Canadian Medical Education Journal

Brief Reports

The Quality of Assessment for Learning score for evaluating written feedback in anesthesiology postgraduate medical education: a generalizability and decision study

Intérêt d'un score de la qualité de l'évaluation pour l'apprentissage pour évaluer la rétroaction écrite dans la formation postdoctorale en anesthésiologie : étude de généralisabilité et de décision

Eugene K Choo,[1] Rob Woods,[2] Mary Ellen Walker,[1] Jennifer M O'Brien,[1] Teresa M Chan[3]

[1]Department of Anesthesiology, College of Medicine, University of Saskatchewan, Saskatchewan, Canada; [2]Department of Emergency Medicine, College of Medicine, University of Saskatchewan, Saskatchewan, Canada; [3]Department of Medicine (Division of Emergency Medicine; Division of Education & Innovation), Michael G. DeGroote School of Medicine, Faculty of Health Sciences, McMaster University and Office of Continuing Professional Development & McMaster Education Research, Innovation, and Theory (MERIT) Program, Faculty of Health Sciences, McMaster University, Ontario, Canada.

Correspondence to: Dr. Eugene Choo, Department of Anesthesiology, College of Medicine, University of Saskatchewan, Saskatoon, SK, Canada; email: eugene.choo@usask.ca

Edited by: Cindy Schmidt (senior section editor); Marcel D'Eon (editor-in-chief)

Published ahead of issue: Oct 10, 2023; CMEJ 2023 Available at https://doi.org/10.36834/cmej.75876

Abstract

Background: Competency based residency programs depend on high quality feedback from the assessment of entrustable professional activities (EPA). The Quality of Assessment for Learning (QuAL) score is a tool developed to rate the quality of narrative comments in workplace-based assessments; it has validity evidence for scoring the quality of narrative feedback provided to emergency medicine residents, but it is unknown whether the QuAL score is reliable in the assessment of narrative feedback in other postgraduate programs.

Methods: Fifty sets of EPA narratives from a single academic year at our competency based medical education post-graduate anesthesia program were selected by stratified sampling within defined parameters [e.g. resident gender and stage of training, assessor gender, Competency By Design training level, and word count (≥17 or <17 words)]. Two competency committee members and two medical students rated the quality of narrative feedback using a utility score and QuAL score. We used Kendall's tau-b co-efficient to compare the perceived utility of the written feedback to the quality assessed with the QuAL score. The authors used generalizability and decision studies to estimate the reliability and generalizability coefficients.

Results: Both the faculty's utility scores and QuAL scores ($r = 0.646$, $p < 0.001$) and the trainees' utility scores and QuAL scores ($r = 0.667$, $p < 0.001$) were moderately correlated. Results from the generalizability studies showed that utility scores were reliable with two raters for both faculty (Epsilon=0.87, Phi=0.86) and trainees (Epsilon=0.88, Phi=0.88).

Conclusions: The QuAL score is correlated with faculty- and trainee-rated utility of anesthesia EPA feedback. Both faculty and trainees can reliability apply the QuAL score to anesthesia EPA narrative feedback. This tool has the potential to be used for faculty development and program evaluation in Competency Based Medical Education. Other programs could consider replicating our study in their specialty.

Résumé

Contexte : La qualité de la rétroaction à la suite de l'évaluation d'activités professionnelles confiables (APC) est d'une importance capitale dans les programmes de résidence fondés sur les compétences. Le score QuAL (Quality of Assessment for Learning) est un outil développé pour évaluer la qualité de la rétroaction narrative dans les évaluations en milieu de travail. Sa validité a été démontrée dans le cas des commentaires narratifs fournis aux résidents en médecine d'urgence, mais sa fiabilité n'a pas été évaluée dans d'autres programmes de formation postdoctorale.

Méthodes : Cinquante ensembles de commentaires portant sur des APC d'une seule année universitaire dans notre programme postdoctoral en anesthésiologie – un programme fondé sur les compétences – ont été sélectionnés par échantillonnage stratifié selon des paramètres préétablis [par exemple, le sexe du résident et son niveau de formation, le sexe de l'évaluateur, le niveau de formation en Compétence par conception, et le nombre de mots (≥17 ou <17 mots)]. Deux membres du comité de compétence et deux étudiants en médecine ont évalué la qualité de la rétroaction narrative à l'aide d'un score d'utilité et d'un score QuAL. Nous avons utilisé le coefficient tau-b de Kendall pour comparer l'utilité perçue de la rétroaction écrite et sa qualité évaluée à l'aide du score QuAL. Les auteurs ont utilisé des études de généralisabilité et de décision pour estimer les coefficients de fiabilité et de généralisabilité.

