Education Research: The Narrative Evaluation Quality Instrument

Development of a tool to assess the assessor

Michael S. Kelly, MD, Christopher J. Mooney, PhD, Justin F. Rosati, MD, Melanie K. Braun, MD, and Robert Thompson Stone, MD

Neurology® 2020;94:91-95. doi:10.1212/WNL.0000000000008794

Correspondence
Dr. Stone
Robert_stone@urmc.rochester.edu

Abstract

Objective

Determining the quality of narrative evaluations to assess medical student neurology clerkship performance remains a challenge. This study sought to develop a tool to comprehensively and systematically assess quality of student narrative evaluations.

Methods

The Narrative Evaluation Quality Instrument (NEQI) was created to assess several components within clerkship narrative evaluations: performance domains, specificity, and usefulness to learner. In this retrospective study, 5 investigators scored 123 narrative evaluations using the NEQI. Inter-rater reliability was estimated by calculating interclass correlation coefficients (ICC) across 615 NEQI scores.

Results

The average overall NEQI score was 6.4 (SD 2.9), with mean component arm scores of 2.6 for performance domains (SD 0.9), 1.8 for specificity (SD 1.1), and 2.0 for usefulness (SD 1.4). Each component arm exhibited moderate reliability: performance domains ICC 0.65 (95% confidence interval [CI] 0.58–0.72), specificity ICC 0.69 (95% CI 0.61–0.77), and usefulness ICC 0.73 (95% CI 0.66–0.80). Overall NEQI score exhibited good reliability (0.81; 95% CI 0.77–0.86).

Conclusion

The NEQI is a novel, reliable tool to comprehensively assess the quality of narrative evaluation of neurology clerks and will enhance the study of interventions seeking to improve clerkship evaluation.

From the Department of Neurology (R.T.S., J.R., M.B.), University of Rochester School of Medicine and Dentistry (C.M., M.K.), NY.

Go to Neurology.org/N for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

Glossary

CI = confidence interval; **ICC** = intraclass correlation coefficient; **NEQI** = Narrative Evaluation Quality Instrument.

Supervising physicians typically complete in-training evaluation reports to document medical student performance on the neurology clerkship and other clinical rotations. In addition to providing personalized feedback, in-training evaluation reports are frequently used for summative assessment, which presents challenges.[1] Numeric scores may be overly simplistic, lacking the breadth and depth necessary to effectively document performance in complex clinical scenarios and across multiple contexts.[1,2] In contrast, recent work has shown that narrative-based evaluations can be used as a valid form of assessment,[3,4] providing equal or better reliability with regard to neurology clerkship grading and ranking.[4] Evidence also suggests that physician trainees have an expressed preference for narrative feedback over numeric scores.[5]

Despite evidence supporting a focus on narrative-based clerkship evaluation, reliability is contingent on assessors' abilities to complete high-quality narratives. Published work examining the quality of narrative evaluations is limited, and has largely focused on singular proxies or correlates of quality (e.g., overall word count), which may not reflect the complexity of narrative quality.[6–8] Given these limitations, this study sought to develop an instrument to comprehensively measure clerkship narrative quality with the goals of (1) informing faculty development around narrative quality improvement and (2) developing a quantitative outcome measure for education research.

Methods

Setting and participants

This retrospective study was conducted using data collected during the 2016–2017 academic year from third-year medical students rotating on the neurology clerkship at the University of Rochester School of Medicine and Dentistry. All students complete a 4-week neurology rotation and are evaluated by supervising faculty and residents (herein assessors) who complete in-training evaluation reports (including both numeric and narrative evaluations) at the rotation's completion. The evaluation includes narrative comments regarding clinical performance, personal and professional qualities, summary comments, and areas for improvement. The study was approved by the University of Rochester research subjects review board.

