Canadian Medical Education Journal

Original Research

Validity evidence for the Quality of Assessment for Learning score: a quality metric for supervisor comments in Competency Based Medical Education

Preuve de la validité du score de la qualité de l'évaluation pour l'apprentissage : une mesure de qualité pour les commentaires des superviseurs dans la formation médicale fondée sur les compétences

Rob Woods,[1] Sim Singh,[2] Brent Thoma,[1] Catherine Patocka,[3] Warren Cheung,[4] Sandra Monteiro,[5] Teresa M Chan,[6] for the QuAL Validation collaborators

[1]Department of Emergency Medicine, University of Saskatchewan, Saskatchewan, Canada; [2]College of Medicine, University of Saskatchewan, Saskatchewan, Canada; [3]Department of Emergency Medicine, University of Calgary, Alberta, Canada; [4]Department of Emergency Medicine, University of Ottawa, Ontario, Canada; [5]Department of Health Research Methods Evidence and Impact, McMaster University, Ontario, Canada; [6]Division of Emergency Medicine and Education & Innovation, Department of Medicine, McMaster University, Ontario, Canada

Correspondence to: Rob Woods, Department of Emergency Medicine, University of Saskatchewan. Room 2689 Royal University Hospital, 107 Hospital Drive, Saskatoon, SK, Canada S7N0W8; email: rob.woods@usask.ca; Twitter: @robwoodsuofs

Published ahead of issue: Aug 16, 2022; published: Nov 15, 2022. CMEJ 2022, 13(6). Available at https://doi.org/10.36834/cmej.74860

Abstract

Background: Competency based medical education (CBME) relies on supervisor narrative comments contained within entrustable professional activities (EPA) for programmatic assessment, but the quality of these supervisor comments is unassessed. There is validity evidence supporting the QuAL (Quality of Assessment for Learning) score for rating the usefulness of short narrative comments in direct observation.

Objective: We sought to establish validity evidence for the QuAL score to rate the quality of supervisor narrative comments contained within an EPA by surveying the key end-users of EPA narrative comments: residents, academic advisors, and competence committee members.

Methods: In 2020, the authors randomly selected 52 de-identified narrative comments from two emergency medicine EPA databases using purposeful sampling. Six collaborators (two residents, two academic advisors, and two competence committee members) were recruited from each of four EM Residency Programs (Saskatchewan, McMaster, Ottawa, and Calgary) to rate these comments with a utility score and the QuAL score. Correlation between utility and QuAL score were calculated using Pearson's correlation coefficient. Sources of variance and reliability were calculated using a generalizability study.

Results: All collaborators (n = 24) completed the full study. The QuAL score had a high positive correlation with the utility score amongst the residents (r = 0.80) and academic advisors (r = 0.75) and a moderately high correlation amongst competence committee members (r = 0.68). The generalizability study found that the major source of variance was the comment indicating the tool performs well across raters.

Conclusion: The QuAL score may serve as an outcome measure for program evaluation of supervisors, and as a resource for faculty development.

Résumé

Contexte : Dans la formation médicale fondée sur les compétences (FMFC), l'évaluation programmatique s'appuie sur les commentaires narratifs des superviseurs en lien avec les activités professionnelles confiables (EPA). En revanche, la qualité de ces commentaires n'est pas évaluée. Il existe des preuves de la validité du score QuAL (qualité de l'évaluation pour l'apprentissage, *Quality of Assessment for Learning* en anglais) pour l'évaluation de l'utilité des commentaires de rétroaction courts lors de la supervision par observation directe.

Objectif : Nous avons tenté de démontrer la validité du score QuAL aux fins de l'évaluation de la qualité des commentaires narratifs des superviseurs pour une APC en interrogeant les principaux utilisateurs finaux des rétroactions : les résidents, les conseillers pédagogiques et les membres du comité de compétence.

Méthodes : En 2020, les auteurs ont sélectionné au hasard 52 commentaires narratifs anonymisés dans deux bases de données d'APC en médecine d'urgence au moyen d'un échantillonnage intentionnel. Six collaborateurs (deux résidents, deux conseillers pédagogiques et deux membres de comités de compétence) ont été recrutés dans chacun des quatre programmes de résidence en médecine d'urgence (Saskatchewan, McMaster, Ottawa et Calgary) pour évaluer ces commentaires à l'aide d'un score d'utilité et du score QuAL. La corrélation entre l'utilité et le score QuAL a été calculée à l'aide du coefficient de corrélation de Pearson. Les sources de variance et la fiabilité ont été calculées à l'aide d'une étude de généralisabilité.