Résultats : Les scores d'utilité et les scores QuAL des enseignants ($r = 0,646$, $p < 0,001$) et ceux des étudiants ($r = 0,667$, $p < 0,001$) étaient modérément corrélés. Les résultats des études de généralisabilité ont montré qu'avec deux évaluateurs les scores d'utilité étaient fiables tant pour les enseignants (Epsilon=0,87, Phi=0,86) que pour les étudiants (Epsilon=0,88, Phi=0,88).

Conclusions : Le score QuAL est en corrélation avec l'utilité de la rétroaction sur les APC en anesthésiologie évaluée par les enseignants et les étudiants. Les uns et les autres peuvent appliquer de manière fiable le score QuAL aux commentaires narratifs sur les APC en anesthésiologie. Cet outil pourrait être utilisé pour le perfectionnement professoral et l'évaluation des programmes dans le cadre d'une formation médicale fondée sur les compétences. D'autres programmes pourraient envisager de reproduire notre étude dans leur spécialité.

Introduction

Competency-based medical education (CBME) in Canada is based on a robust program of assessment, relying on direct observation assessments that combine a numeric rating of entrustment with written narrative comments.[1,2] Canadian residency programs have adopted Entrustable Professional Activities (EPAs) for work-based assessment to facilitate frequent, formative, and low-stakes assessment for learning. Effective feedback for learning provides the learner with guidance and direction through targeted, specific, and actionable narratives.[3] Despite their importance, written narrative comments provided to trainees are not currently measured for quality.[4,5] Canadian anesthesiology residency programs have fully transitioned to CBME, pushing the research focus away from CBME implementation, and toward CBME outcomes.[6] A tool to assess the quality of written narrative feedback in anesthesia residency education would be useful for program evaluation and faculty development initiatives.[7] and improved training experiences and learning for residents.

The Quality Assessment for Learning (QuAL) score was developed to measure the perceived utility of narrative comments. It consists of three domains with a total score of 5 (Table 1). The QuAL score has been assessed for use by emergency medicine faculty involved in resident assessments (program directors, competency chair members) on emergency medicine written narrative feedback and EPA work-based assessment tools.[8,9] In this setting, the QuAL score has demonstrated acceptable reliability with only two raters and high correlation with the perceived utility of the narrative feedback for both trainees and meta-raters (faculty members interpreting others' comments, such as competency committee members).[8,9] We approached the idea of validity as an evidentiary chain, requiring ongoing analysis and interpretation of assessment results. Validity is not an inherent property of the QuAL score tool itself.[10] As such, with the changing of context (such as applying a tool to a new population), new validity evidence for a tool's usage in this new context is imperative. Although there is emerging validity evidence that the QuAL score is useful in emergency medicine (both pre- and post-EPA implementation),[8,9] we do not know whether the QuAL score can be applied to: (1) written narrative feedback in the context of anesthesia EPA work-based assessments, (2) to those not responsible for resident assessment such as learners, and (3) whether learners' perceptions of feedback utility are similar to faculty perceptions of feedback utility.[11] We had two aims within this study: 1) to assess the inter-rater reliability of QuAL scores within the anesthesia context (as measured by a decision study); and 2) to gather validity evidence for the QuAL score within a dataset of anesthesiology EPA feedback by mapping the score to perceptions of utility from trainees (n = 2) and anesthesia program competency committee members (n = 2) (as measured by both the correlation, generalizability, and decision studies).

Methods

In 2021, we conducted a single-centre rating study of EPAs from a Canadian Anesthesia Training program aimed at examining the reliability of the QuAL score and to further establish validity evidence for the QuAL score's ability to discern usefulness of the narratives within EPAs. See Figure 1 for a graphical representation of the study design.

Setting

Our study was conducted via online survey with participant-rater contributors from a single university (University of Saskatchewan).

Ethical Considerations

This study was deemed exempt from ethical review by the University of Saskatchewan Research Ethics Board under Article 2.5 of the Tri-Council Policy Statement. Trainees were informed of the study and given the opportunity to dissent; one resident dissented, and their assessment data were removed from the dataset. Names and gender of trainees and faculty embedded in the comment were removed. Raters signed a confidentiality agreement to ensure the data (although anonymized) was not used for any other purposes.