Creation, testing, and analysis of the Narrative Evaluation Quality Instrument (NEQI)

Development of the NEQI (figure 1) was informed by existing work in medical student performance assessment, which suggested the importance of 3 distinct components: (1) performance domains commented on, (2) specificity of comments, and (3) usefulness to trainee.[3,6,9,10] Select performance domains that correlated with a significant effect ($p < 0.05$) on preceptor grade in a study by Plymale et al.[9] were included in the NEQI (figure 1). For the specificity component arm, assessors' comments on student performance were classified as follows: qualifying statements (e.g., "strong physical exam skills"), statements citing evidence (e.g., "his physical examination was well organized and complete"), and inclusions of a specific example (e.g., "she picked up on key exam findings, including ataxia in a patient with cerebellar stroke"). Five scoring options were then created to reflect the overall quantity of these statements in each narrative (figure 1). The usefulness to trainee component arm was modeled after work by Gulbas et al.,[10] where evaluations were pile-sorted based on helpfulness into 4 distinct clusters. These were consolidated into 3 clusters after pilot testing suggested decreased reliability with more options. The clusters were as follows: "very useful" (e.g., gives examples, provides useful feedback on how to improve), "moderately useful" (e.g., uses full sentences, comments on multiple domains, without examples or specific areas for improvement), and "minimally useful" (e.g., use of third person, sentence fragments, vague information) (figure 1).

As a pilot, 5 study team members (herein investigators) scored a sample of 10 narrative in-training evaluation reports. Investigators subsequently met to resolve discrepancies in using the tool and clarified the NEQI instructions. Following the pilot, the investigators were provided with narrative in-training evaluation reports from 20 neurology clerkship students chosen by a random selection process to ensure equivalent representation across clerkship blocks. The investigators did not write any narratives used in the study. In total, 123 de-identified retrospective evaluations from 53 unique assessors were independently scored by each investigator.

Interrater reliability was estimated for each NEQI component arm (performance domains, specificity, usefulness) and total NEQI score with intraclass correlation coefficients (ICCs) using a 2-way random effects model (ICCs 2, k) to measure homogeneity of measurements. Bootstrapped 95% confidence intervals (CIs) with 100 replications were calculated for each ICC estimate.

Results

A total of 615 NEQI scores were obtained from the 123 narrative evaluations scored by investigators. Analysis revealed that NEQI scores were normally distributed, with an average score

Figure 1 Narrative Evaluation Quality Instrument

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 =

The Narrative Evaluation Quality Instrument is designed to quantitatively assess the quality of a medical trainee's narrative evaluation. The instrument includes 3 component arms: (1) performance domains commented on, (2) specificity of comments, and (3) usefulness to trainee. Each component arm has a possible score range from 0 to 4, resulting in an overall possible total score ranging from 0 to 12.

of 6.4 (SD 2.9). Regarding the NEQI component arms, the performance domain arm had the largest score (mean 2.6; SD 0.9) and was less variable than the specificity (mean 1.8; SD 1.1) and usefulness (mean 2.0; SD 1.4) arms. The frequency of NEQI scores across the component arms is displayed in figure 2.

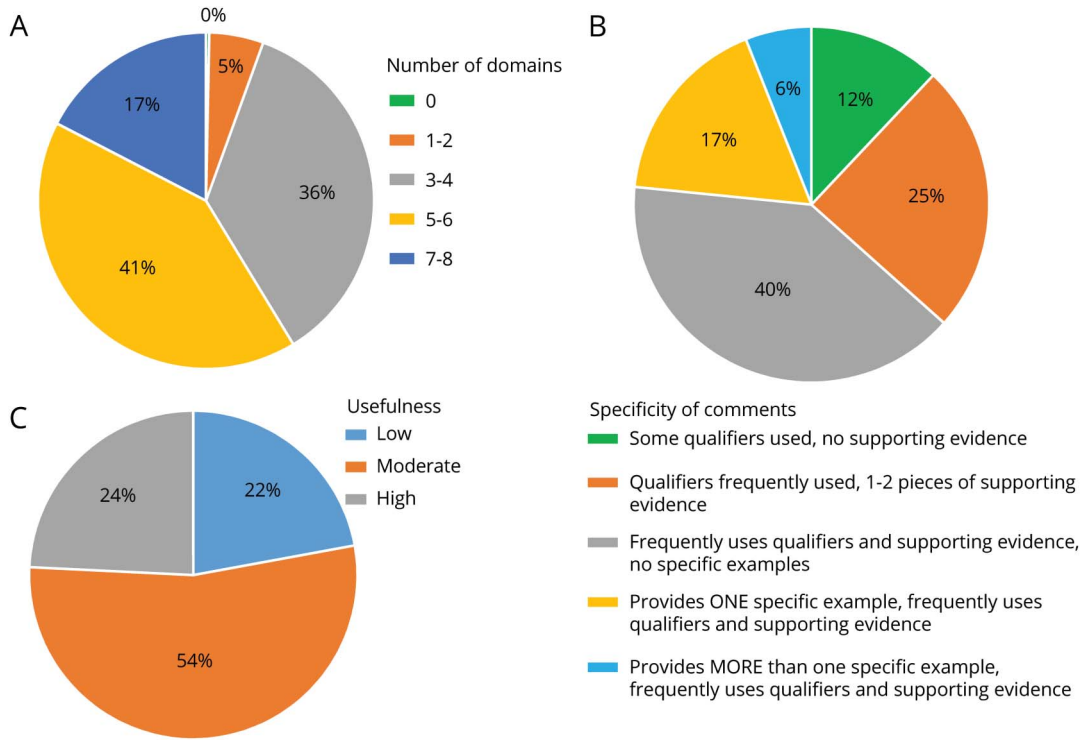
Results further showed that assessors commented on an average of 4.7 (SD 1.7) performance domains per narrative evaluation. The "professionalism" (78%) and "overall performance" (78%) domains were most frequently commented on by assessors. In contrast, assessors commented least frequently on the "prepares and participates in patient care" (33%) and "clinical reasoning" domains (31%).