Résultats : Tous les collaborateurs (n=24) ont réalisé l'étude complète. Le score QuAL présentait une corrélation positive élevée avec le score d'utilité parmi les résidents (r=0,80) et les conseillers pédagogiques (r=0,75) et une corrélation modérément élevée parmi les membres du comité de compétence (r=0,68). L'étude de généralisation a révélé que la principale source de variance était le commentaire, ce qui indique que l'outil a fonctionné avec une efficacité égale pour tous les évaluateurs.

Conclusion : Le score QuAL peut servir de mesure des résultats pour l'évaluation des superviseurs par les programmes, et de ressource pour le perfectionnement du corps professoral.

Introduction

Competency-based medical education (CBME) was developed to align the outcomes of training with the needs of society,[1] but requires, among other things, a programmatic assessment model to achieve this goal.[2] The promise of CBME and programmatic assessment entails trainees receive frequent assessments and useful feedback throughout their training. Unfortunately, providing useful feedback to trainees has consistently been identified as a challenge for frontline faculty.[3-5]

Competence by design (CBD), used by Canadian specialty training programs, relies heavily on direct observation feedback organized around entrustable professional activities (EPA).[6] These assessments contain both quantitative and qualitative information captured via a rating scale with entrustment anchors and a free-text narrative comment respectively. Entrustment scores on the rating scale are less discriminating of individual performance in Emergency Medicine (EM),[7] leaving programs to rely heavily on the narrative comments for coaching and assessment.[8-10]

Various stakeholders use the narrative component of EPA assessments for their own purposes. Trainees use them to guide their own development, academic advisors (AA) to inform their coaching of the trainees, and competence committee (CC) members to make decisions on the trainee's progress through the program.[2,11] While EPA assessments are essential to the core business of CBME, the field is only beginning to examine these uses. Given how widely they are utilized in CBME, the need for these comments to be useful or helpful for each stakeholder group is essential. A measure of the usefulness of the supervisor narrative comments contained within them could support continuous improvement in CBME.

The Quality of Assessment for Learning (QuAL) score is a simple three-item scoring tool that was developed in an Emergency Medicine assessment program to evaluate a supervisor's ability to provide a useful narrative comment within a workplace-based assessment.[12] Developed on Messick's validity framework elements, it was subsequently tested through a multi-centre meta-rating exercise. All of the raters were faculty and not all of them were CC members. It is unclear whether it identifies useful narrative comments for each stakeholder group (residents, AAs, and CC members) that use EPA data in graduate education.[11] Additionally, it has not been evaluated for use with the supervisor narrative comments contained specifically within EPA assessments. We conducted this analysis with the goal of building validity evidence for the QuAL score in keeping with Kane's concepts of scoring (the richness of qualitative data) and generalization (consistency of interpretations).[13] This was accomplished by evaluating it as a measure of the usefulness of supervisor narrative comments within EPAs for the three main end-user groups in CBME: residents, AAs, and CC members.

Methods

In 2020, we conducted a multi-centre rating study of EPAs aimed to build validity evidence for the QuAL score to rate the usefulness of supervisor narrative comments contained within EPAs. Please see Figure 1 for a graphical depiction of our study protocol.

Setting. Our study was conducted via an online survey with participant-rater collaborators from four sites (University of Saskatchewan, McMaster University, University of Calgary, and University of Ottawa).

Ethical approval. This study was deemed exempt from ethical review by the Behavioural Research Ethics Boards at both the University of Saskatchewan and McMaster University.[14]

Data selection. Two of our investigators (RW, TC) extracted and de-identified EPA assessments from the databases of the University of Saskatchewan and McMaster University emergency medicine programs. Each EPA assessment included both a rating of performance using entrustment anchors and a narrative comment (Appendix A).[6] The narrative comments were anonymized by replacing names and specific pronouns (he or she) with general pronouns (they). As real trainee assessment data was used within the study, all study authors and survey participants signed a non-disclosure agreement to maintain the confidentiality of the assessment data.