Data selection

De-identified program-level data were obtained for anesthesia residents (n = 29) within a single Royal College accredited five-year post-graduate residency training program in anesthesiology. One researcher (EC) pulled the narrative feedback from EPAs from the 2020-2021 academic year (n = 1591). After classifying these EPA data by a number of facets [resident gender and stage of training (man/woman; junior/senior), assessor gender (man/woman), Competency By Design training level (transition to discipline/foundations/core/transition to practice), and word count (≥17 or <17 words)], we selected EPA assessments using stratified sampling to create our rating dataset (n = 50).

Sample size estimates

Based on multiple previous studies, at least two raters are required to establish reliability with the QuAL score ratings.[8,9] Since we were testing the score in a new context, we doubled the number of raters (n = 4 in total) to ensure that we were more conservative in our rating exercise. Furthermore, to obtain unbiased Phi and G coefficients, a minimum sample size of at least 50 has been recommended by methodological experts and researchers performing similar studies.[8,9]

Survey design and rating activity

The data were compiled into two surveys that asked raters to score 50 written comments for utility ("Do you think the resident who received this feedback found it useful?") using a 3-point ordinal scale (2=Yes; 1=Maybe; 0=No) and the QuAL score (Table 1). The QuAL score is calculated in three domains with a total score of 5.[8] One team member (EC) created the survey tool in Survey Monkey (Momentive Inc, San Mateo, California, USA). Another member (JO) provided feedback on the survey tool.

We enlisted two competency committee (CC) members and two learners to rate the utility and quality of narrative EPA comments. Each rater completed utility scoring in a single sitting followed by the QuAL scoring a week later to minimize recall bias.

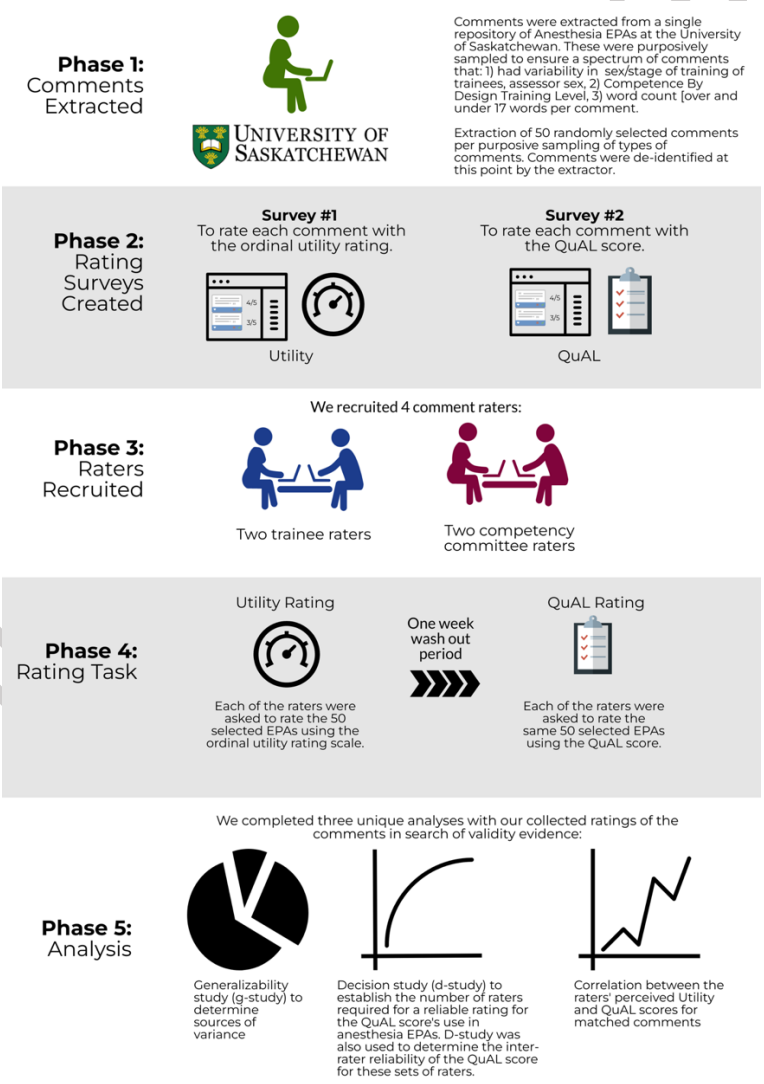

Figure 1. A graphical depiction of our study protocol.