Overall, the NEQI showed good reliability, with an ICC of 0.81 (95% CI 0.77–0.86). Individual NEQI component scores were found to have moderate reliability: performance domain score = 0.65 (95% CI 0.58–0.72); specificity score = 0.69 (95% CI 0.61–0.77); usefulness score = 0.73 (95% CI 0.66–0.80).

Discussion

Our results indicate that the NEQI can reliably differentiate neurology clerkship narrative evaluation quality along several dimensions including performance domains, specificity of comments, and usefulness to trainee. Reliability was improved when overall NEQI score was used, which suggests that using a more holistic approach to defining narrative quality provides better reliability than single measures. These results build significantly upon prior work, and culminate in a useful tool to systematically and comprehensively assess the quality of neurology clerks' narrative evaluations.[3–13] While the tool was tested on neurology clerkship evaluations, it was derived from the broader medical education literature with the goal of being generalizable to a broad range of student in-training evaluation reports.

Figure 2 Frequencies of Narrative Evaluation Quality Instrument component arm scores



(A) Percentage of narratives with different number of student performance domains commented on. (B) Percentage of narratives meeting different criteria for specificity of comments. (C) Percentage of narratives scored in different categories of usefulness to student.

With respect to student performance domains, those that were included in the prompts of the in-training evaluation reports were most commonly referenced in evaluator narratives. Conversely, the 2 performance domains commented on the least ("prepares and participates in patient care" and "clinical reasoning") were not mentioned explicitly in the prompt. Given the importance of clinical reasoning in teaching and evaluating neurology clerks, it is concerning that few evaluators commented on this domain. Our results suggest that providing assessors with examples of performance domains in the evaluation prompt may increase the number of comments on specific domains. Such prompts could also reasonably extrapolate to other component arms (e.g., using a specific example), which would improve the overall quality of a narrative evaluation.

Given that our study demonstrates reliability evidence for assessing narrative evaluation quality, the NEQI has several applications. First, the tool could help examine effectiveness of interventions designed to enhance narrative quality. Second, the NEQI could be leveraged to study factors associated with narrative quality including relevant demographic factors of evaluation authors and students, time to in-training evaluation report completion, or time assessor spent with a student on the rotation. The NEQI has implications for faculty development. For instance, NEQI feedback could be provided to evaluators as an intervention to improve narrative quality, and through teaching the salient aspects of evaluation, perhaps improve evaluation efficiency as well. Indeed, we are currently in the process of creating a narrative translation of NEQI quantitative data (the narrative evaluation quality report) to use in faculty development efforts.

