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 Chan3

1Department of Anesthesiology, College of Medicine, University of Saskatchewan, Saskatchewan, Canada; 2Department of Emergency Medicine, College of Medicine, University of Saskatchewan, Saskatchewan, Canada; 3Department 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.

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 medical education 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 anesthesiology 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 reliably 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 prédéfinis (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 professionnel 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. 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. Despite their importance, written narrative comments provided to trainees are not currently measured for quality. Canadian anesthesiology residency programs have fully transitioned to CBME, pushing the research focus away from CBME implementation, and toward CBME outcomes. A tool to assess the quality of written narrative feedback in anesthesia residency education would be useful for program evaluation and faculty development initiatives.

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

<table>
<thead>
<tr>
<th>Evidence</th>
<th>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)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suggestion</td>
<td>Does the rater provide a suggestion for improvement? (0: no; 1: yes)</td>
</tr>
<tr>
<td>Connection</td>
<td>Is the rater’s suggestion linked to the behavior described? (0: no; 1: yes).</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Absolute g-coefficient (Phi)</th>
<th>Interrater reliability (Epsilon)</th>
<th>Generalizability Study Results (Variance Components)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment</td>
<td>Rater</td>
<td>Rater type (nested in rater)</td>
</tr>
<tr>
<td>0.90</td>
<td>0.90</td>
<td>92%</td>
</tr>
</tbody>
</table>

*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.

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

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.
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, 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. 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.

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. 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.

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. While procedural-leaning specialties such as anesthesia may be prone to gender bias, other specialties should examine their assessment systems for such gender-related bias.

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