Table 1. Quality of Assessment of Learning (QuAL) score components

Evidence	Does the assessor provide sufficient evidence about resident performance? (0= no comment at all; 1= no, but comment present; 2= somewhat; 3 = yes, full description)
Suggestion	Does the rater provide a suggestion for improvement? (0=no; 1=yes)
Connection	Is the rater's suggestion linked to the behavior described? (0=no; 1=yes).

Analysis

Generalizability & decision studies. We conducted generalizability studies (G-Study) and decision studies (D-Study) to determine sources of variance and reliability in ratings of QuAL scores among learners and faculty. A generalizability study (G-study) allows for assessment of multiple sources of error (called facets) that affect reliability.[12] We used a G-study to explore how facets contribute to the proportion of variance attributed to the object of measurement and the suspected sources of error. G-studies allowed us to assess the validity of the QuAL score and the utility score in rating feedback on anesthesiology resident EPAs. We assessed the reliability of trainees (medical students, *n* = 2) vs. CC members (*n* = 2) in their ratings. Further, we explored whether QuAL scores were correlated with the perceived utility of the narrative comments on EPA assessments for anesthesia trainees. We used G string VI software (Hamilton, ON, Canada) to perform the G-studies and D-studies.

Correlation studies. To compare the QuAL scores and Utility ratings, we use Friedman's ANOVA and Wilcoxon matched-pairs sign rank tests to test for differences in scores between raters. We calculated Kendall's tau-b correlation coefficient to assess relationships between two utility scores or between two QuAL scores, and we used Kendall's tau-c correlation coefficient to assess relationships between utility scores and QuAL scores. Kendalls' tau-b and Kendall's tau'c were used because data were not normally distributed and had many tied ranks.[13]

We used IBM SPSS Statistics (version 28) for the above analyses.

Results

Four raters completed 100% of rating activities (both rounds).

Differences between raters scores

When exploring differences between raters, we found no significant difference between raters in QuAL score ratings (chi-squared = 6.7(3), p = 0.082). There were significant differences between raters in utility scores (chi-squared = 63.6(3), p<0.001). Pairwise testing showed no significant differences between utility scores within the two rater groups, but the CC members' utility ratings (median=1.0, interquartile range [IQR]=1.0 to 2.0) were lower than the trainee raters (median=2.5, IQR=1.0 to 3.0).

Generalizability theory analysis. Utility scores were reliable with two raters for both CC members (phi=0.86) and trainees (phi=0.88). A phi value (absolute g-coefficient) greater than 0.80 is usually considered a minimum standard for high stakes assessments.[14] QuAL scores were also reliable with all four raters (phi = 0.90) with two raters for both CC members (phi=0.90) and trainees (phi=0.90). The generalizability results are summarized in Table 2.

Decision Study. The D-study shows that both CC members and trainees require two raters for both utility scores and QuAL scores to get a g-coefficient >0.80 (Figure 2). Overall, both QuAL scores and utility scores have similar levels of reliability in both CC members and trainees.

Correlation Studies. After calculating the various pairwise results for the QuAL vs. Utility scores of each rater, we found that these values paired well across all four raters. Overall, the utility score and the QuAL score had a moderate-to-strong correlation with each rater (Table 3).

Table 2. Variance components for expert and learner QuAL scores.

Absolute g-coefficient (Phi)	Interrater reliability (Epsilon)	Generalizability Study Results (Variance Components)				
		Comment	Rater	Rater type (nested in rater)	Comment x Rater	Error
0.90	0.90	92%	0%	1%	6%	6%

Absolute g-coefficient (phi) – indicates the generalizability of scores of these findings to another study.
Interrater reliability – d-study estimate of reliability of scores from one to another potential rater
Rater – the four raters of our study
Rater type – Competency Committee Faculty Member vs. Medical Student Rater
Comment x Rater – the % variance contribution by interaction between the comment and the raters
Error – remaining variability that is not explained; represents interaction between all the facets (comment x rater:type)