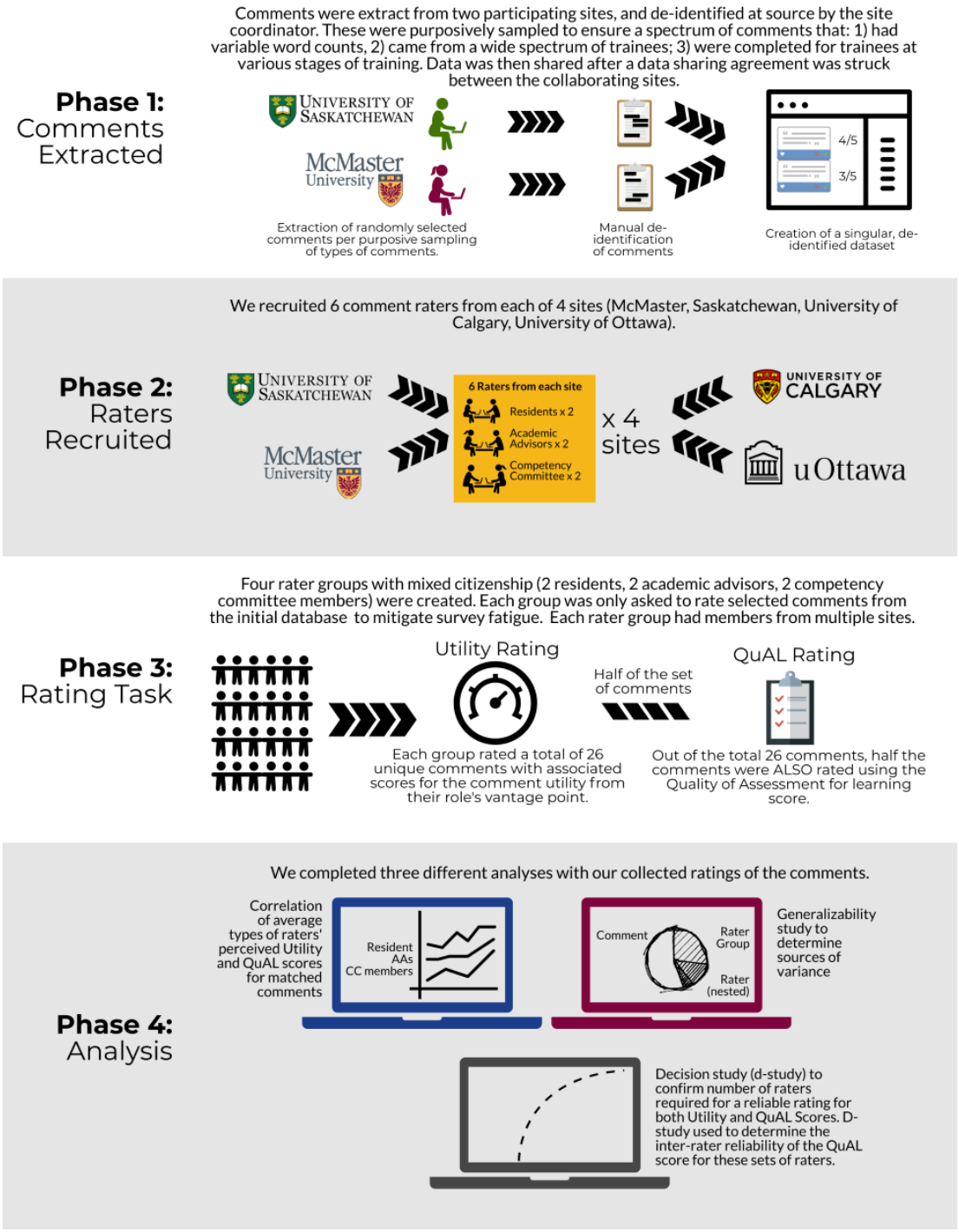


Figure 1. An infographic depicting our study protocol phases and steps

To ensure that a diverse and representative sample of EPAs was selected, 26 categories of EPA assessments were established using stage of training, gender of resident, gender of assessor, and word count[15] and selected using a randomization protocol that was established *a priori*. This process used the random number generator function in Excel (Microsoft Corporation Inc. Washington, DC) to select EPAs from each category (Appendix B). We extracted and de-identified one EPA within each category from each institution, 26 from McMaster and 26 from the University of Saskatchewan (Appendix C). To evaluate the similarity between the sample EPAs and the broader EPAs within the two databases, the frequency of each entrustment score rating was compared (Appendix D).

