness:** • Gives examples from trainee's rotation, and demonstrates knowledge of trainee • Helps trainee understand how to excel; reinforces good behaviors or gives constructive criticism for how to change

Total Score =

Figure 1. Narrative evaluation quality instrument (NEQI). The NEQI is a quantitative, manually scored tool that is used to rate the quality of a narrative assessment on three dimensions—the number of performance domains commented on, the specificity of comments, and the usefulness to the trainee.



RESIDENT IDENTIFIED LEARNING OBJECTIVES AND GOALS:

1. ____
   ____
2. ____
   ____
3. ____
   ____

ATTENDING ASSESSMENT OF RESIDENT:

____

____

____

____

Figure 2. Our novel paper ITER. Our learners proactively commit to 3–5 individualized goals at the beginning of their 2- to 4-week clinical rotations. Subsequently, faculty reflect on the learner's individualized goals and incorporate them into a summative narrative assessment.

To improve the quality of our residency program's narrative assessments, we developed a novel ITER (Figure 2) incorporating self-set learner goals and goal progress feedback from faculty to replace our program's previous ITER. Before our intervention, the narrative assessment portion of our program's ITER consisted of a simple blank text comment box on an electronic form. Due to limitations in our residency management suite, we were unable to develop an electronic version of our novel ITER that allowed learners to set learning goals before sending them to faculty for feedback. Additionally, one of our three clinical sites had limited compatibility with our residency management suite. To allow learners to set goals and to ensure equitable implementation at all three of our residency sites, we decided to develop our novel ITER as a hand-delivered paper form. The result amounted to a bundle of interventions, consisting of a switch from electronic to paper format, timely hand delivery of evaluations to faculty by residents, and the incorporation of resident learning objectives to guide faculty feedback. In this retrospective study, we graded narrative assessments to investigate (a) if the NEQI could reliably measure the quality of ITER comments regarding resident performance and (b) if our novel paper ITER and the associated bundle of interventions improved the quality of narrative assessments.

Methods

Setting

This study was conducted using de-identified narrative assessments collected from the Wright State University/Wright-Patterson Air Force Base (WPAFB) Internal Medicine Residency Program over 4 years from 2017 to 2021. No other shifts in faculty evaluation of residents occurred during this time.

Study design

This dataset included two distinct subsets: (a) the electronic resident evaluation forms on the internet-based New Innovations database from academic years 2017–2018 and 2018–2019 and (b) the novel paper resident in-training evaluation forms maintained by the Program Coordinator in academic years 2019–2020 and 2020–2021. With the electronic forms, faculty received emails prompting them to complete assessments of each resident at the end of each 2- to 4-week rotation. Each electronic form included numeric competency-based metrics followed by a comment box for narrative feedback. With the paper evaluation forms, residents proactively committed to 3–5 individualized goals at the beginning of each clinical rotation. Residents were then required to hand-deliver each paper form to their attending faculty at the end of each 2- to 4-week clinical rotation. The paper form included a numeric competency-based metric followed by an area for summative narrative assessment. Faculty were asked to reflect on the resident-identified learner goals before writing the narrative assessment. Of note, the clinical rotations in both arms comprised inpatient ward rotations, ICU rotations, and all subspecialty rotations at all three of our clinical sites. Night float and outpatient primary care clinics are evaluated by a different system in our program and were not included in the study.

Before inclusion in our study, the administrative staff and research team screened the resident evaluations for evaluation forms that were duplicates, declined, or blank and removed them from the study. Evaluation forms that were annotated "NA," "I did not work with this trainee," or similar were also excluded. ITERs submitted with a completed numeric metric system but without narrative assessments were included and scored "0."

Scoring

After applying inclusion and exclusion criteria, two researchers independently graded each narrative assessment using the NEQI (Figure 1). Before grading, the researchers were all trained to use the NEQI during a faculty development session. Each of the four data subsets was scored by a unique combination of researchers to minimize bias. The 2017–2018 dataset was scored by KW and MG; 2018–2019 by KW and BB; 2019–2020 by BB and JY; and 2020–2021 by JY and MG. The NEQI requires each researcher to assign a score from 0 to 4 on each of three dimensions: "Performance Domains Commented On," "Specificity of Comments," and "Usefulness to the Trainee." For "Performance Domains Commented On," each researcher assigned a point value based on the number of performance domains commented on by the faculty member using the domains explicitly listed in the NEQI. For "Specificity of Comments," each researcher assigned a score based on the number of qualifiers, that is "advanced," "lacking," supporting evidence, that is, "treated supporting staff with respect," "was not prepared for rounds," and specific examples, that is, "her discharge summaries are outstanding," "very skilled with end-of-life discussions" used by the assessor. "Usefulness to the trainee" was scored from low to high usefulness based on the descriptions provided in the NEQI. The resultant scores are added together for a total score of between 0 and 12. When the two raters differed by 3 points or more, the researchers discussed their scoring of these narrative assessments until a consensus was reached on final scores that differed by less than 3 points.