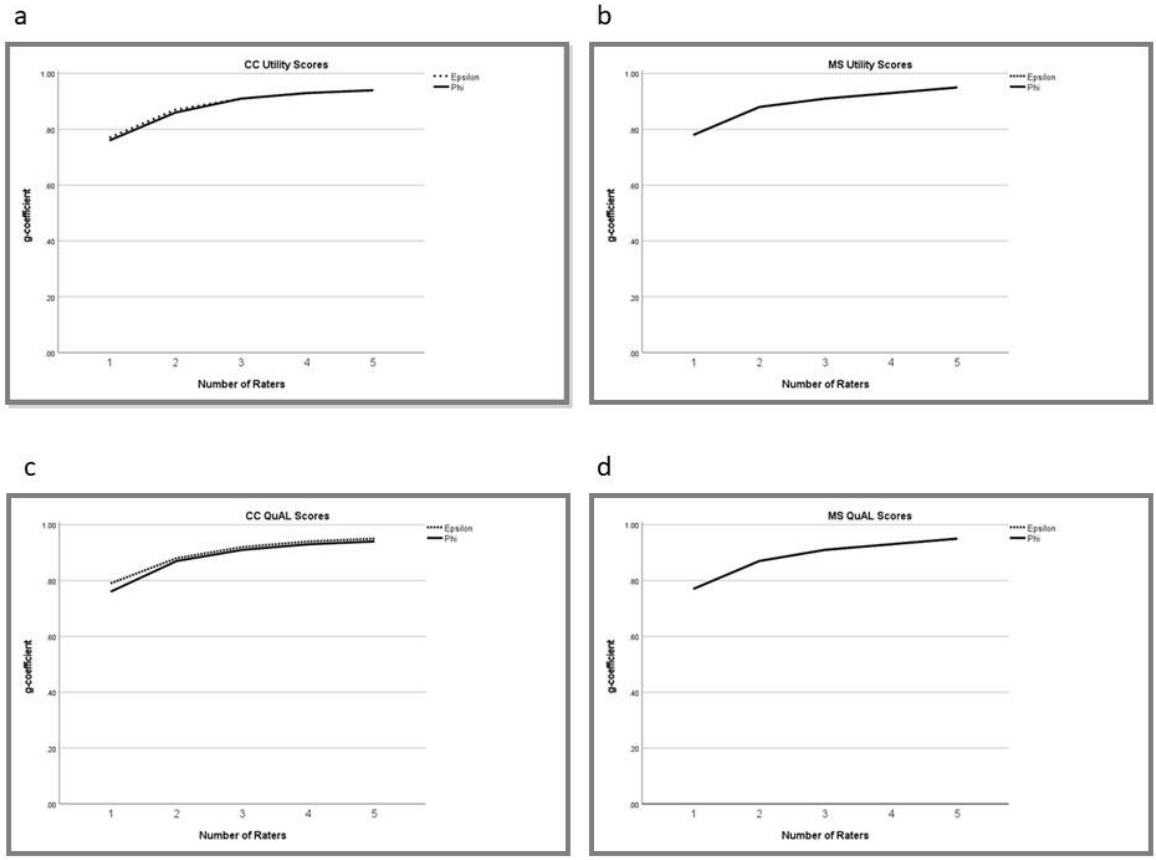

Figure 2. G-coefficients for: a) Competency Committee (CC) Utility scores; b) Trainee (MS) utility scores; c) CC QuAL scores; and d) MS QuAL scores. Note: the phi lines are covering the epsilon lines in panels b) and d) because their coefficients are the same

Table 3. Pairwise correlation results as calculated using Kendall's Tau

		Utility Rating			QuAL Score			
		CC2	MS1	MS2	CC1	CC2	MS1	MS2
DoneUtility Rating (95% CI)	CC1	0.78† (0.69- 0.84)	0.64†(0.52-0.74)	0.69† (0.57- 0.78)	0.68* (0.51-0.84)	0.46* (0.30- 0.62)	0.64* (0.51 to 0.78)	0.60* (0.46-0.74)
	CC2		0.64† (0.52-0.74)	0.66† (0.54- 0.76)	0.61* (0.42-0.80)	0.56* (0.40- 0.73)	0.53* (0.36 to 0.71)	0.57* (0.39-0.76)
	MS1			0.708† (0.60-0.79)	0.67* (0.55-0.79)	0.56* (0.41- 0.71)	0.70* (0.59- 0.81)	0.74* (0.61-0.88)
	MS2				0.67* (0.54-0.79)	0.50* (0.36-0.63)	0.68* (0.58-0.77)	0.72* (0.53-0.91)
QuAL Score (95% CI)	CC1					0.72† (0.62-0.80)	0.76† (0.67 to 0.83)	0.79† (0.71-0.86)
	CC2						0.70† (0.58 to 0.78)	0.66 (0.54-0.76)
	MS1							0.79† (0.71-0.78)