Survey design. We required participants to complete the survey in a single sitting between June 1st and August 20th, 2020. Survey fatigue was mitigated by capping the number of items rated by each participant. We divided the EPAs into four shorter surveys, each containing a balanced sample of low and high word count narratives. Through their extensive experience of reviewing EPA assessments on competence committees, the authors have found 20 words to be an approximate median and therefore used this number for the cut-off. The raters in each cohort were gender balanced and from more than one institution. Within each survey, we limited the number of raters evaluating each item with the QuAL Score to two due to its demonstrated reliability with two raters.[12] This resulted in four survey versions (Appendix E) that required participants to rate 26 EPA narratives for utility with a 3-point Likert scale.[12] Each participant was asked if the comment was useful for their purpose: residents (informing their development), AAs (coaching over time) and CC members (progress decisions). Then 13 of the same 26 EPA narratives were rated using the three elements of the QuAL Score, resulting in a cumulate single score. (Table 1). One study team member (SS) created the surveys in Survey Monkey (SVMK Inc. San Manteo, California). The rest of the authors piloted the surveys and revised the surveys for content and clarity.

Recruitment of participant-raters. Site leads (CP, WC, TC, RW) recruited six participants (two residents, two AAs, and two CC members each) from each of the participating Universities.

Analysis

Descriptive statistics. Demographic and item characteristics were analyzed using descriptive statistics.

Correlation analysis. Pearson correlations were calculated to determine the strength of the relationship between utility and the QuAL score ratings for all rater cohorts and for each rater subgroup (residents, AAs, CC members).

Generalizability & decision studies. Four rater cohorts containing three subgroups (residents, AAs, CCs) rated *different* sets of EPAs (Appendix E). Using this data, four separate generalizability studies were conducted to evaluate sources of variance in QuAL ratings; one for each of the four rater cohorts.

Table 1. Utility score and Quality of Assessment of Learning (QuAL) score criteria

Component of Survey	Question	Scoring System
Utility Score	Is this a useful comment for the purpose of Development (Resident), Coaching (Academic Advisor) or Progress Decision (Competence Committee)?	Yes (2 points) Maybe (1 point) No (0 points)
QuAL Score	Item 1: Does the rater provide sufficient evidence about performance?	Yes (3 points) Maybe (2 points) No (1 Point) No comment (0 points)
	Item 2: Does the rater provide a suggestion for improvement?	Yes (1 point) No (0 points)
	Item 3: Is the rater's suggestion linked to the behaviour described?	Yes (1 point) No (0 points)

In a generalizability study (G-study), individual variables are called facets. The goal of a g-study is to describe how the data, in this case the QuAL scores, vary in each facet.[16] The facets in this study were the narrative comment, participant rater and participant-rater group. Participant rater group indicated the participant rater's status (AA, CC or resident). The narrative comment was the facet of differentiation, indicating that we hoped to differentiate between narrative comments based on the QuAL score data collected. The participant rater and participant rater group were facets of generalization, indicating that we hope to generalize QuAL scores across raters; ideally different raters agree on the QuAL score assigned to a comment. The participant and participant-rater group was treated as a random facet. Each G-study was cross designed with a nested facet; narrative comment was crossed with each participant rater, who were all nested in their participant rater group (AA, CC, or resident). Subsequent decision studies (D-study) estimated the inter-rater reliability between participant raters and the reliability coefficients with varying numbers of participant raters.

Results

All of the study participant-raters completed their assigned ratings for utility and QuAL. The participant-raters reported that the survey took approximately 75 minutes (*M* = 74, *SD* = 28) to complete. Participant demographics are presented Appendix F.

Our purposeful sampling resulted in a broad range of EPAs extracted across all four stages of training (Appendix C). To demonstrate the similarity between the sample EPAs and the EPAs completed during the 2019-20 academic year from each of the two participating sites, the frequency of each entrustment rating was compared (Appendix D). The scores were found to be relatively similar, with 86.5% having entrustment scores of 4 or 5. The items were found to have a broad range of utility scores with nine items (17.3%) rated as having no utility, 20 items (38.5%) rated as having moderate utility, and 23 items (44.2%) rated as having high utility.

Utility and QuAL correlation

Overall, the QuAL scores had a positive correlation with utility scores across the four rater cohorts (Appendix E). The QuAL Score had a high positive correlation with the utility scores in the rater subgroups of residents (0.80) and AAs (0.75), and had a moderate positive correlation with the utility scores of the rater subgroup of CC members (0.65).[17] Notably, there was no correlation between the utility and QuAL scores for the Cohort 1 CC participants (Table 2).