Outcomes

To determine if our novel paper ITER improved the quality of narrative assessments, we evaluated the mean difference in NEQI score between the electronic pre-intervention period and the paper post-intervention period. To determine the reliability of the NEQI tool, the internal consistency reliability of NEQI scores was calculated.

Statistical analysis

The independent samples Mann–Whitney Test was used to compare the ITER means (using the NEQI) for the pre-intervention electronic comment period and the post-intervention novel paper comment period. The effect size was calculated using Cohen's d. Ebel's intraclass correlation coefficient (ICC)[9] was used to determine the internal consistency reliability (agreement) between the two raters.

IRB

The Institutional Review Board (IRB) at USAF–88 MDG/ Wright-Patterson Medical Center exempted the study from full IRB review by declaring it non-research and waived the requirement of Informed Consent on August 10, 2021.

Results

A total of 1468 comments were de-identified, compiled, and sorted into the pre-intervention arm (electronic resident evaluation form) and the post-intervention arm (our novel ITER; Figure 3). A total of 105 evaluation forms, all from the pre-intervention arm, were excluded. The remaining 1363 evaluations, 803 in the pre-intervention period and 560 in the post-intervention period were graded by two researchers as described above in Methods, *Scoring*. Of note. no post-intervention narrative assessments were rejected for illegible handwriting.

Table 1 shows that the mean NEQI score for the pre-intervention electronic period was 2.43 ± 3.34 and the mean score for the post-intervention paper period was 3.31 ± 1.71, with a mean difference of 0.88 ($p < 0.001$). The median NEQI score was 0.5 ± 4.5 for the pre-intervention period and 3.0 ± 2.5 for the post-intervention period. The calculated effect size was 0.3, indicating a medium effect size. In the pre-intervention period, 367 of 803 (46%) evaluations were submitted without a narrative assessment (scored as a zero) while 6 of 560 (1%) post-intervention period evaluations had no narrative assessment. The ICC showed high agreement among the two raters: 0.92 (95% CI = 0.91 to 0.94).

Discussion

Narrative assessment of clinical performance is a central element of medical education at all levels of training. Multiple studies support the assertion that the quality of narrative assessments has a direct impact on learning.[4] In a novel approach, our group utilized the NEQI to quantitatively evaluate the effect of an intervention on the quality of narrative assessments in a graduate medical education program. We found that implementation of our bundle of interventions—a switch from electronic to paper format, timely hand delivery of evaluations to faculty by residents, and incorporation of resident learning objectives to guide faculty feedback—was associated with a modest yet statistically significant increase in NEQI score as well as a simultaneous increase in the percentage of evaluations with completed narrative assessments from 54% to 99%. The use of the NEQI will allow us to track the quality of narrative assessments for future interventions and contribute to an iterative process of faculty development and program improvement.

Many physician trainees express a preference for narrative assessment over numeric scores.[5] Most studies on the quality of narrative assessments have been qualitative in design,[6] which provide useful insights but can be more difficult to track over time than quantitative data. It would be useful for program administrators to assign quantitative scores to narrative assessments, identify faculty with low scores, and provide targeted faculty development interventions. Dudek et al developed a quantitative scoring tool to rate narrative assessments of residents called the Completed Clinical Evaluation Report Rating (CCERR).[7] While reliable and valid in settings where appraised, the CCERR is resource-intensive, which may inhibit its use to examine larger amounts of data.[7,8] The same group developed a computer-based proxy-CCERR screening algorithm that estimates human CCERR scorers with high levels of reliability.[6] However, the NEQI is simple to complete, intuitive, and less resource-intensive than the CCERR or proxy-CCERR.[8]