*Notes: Correlations with 95% confidence intervals. Legend: † denotes calculation using Kendall's Tau-b. * denotes calculation using Kendall's Tau-c. All pairings listed in the above table had a p-value significance of <0.001.*

		Utility Rating			QuAL Score			
		CC2	MS1	MS2	CC1	CC2	MS1	MS2
Utility Rating	CC1	0.778†	0.642†	0.687†	0.677*	0.456*	0.644*	0.600*
	CC2		0.644†	0.660†	0.605*	0.562*	0.533*	0.574*
	MS1			0.708†	0.672*	0.556*	0.698*	0.744*
	MS2				0.665*	0.496*	0.678*	0.721*
QuAL Score	CC1					0.720†	0.762†	0.794†
	CC2						0.696†	0.661†
	MS1							0.791†

*Notes: Legend: † denotes calculation using Kendall's Tau-b. * denotes calculation using Kendall's Tau-c. All pairings listed in the above table had a p-value significance of <0.001.*

Discussion

We found that the QuAL score was reliable and shows validity evidence as compared to a simple ordinal utility rating tool for anesthesia EPA assessments. We found that the QuAL score was usable with a high level of reliability with both inexperienced trainees (e.g. medical students, MS) and faculty members who were part of a CC. Importantly, the QuAL score had a strong association with both the trainees' and faculty members' perceptions of feedback utility. Our results show a robust correlation between QuAL scores and utility ratings, providing validity evidence that the QuAL score is useful for discerning the quality of anesthesia narrative feedback.

Interestingly, as opposed to previous studies on the QuAL score,[8,9] our group sought to also compare the rater groups' perceptions of utility as another check of our process. The difference in utility ratings between CC members and trainees suggests there may be differences in how narrative feedback is perceived, which lends a certain level of validity evidence for the ordinal utility ratings scale that was previously used in both the derivational QuAL study by Chan and colleagues and the subsequent study by Woods et al.[8,9] Previous literature suggests that competency committee members act as "meta-raters" (i.e. faculty members who seek to make judgements based upon the ratings' of others). Whereas most scoring tools simply look at the efficacy of comments for learners, the QuAL score has a dual purpose of being useful for feedback to the trainee *and* useful for meta-raters.[15-16]

Implications for Practice

We believe that using the QuAL score can serve to inform faculty development and program evaluation. There is increasing literature suggesting that written comments are very useful.[17-28] The QuAL score may scaffold faculty members who may not have had substantive training to achieve a higher quality comment. Moreover, when meta-raters are faced with the complexities of making summative determinations within competency committees about larger swaths of narratives, the QuAL score may act as a tool that allows them to compare the quality of narrative feedback.

Limitations

Our study had a few limitations. In our study, raters were not instructed on how to apply the QuAL score; calibrating raters by providing examples of different narrative quality and QuAL scores may further improve rater reliability. While we attempted to represent all EPA facets by using a stratified sample, the overall quality of the narrative feedback in our dataset was low which may have skewed the utility scoring due to comparative bias. If we failed to capture the true variability of the data in our sample, this may have impacted the correlation coefficients or G-coefficients. This was also a single centre program evaluation limits the generalizability of our study, however, as we aimed to build more validity evidence for a known scoring tool, such replication work is often required to build the body of evidence around a particular tool.[10]

Next steps

Since this is the first study to date that seeks to apply the QuAL score to the anesthesia context, further work can be done to improve the discriminatory ability of the QuAL score for narrative EPAs in our specialty. Due to their ability to also map to trainee utility scores, the QuAL score may be useful for programs to engage in quality improvement audit and feedback processes for individual faculty members or all faculty members within a program.[29] While procedural-leaning specialties such as anesthesia may be prone to gender bias,[30-31] other specialties should examine their assessment systems for such gender-related bias.[32-35]

Conclusions

The QuAL score is correlated with both faculty and trainee perceived utility in anesthesia EPA feedback. Both faculty and trainees can reliably apply the QuAL score to anesthesia EPA narrative feedback.