Table 2. Correlation of the Quality of Assessment of Learning (QuAL) score and utility by end user sub-group

Rater Cohort	Sub-Group			
	Residents	AA	CC	Cohort's Average
1	0.71	0.82	0.26*	0.80
2	0.88	0.84	0.73	0.91
3	0.95	0.83	0.75	0.90
4	0.91	0.86	0.78	0.83
All Raters across all four cohorts	0.80	0.75	0.68	0.85

*Values above are Pearson's R, where 0.9-1 = very high, 0.7-0.9 = high positive, 0.5-0.7 = moderate positive, 0.3-0.5 = low positive, 0.0-0.3 = negligible correlation. All results were significant < 0.001 unless indicated by **

Generalizability theory analysis

In our analysis, we compared the four rater cohorts. Each cohort consisted of two raters from each stakeholder subgroup (residents, AAs, CC members) from different sites. The generalizability theory analysis revealed g-coefficients ranging from 0.79 to 0.95 for the four rater cohorts. A coefficient greater than 0.80 is generally considered the minimum standard for high stakes assessments.[18,19] Although this tool is not used for high stakes assessment of trainees, it will be used to flag poor faculty performance, and similar level of rigor would be valuable. Our study suggests an acceptable level of reliability was achieved with six raters (Table 3 & Supplemental Digital File – Appendix G), however D-studies indicated that a set of six raters is not always required. Table 3 also shows the results of the D-studies evaluating the interrater reliability between any two randomly selected raters had a range of 0.72-0.94. Notably, the major source of variance for each rater cohort was the comment, rather than the raters or the subgroup of origin (AA, CC, residents). Table 3 also outlines the percentage variance contributed by each of the facets in each cohort.

Table 3. Generalizability study results

Cohort	Absolute g-coefficient	Interrater Reliability	Generalizability Study Results (Variance Components)				
			Comment	Rater Group	Rater (nested in group)	Comment x Rater Group	Error
1	0.86	0.77	52.2%	0.0%	10.2%	1.4%	36.2%
2	0.91	0.94	71.3%	7.0%	0.1%	6.8%	14.8%
3	0.79	0.72	42.2%	7.7%	7.2%	0.7%	42.2%
4	0.95	0.90	75.2%	0.0%	0.0%	0.0%	24.8%

Legend:
Absolute g-coefficient (aka Phi coefficient)- indicates the estimated generalizability of scores from this study design to another similar study or universe of scores
Interrater Reliability - d-study estimate of generalizability of scores from one rater to any other potential rater
Comment - % variance contributed by comments for EPAs, based on scores using the QuAL scale
Rater Group - % variance contributed by the 3 types of participant raters: Academic advisors, competency committee member or trainee
Rater - % variance contributed by individual participant-raters as nested within their rater group
Comment x Rater Group - % variance contributed by the interaction between comment and participant rater group
Error - essentially all remaining variability; representing the interaction between QuAL score and participant rater, nested in participant rater group

Discussion

Our study found the QuAL score to have high inter-rater reliability across a broad sample of EPAs in EM and to correlate positively with utility across stakeholder groups including residents, AAs, and CC members. Our findings add to the findings by Chan et al.,[12] adding validity evidence for the QuAL Score as a measure of the usefulness of supervisor narrative comments, specifically in the context of CBME and EPA-based assessments. Specifically, our findings support the *Scoring* inference of a validity argument by demonstrating the usefulness of the narrative comment to multiple stakeholders and the *Generalization* inference through the high reliability of the ratings.[13]

Notably, there was a stronger correlation between the QuAL score and perceived utility for residents and AAs as compared to CC members. This suggests that the QuAL score may be more aligned with the *assessment for learning* paradigm, detecting useful comments more in line with coaching or good feedback. Alternatively, it may mean that individual comments are less useful for CC members since these individuals are usually acting as meta-raters. Therefore, the utility of individual comments is less because they are looking for trends *across* multiple EPAs.[8] This finding is in line with recent literature suggesting that there is an inherent tension that feedback in the moment is advertised as coaching or formative assessment, but also used for progress decisions or cumulative or summative assessment.[20] Because of this, it is unlikely that any one tool will have equal correlation with utility for all end-user groups.

