Effects of a Faculty Educational Intervention to Improve Quality of Written Evaluations of Pediatrics Clerkship Medical Students

Michelle Prong, MD, MPH, Kunali Gurditta, MD, Grace Ng, MD, and Margarita Corredor, MD
Departments of Internal Medicine and Pediatrics, University of Rochester Medical Center, Rochester NY

INTRODUCTION

- Written evaluations are a cornerstone of clinical assessment in undergraduate medical education; they account for the majority of a student's clerkship grade and, more importantly, students depend on written evaluations to improve their clinical performance.
- Improving the quality of faculty-written narrative evaluations remains a challenge.
- The Narrative Evaluation Quality Instrument (NEQI) is a validated tool to assess quality of medical student narrative evaluations[1].
- A randomized, controlled faculty educational intervention using NEQI has demonstrated marked improvement in Internal Medicine clerkship faculty written evaluation scores post-intervention[2].

OBJECTIVE

To determine if the quality of faculty-written narrative evaluations of medical students can be improved through trainee-delivered feedback based on NEQI scoring principles.

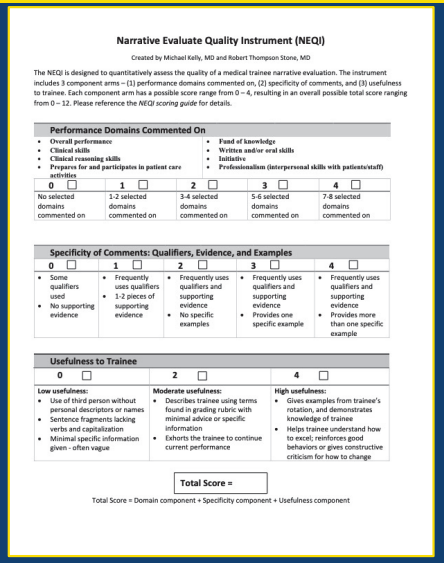
METHODS

SETTING, PARTICIPANTS, & INTERVENTION

- All Pediatric Hospital Medicine Faculty at the University of Rochester Medical Center were invited to participate.
- Participation was solicited via email and in-person at faculty meetings.
- Three Internal Medicine-Pediatrics senior residents trained in NEQI scoring reviewed each participating faculty's three most recent narrative evaluations of pediatrics clerkship medical students.
- Resident reviewers completed one 30-minute, in-person feedback session with each faculty participant to review the participant's NEQI scores, current evaluation strengths, and areas for growth.
- Following feedback sessions, resident reviewers scored at least two subsequent, de-identified narrative evaluations for each faculty participant.

NEQI TOOL, SCORING, & ANALYSIS

- The NEQI tool includes three component arms: breadth of performance domains evaluated, specificity of comments, and usefulness to the trainee. Each component arm has a score range from 0-4; the maximum overall score is 12 for a particular written evaluation (Image 1).
- Narrative evaluations were de-identified prior to analysis.
- Each evaluation was independently scored by two separate resident reviewers.
- Descriptive statistics for pre- and post-intervention scores including mean, standard deviation (SD), and 95% confidence intervals (CI) were calculated using Microsoft Excel.
- A p-value for pre- and post-intervention scores was calculated using a one tailed, paired T-test.



Narrative Evaluate Quality Instrument (NEQI)

Created by Michael Kelly, MD and Robert Thompson Stone, MD

The NEQI is designed to quantitatively assess the quality of a medical trainee 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 – 4, resulting in an overall possible total score ranging from 0 – 12. Please reference the *NEQI scoring guide* for details.

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 =

Total Score = Domain component + Specificity component + Usefulness component

Image 1: NEQI Scoring Rubric

	Pre-Intervention (SD, 95% CI)	Post-Intervention (SD, 95% CI)	P-value
NEQI Total Average Score	8.79 (0.96, 8.45-9.13)	9.56 (1.89, 9.25-9.87)	0.15
Performance Domains Commented On	3.40 (0.55, 2.85-3.95)	3.42 (0.45, 2.97-3.86)	0.44
Specificity of Comments	2.80 (0.44, 2.35-3.25)	3.15 (0.96, 2.19-4.11)	0.18
Usefulness to Trainee	2.60 (0.60, 2.00-3.20)	3.00 (1.00, 2.00-4.00)	0.14

Table 1: Pre- and Post-Intervention NEQI Scores

RESULTS

- Five out of 17 (29.4%) eligible Pediatric Hospital Medicine faculty participated in the study.
- Pre- and Post-Intervention NQEI total average scores and subcategory scores are reported in Table 1.
- The results demonstrated a p-value of 0.15.

DISCUSSION & CONCLUSIONS

- There was most improvement in the "specificity of comments" and "usefulness to the trainee" component arms pre- and post-intervention. As medical education tends to encourage trainees to practice individual reflection and continual improvement through formative assessment, the gains these two component arms of evaluations are encouraging and empowering to the learner.
- There are a myriad of factors influencing the quality of written evaluations beyond faculty knowledge of what constitutes a quality evaluation. Some factors are likely to include numerous other administrative tasks of educators, lag time between working with a learner and writing their evaluation, and the time required to complete detailed narrative evaluations (compared with simpler Likert scale evaluations).
- The results do not suggest a strong statistical difference in average total NEQI scores or subcategory scores pre- and post-intervention, which may represent a type II error owing to the small sample size.
- Despite our results lacking statistical significance, many faculty commented anecdotally that the current intervention was helpful for them to reflect on their own practices for writing narrative evaluations and learn new ways to make their evaluations more helpful for learners. This suggests that the intervention was well-received by faculty and may represent a future direction for faculty development.

LIMITATIONS

- Small sample size; only five eligible faculty members elected to participate.
- Opt-in recruitment strategy may have selected for faculty members who intrinsically value or who were already motivated to write high-quality evaluations.
- Restricting participants to Pediatric Hospital Medicine faculty may limit generalizability to faculty in other departments or divisions within Pediatrics.
- Discrepancies in the number of completed evaluations in the post-intervention period was associated with larger standard deviations and therefore less precise estimates of the intervention's effects.

REFERENCES

1. Kelly, M. S., Mooney, C. J., Rosati, J. F., Braun, M. K., & Thompson Stone, R. (2020). Education research: the narrative evaluation quality instrument: development of a tool to assess the assessor. Neurology, 94(2), 91-95.
2. Bergin, C, Blatt, A, Pascoe J, & Skehan, N. Evaluating the Evaluators: Strategies to Improve the Quality of Summative Evaluations. Workshop given at Academic Internal Medicine Week Conference, April 2022.