Conflicts of Interest: None

Funding: None.

Acknowledgements: We gratefully acknowledge the contributions of Vivian Murungi and Devin Edwards, medical students, who rated the quality of narrative feedback using a utility score and QuAL score.

References

1. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Med Teach*. 2010;32(8):638–45. https://doi.org/10.3109/0142159x.2010.501190
2. Holmboe ES, Sherbino J, Long DM, Swing SR, Frank JR. The role of assessment in competency-based medical education. *Med Teach*. 2010 Aug 1;32(8):676–82. https://doi.org/10.3109/0142159x.2010.500704
3. Caverzagie KJ, Nousiainen MT, Ferguson PC, et al. Overarching challenges to the implementation of competency-based medical education. *Med Teach*. 2017 Jun 3;39(6):588–93. https://doi.org/10.1080/0142159x.2017.1315075
4. Jensen AR, Wright AS, Kim S, Horvath KD, Calhoun KE. Educational feedback in the operating room: a gap between resident and faculty perceptions. *Amer J Surg.* 2012 Aug

1;204(2):248–55. https://doi.org/10.1016/j.amjsurg.2011.08.019
5. Upadhyaya S, Rashid M, Davila-Cervantes A, Oswald A. Exploring resident perceptions of initial competency based medical education implementation. *Can Med Educ J.* 2021 Apr;12(2):e42–56. https://doi.org/10.36834/cmej.70943
6. Weller JM, Naik VN, Diego RJS. Systematic review and narrative synthesis of competency-based medical education in anaesthesia. *Brit J Anaesthesia.* 2020 Jun 1;124(6):748–60. https://doi.org/10.1016/j.bja.2019.10.025
7. Yilmaz Y, Carey R, Chan T et al. Developing a dashboard for faculty development in competency-based training programs: a design-based research project. *Can Med Ed J*. 2021 Sep 15;12(4):48–64. https://doi.org/10.36834/cmej.72067
8. Chan TM, Sebok-Syer SS, Sampson C, Monteiro S. The Quality of Assessment of Learning (Qual) Score: validity evidence for a scoring system aimed at rating short, workplace-based comments on trainee performance. *Teach Learn Med*. 2020 Jul;32(3):319–29. https://doi.org/10.1080/10401334.2019.1708365
9. Woods RA, Singh S, Thoma B, et al. Validity evidence for the QuAL (Quality of Assessment for Learning) score: a quality metric for supervisor comments in Competency Based Medical Education. *Can Med Ed J*. 2022; 13(6); 19-35. https://doi.org/10.36834/cmej.74860
10. St-Onge C, Young M, Eva KW, Hodges B. Validity: one word with a plurality of meanings. *Adv in Health Sci Educ*. 2017 Oct;22(4):853–67. https://doi.org/10.1007%2Fs40037-018-0433-x
11. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: a practical guide to Kane's framework. *Med Educ.* 2015 Jun;49(6):560–75. https://doi.org/10.1111/medu.12678
12. Streiner DL, Norman GR. *Health measurement scales: a practical guide to their development and use.* 4th ed. Oxford: Oxford University Press; 2008. 452 p. Available from: https://oxford.universitypressscholarship.com/10.1093/acprof:oso/9780199231881.001.0001/acprof-9780199231881 [Accessed on Oct 3, 2021].
13. Field A. *Discovering statistics using SPSS*. 3rd ed., SAGE Publications.2009
14. Streiner DL. Starting at the beginning: an introduction to coefficient alpha and internal consistency. *J Pers Assess*. 2003 Feb;80(1):99-103. https://doi.org/10.1207/S15327752JPA8001_18
15. Telio S, Regehr G, Ajjawi R. Feedback and the educational alliance: examining credibility judgements and their consequences. *Med Educ*. 2016 Sep;50(9):933–42. https://doi.org/10.1111/medu.13063
16. Acai A, Li SA, Sherbino J, Chan TM. Attending emergency physicians' perceptions of a programmatic workplace-based assessment system: the McMaster Modular Assessment Program (McMAP). *Teach Learn Med*. 2019 Aug 8;31(4):434-44. https://doi.org/10.1080/10401334.2019.1574581
17. Chan T, Oswald A, Hauer KE, et al. Diagnosing conflict: conflicting data, interpersonal conflict, and conflicts of interest in clinical competency committees. *Med Teach*. 2021 Jul 3;43(7):765-73. https://doi.org/10.1080/0142159x.2021.1925101
18. Dudek N, Dojeiji S. Twelve tips for completing quality in-training evaluation reports. *Med Teach*. 2014 Dec;36(12):1038–42. https://doi.org/10.3109/0142159x.2014.932897
19. Gray JD. Global rating scales in residency education. *Acad Med.* 1996 Jan;71(1 Suppl):S55-63. https://doi.org/10.1097/00001888-199601000-00043
20. 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 Jun 1;92(6):868–79. https://doi.org/10.1097/acm.0000000000001506
21. Ginsburg S, van der Vleuten C, Eva KW, Lingard L. Hedging to save face: a linguistic analysis of written comments on in-training evaluation reports. *Adv Health Sci Educ*. 2016 Mar;21:175-88. https://doi.org/10.1007/s40037-021-00681-w
22. Ginsburg S, Regehr G, Lingard L, Eva KW. Reading between the lines: faculty interpretations of narrative evaluation comments. *Med Ed*. 2015 Mar;49(3):296-306. https://doi.org/10.1111/medu.12637
23. 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 Oct 1;88(10):1539-44. https://doi.org/10.1097/acm.0b013e3182a36c3d
24. Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: residents' interpretations of written assessment comments. *Med Ed*. 2017 Apr;51(4):401-10. https://doi.org/10.1111/medu.13158
25. Ginsburg S, Watling CJ, Schumacher DJ, Gingerich A, Hatala R. Numbers encapsulate, words elaborate: toward the best use of comments for assessment and feedback on entrustment ratings. *Acad Med*. 2021 Jul 1;96(7S):S81-6. https://doi.org/10.1097/acm.0000000000004089
26. Li SA, Sherbino J, Chan TM. McMaster Modular Assessment Program (McMAP) through the years: residents' experience with an evolving feedback culture over a 3-year period. *AEM Educ Training.* 2017 Jan;1(1):5-14. http://dx.doi.org/10.1002/aet2.10009
27. Chan TM, Sherbino J, Mercuri M. Nuance and noise: lessons learned from longitudinal aggregated assessment data. *J Grad Med Ed*. 2017 Dec;9(6):724-9. https://doi.org/10.4300/jgme-d-17-00086.1
28. Sebok-Syer SS, Klinger DA, Sherbino J, Chan TM. "It's complicated": understanding the relationships between checklists, rating scales, and written comments in workplace-based assessments. *Acad Med*. 2016 Nov 1;91(11):S10. https://doi.org/10.1097/ACM.0000000000001373
29. Yilmaz Y, Carey R, Chan TM et al. Developing a dashboard for faculty development in competency-based training programs: a design-based research project. *Can Med Ed J*. 2021 Oct 20;12(4):48-64. https://doi.org/10.36834/cmej.72067
30. Miller J, Katz D. Gender differences in perception of workplace experience among anesthesiology residents. *JEPM*. 2018 Jan;20(1).
31. Pearce G, Sidhu N, Cavadino A, Shrivathsa A, Seglenieks R. Gender effects in anaesthesia training in Australia and New