Notably, the pre-intervention electronic form had 367 submissions with no comments (46%) while the post-intervention paper form had only 6 no comment submissions (1%). This difference may be attributable to multiple factors. First, our novel ITER, in contrast to the pre-invention ITER, elicits self-directed learner goals and directly prompts faculty to comment on their assessment of the learner. Secondly, the transition from electronic to paper forms affected the timing and logistics of the evaluation process. Our novel ITER requires residents to hand-deliver the evaluation to their preceptors, usually on the last day of the rotation. This physical interaction prompts evaluators to generate timely narrative assessments while they are readily able to recall their interactions with the resident over the preceding weeks. In contrast, the previous electronic evaluation system sent a notification to the preceptor after the rotation had ended. Without residents being physically present and with preceptor-resident interactions less retrievable from memory (sometimes weeks later), narrative assessments were less likely to be made. Interestingly, a 2022 follow-up study from the original developers of the NEQI found that "significantly higher narrative evaluation quality was associated with a shorter time to evaluation completion."[10] Consequently, the incorporation of learner goals in combination with the logistical changes to the evaluation process resulted in both more and higher quality narrative assessments.

While our intervention was associated with a statistically significant increase in NEQI score with medium effect size, the absolute difference in mean scores of 0.88 is of questionable practical significance given no threshold for meaningful

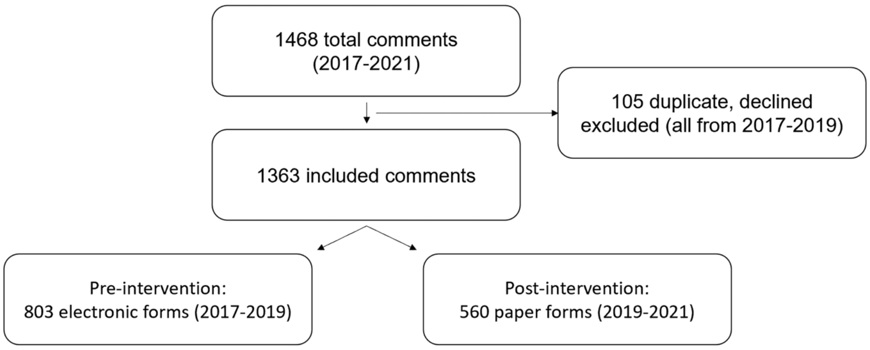


Figure 3. Flow diagram of comments in the study of the pre-intervention period and post-intervention period.

Table 1. Comparison between pre-intervention electronic ITER and post-intervention paper ITER.

NEQI score	Pre-intervention	Post-intervention
	N = 803	N = 560
Mean ± standard deviation	2.43 ± 3.34	3.31 ± 1.71
Median ± interquartile range	0.5 ± 4.5	3.0 ± 2.5

ITER = in-training evaluation report; NEQI = narrative evaluation quality instrument.

change in NEQI score has ever been established. Furthermore, the mean NEQI score of 3.31 for the quality of our novel paper narrative assessments was well below the standard for a comment being "moderately useful" (≥7)2 per Kelly et al, however, this threshold appears to be subjective. However, the overall low quality of our program's narrative assessments aligns with past research by Herbers et al, who found that narrative assessments in residency are of low quality or, at best, inconsistent,[11] and the modest impact of our intervention on narrative assessment likely underscores the need for more comprehensive interventions, including faculty development. Formal faculty training can increase the quality of feedback from narrative assessments,[12,13] and we interpret these findings as a useful baseline for further interventions rather than an overall assessment of low versus high quality.

Our study had limitations. First, the NEQI was developed by retrospective evaluation of narrative assessments from a single clerkship at a single institution,[2] and our use of this tool represents the first published study using the NEQI outside of the developer's institution. Additional study is needed to ensure the NEQI is generalizable and demonstrates valid evidence. Second, the incorporation of self-identified learning goals into our novel ITER necessitated the transition from an electronic to a paper medium. As described above, this transition increased the logistical burden of our residents and raised questions about sustainability. We are currently evaluating options to incorporate self-identified learning goals into an electronic residency management suite in a way that is feasible, equitable, and sustainable. Third, we used a locally constructed instrument (our novel paper ITER) to collect narrative assessments. However, the 0.92 ITER/NEQI internal consistency reliability indicates that faculty are likely to agree on a comment's quality from our paper ITER. Fourth, previous researchers have reported a positive correlation between the length of comments and the quality of written feedback.[4,5] Since our paper form's comment box limited the number of words while an electronic form is, in practical terms, open-ended, investigating the correlation between the number of words and the quality of comment would have been biased. Fifth, we did not report on or perform secondary analysis on NEQI subcomponent scores. Further analysis of subcomponent scores in future work may provide additional insights into narrative assessment quality. Sixth, the pre-invention and post-intervention periods comprised separate groups of residents and attending physicians, which may have biased results. Lastly, the researchers scoring the narrative assessments were not blinded to the conditions (ie, paper vs. electronic), which is a potential threat to validity.