Further consensus is needed to define a narrative of sufficient quality. Although we did not study what score represented the minimum standard of quality, a minimum NEQI score of 7 is suggestive of a level of quality evaluation to be of moderate usefulness to the student and grading committee. Of the 615 in-training evaluation reports analyzed, 313 (51%) of our narratives reached that threshold of ≥7, indicating at least minimum quality. Defining narrative quality and minimum standards for useable narrative evaluations warrants further research as narrative-only evaluations grow in their use and emphasis in medical education.

There are several limitations to this study. First, the NEQI's development was accomplished by retrospectively assessing evaluations from a single clerkship at a single institution. Thus, additional study employing narrative evaluations utilizing different learner groups across different educational settings and institutions is necessary to ensure the NEQI is generalizable and demonstrates validity evidence. As noted previously, not all performance domains scored in the NEQI were noted in the narrative evaluation prompts. The omission

of these prompts may have influenced the domains that assessors commented on and ultimately, overall NEQI scores. Finally, the study examined only the narrative, and not the numeric, portion of the evaluation. It is possible that assessors intended their narrative evaluations to be viewed in conjunction with their numeric data.

The NEQI provides reliability evidence to comprehensively assess the quality of narrative evaluations, which can gauge effectiveness of interventions targeting narrative quality and aid faculty development. Enhancing the quality of narrative performance evaluations will improve summative and formative methods of assessing neurology clerkship performance.

Study funding

No targeted funding reported.

Disclosure

The authors report no disclosures relevant to the manuscript. Go to Neurology.org/N for full disclosures.

Appendix Authors

Name	Location	Role	Contribution
Michael S. Kelly, MD	University of Rochester	Author	Design and conceptualization of the study, analysis and interpretation of the data, drafting and revising the manuscript
Christopher J. Mooney, PhD	University of Rochester	Author	Design and conceptualization of the study, analysis and interpretation of the data, drafting and revising the manuscript
Justin F. Rosati, MD	University of Rochester	Author	Analysis and interpretation of the data, aided in drafting the manuscript
Melanie K. Braun, MD	University of Rochester	Author	Analysis and interpretation of the data, aided in drafting the manuscript
Robert Thompson Stone, MD	University of Rochester	Author	Design and conceptualization of the study, analysis and interpretation of the data, drafting and revising the manuscript

References

1. Schuwirth LWT, Vleuten CPM. A plea for new psychometric models in educational assessment. Med Educ 2006;40:296–300.
2. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. Med Teach 2013;35:564–568.
3. Hatala R, Sawatsky AP, Dudek N, Ginsburg S, Cook DA. Using in-training evaluation report (ITER) qualitative comments to assess medical students and residents. Acad Med 2017;92:868–879.
4. Bartels J, Mooney CJ, Stone RT. Numerical versus narrative: a comparison between methods to measure medical student performance during clinical clerkships. Med Teach 2017;39:1154–1158.
5. Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: residents' interpretations of written assessment comments. Med Educ 2017;51:401–410.
6. Littlefield JH, Darosa DA, Paukert J, Williams RG, Klamen DL, Schoolfield JD. Improving resident performance assessment data: numeric precision and narrative specificity. Acad Med 2005;80:489–495.
7. Pelgrim EA, Kramer AW, Mokkink HG, Van der Vleuten CP. Quality of written narrative feedback and reflection in a modified mini-clinical evaluation exercise: an observational study. BMC Med Educ 2012;12:97.
8. Pelgrim EAM, Kramer AWM, Mokkink HGA, van der Vleuten CPM. Written narrative feedback, reflections and action plans in single-encounter observations: an observational study. Perspect Med Educ 2013;2:106–108.
9. Plymale MA, Donnelly MB, Lawton J, Pulito AR, Mentzer RM. Faculty evaluation of surgery clerkship students: important components of written comments. Acad Med 2002;77:S45–S47.
10. 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.
11. Ginsburg S, Eva K, Regehr G. Do in-training evaluation reports deserve their bad reputations? A study of the reliability and predictive ability of ITER scores and narrative comments. Acad Med 2013;88:1539–1544.
12. Ginsburg S, Regehr G, Lingard L, Eva K. Reading between the lines: faculty interpretations of narrative evaluation comments. Med Educ 2015;49:296–306.
13. Schwind C, Williams R, Boehler M, Dunnington G. Do individual attendings' post-rotation performance ratings detect residents' clinical performance deficiencies? Acad Med 2004;79:453–457.