Zealand. *Brit j anaesthesia.* 2020 Mar 1;124(3):e70-6. https://doi.org/10.1016/j.bja.2019.12.020

32. Dayal A, O'Connor DM, Qadri U, Arora VM. Comparison of male vs female resident milestone evaluations by faculty during emergency medicine residency training. *JAMA intern med.* 2017 May 1;177(5):651-7. https://doi.org/10.1001/jamainternmed.2016.9616
33. Mamtani M, Shofer F, Scott K, et al. Gender differences in emergency medicine attending physician comments to residents: a qualitative analysis*. JAMA Network Open.* 2022 Nov 1;5(11):e2243134-. https://doi.org/10.1001/jamanetworkopen.2022.43134
34. Menchetti I, Eagles D, Ghanem D, Leppard J, Fournier K, Cheung WJ. Gender differences in emergency medicine resident assessment: a scoping review. *AEM Educ Training.* 2022 Oct;6(5):e10808. https://doi.org/10.1007/s11606-019-04884-0
35. Santen SA, Yamazaki K, Holmboe ES, Yarris LM, Hamstra SJ. Comparison of male and female resident milestone assessments during emergency medicine residency training: a national study. *Acad Med*. 2020 Feb;95(2):263. https://doi.org/10.1097/acm.0000000000002988