Supported by the findings from this study, we plan to develop and implement faculty and resident development workshops designed to increase the quality of narrative assessments in our resident ITERs. More specifically, we plan to use the overall low NEQI scores as a catalyst to teach faculty how to increase the quality of their narrative assessments. We also intend to teach residents how to develop specific, measurable, achievable, realistic, and timely (SMART) learning goals to better guide the narrative assessments they receive from faculty. In future research, we intend to use the NEQI to examine whether interventions, such as these, result in improvements in the quality of our faculty's narrative assessments. Future educational research may benefit from examining the same resident cohort over a shorter time interval to avoid confounding and bias.

Conclusion

Narrative assessments serve both the educator and the learner by providing a versatile method that complements numeric evaluation scores. Application of the NEQI to ITERs can provide practical information on the reliability and quality of these assessments. Our findings suggest that implementation of a bundle of interventions including a switch from electronic to paper format, timely hand delivery of evaluations to faculty by residents, and incorporation of resident learning objectives to guide faculty feedback can improve the quality and consistency of narrative feedback, but will likely require further interventions to achieve practically meaningful results.

Acknowledgment

The authors would like to thank the academic faculty of Wright Patterson AFB and Wright State University Internal Medicine Program for their tireless support and dedication to our residents. The views expressed in this work are those of the authors and do not reflect the official policy or position of the Department of the Army, Navy, Air Force, Department of Defense, or the United States Government.

ORCID iDs

Marc Gutierrez https://orcid.org/0000-0001-6468-0971
Kathryn M Burtson https://orcid.org/0000-0003-1259-833X

REFERENCES

1. Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA. Using in-training evaluation report (ITER) qualitative comments to assess medical students and residents: a systematic review. *Acad Med*. 2017;92(6):868-879.
2. Kelly MS, Mooney CJ, Rosati JF, Braun MK, Thompson Stone R. Education research: the narrative evaluation quality instrument: development of a tool to assess the assessor. *Neurology*. 2020;94(2):91-95.
3. Tekian A, Park YS, Tilton S. Etal competencies and feedback on internal medicine Residents' End-of-rotation assessments over time: qualitative and quantitative analyses. *Acad Med*. 2019;94(12):1961-1969.
4. Canavan C. The quality of written comments on professional behaviors in a developmental multisource feedback program. *Acad Med*. 2010;85(10Supp):S106-S109.
5. 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.
6. 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;16:145.
7. Dudek NL, Marks MB, Wood TJ, Lee AC. Assessing the quality of supervisors' completed clinical evaluation reports. *Med Educ*. 2008;42(8):816-822.
8. Bismil R, Dudek NL, Wood TJ. In-training evaluations: developing an automated screening tool to measure report quality. *Med Educ*. 2014;48(7):724-732.
9. Ebel RL. Estimation of the reliability of ratings. *Psychometrika*. 1951;16(4):407-424.
10. Mooney CJ, Pascoe JM, Blatt AE, et al. Predictors of faculty narrative evaluation equality in medical school clerkships. *Med Educ*. 2022;56(12):1223-1231.
11. Herbers JE Jr., Noel GL, Cooper GS, Harvey J, Pangaro LN, Weaver MJ. How accurate are faculty evaluations of clinical competence? *J Gen Intern Med*. 1989;4(3):202-208.
12. Holmboe ES, Fiebach NH, Galaty LA, Huot S. Effectiveness of a focused educational intervention on resident evaluations from faculty a randomized controlled trial. *J Gen Intern Med*. 2001;16(7):427-434.
13. Warm E, Kelleher M, Kinnear B, Sall D. Feedback on feedback as a faculty development tool. *J Grad Med Educ*. 2018;10(3):354-355.