The idea of using narrative comments for assessment can be met with initial reservation due to their perceived subjectivity. They have been described as a form of coded language,[21]however faculty and residents do have the skills to decode them.[22] A study of internal medicine in-training evaluation reports found that faculty can reliability rate trainees after reading only a few sets of narrative comments.[21] Similarly internal medicine residents can rank-order anonymous trainee narrative comments with high reliability. Attending physicians may not appreciate the impact their comments have on the assessment process, although the comments do need to be sufficiently detailed to be useful for this purpose. Providing feedback to faculty on the usefulness of the narrative comments they create may be another way to communicate to faculty that these comments are very important for assessment. Even if we are able to significantly improve the usefulness of supervisor comments within a program, the trend of using increasing amounts of qualitative data brings on new challenges; narrative comments will take more time to analyze and integrate, and they will require skills that current faculty may not possess.[23,24]

Given the importance of having useful supervisor narrative comments within CBME,[1,2] and that the QuAL demonstrates validity evidence for measuring utility, it is a tool that can serve multiple purposes. One aspect of potential program evaluation in CBME is to determine whether or not the narrative comments in a program or at an institution meet a certain standard over time.[25,26] CBME relies heavily on multiple low stakes assessments that form a comprehensive image. If most comments within that program fail to meet a certain standard, it may call into question the validity of competence committee decisions.[13] Since CCs are essentially meta-raters meta-analytic techniques are governed by similar rules of other types of meta-analyses: garbage in, garbage out.[8,23,27] To ensure the validity of group decision making on aggregate data within CCs, we must first ensure that we have a method to determine the rigor of the data these groups will use.

Elements of the QuAL Score could also be used by faculty as a scaffold for constructing high-quality narrative comments.[28] We hope that by providing a three-item mental checklist for narrative comments using the QuAL score, we can support faculty in providing a useful comment, potentially mitigating the phenomenon of hedging or staying quiet in assessment.[4,5] For ongoing quality assurance measures after initial training with EPAs, the QuAL can be used in different ways. Simple force function interventions could be implemented, such as creating two comment boxes titled with the QuAL elements.[29] Report cards of performance through faculty development initiatives have shown improvement in In-Training Evaluation Report completion.[29] The QuAL score could be used for this purpose. In order to create these report cards, we need efficient methods for rating EPA comments, as the number of EPA assessments in CBME is massive.[7,30] Natural language processing models of narrative comments using QuAL may be able to accomplish this.[31] A recently published tool called EFeCT (Evaluation of Feedback Captured Tool) has been described. It was developed in the context of field notes in Family Medicine residency training and they have used non clinician educators to provide the scoring. Training non-clinicians to

rate narrative comments with the QuAL score could also be explored.[32]

Limitations

Our sample of four training sites may not be generalizable outside of EM in Canada. It would have been preferable for all study collaborators to have rated all of the EPAs, however we felt this was not realistic given time constraints. Having raters rate the same set of comments in one sitting may result in raters remembering how they scored a comment for utility and biasing them when using the QuAL score. We had them do all 26 utility ratings first before they knew about the QuAL score mitigating this potential bias.

Despite our enthusiasm for attempting to measure the usefulness of narrative comments as a surrogate for the feedback encounter, this will never represent the entirety of the complex social interaction of workplace assessment. Raters are prone to subjectivity, emotional influence, role conflict and the competing demands of caring for patients while being a coach.[33-35] Additionally, the acceptance of the feedback is variable based on the importance of the relationship between the trainee and assessor as well as the perceived credibility of the assessor.[36]

Because our dataset was de-identified, our raters determined utility without knowing the identity of the author of the EPA; in CBME the trainee will know their rater, and AAs and CC members are likely to know them as well. The influence of relationship and reputation on interpretation of an EPA was not analyzed in our study. Additionally, the residents in our study did not have the perception of personal consequence[37]in assessing the comments as they would in real life, possibly introducing a bias in their ratings.

Conclusions

This study presents evidence for the validity of QuAL scores for determining the usefulness of supervisor narrative comments in EPA assessments for three different end user groups: residents, academic advisors, and competence committee members. It can serve an outcome measure for program evaluation in CBME and as a resource for faculty development.

Conflicts of Interest: SM, CP, and TMC have received an unrelated research funding from the Royal College of Physicians and Surgeons of Canada. WJC and BT are employed by the Royal College of Physicians and Surgeons of Canada as Clinician Educators. RW & SS have no conflicts of interest to declare.

Funding: University of Saskatchewan, College of Medicine Dean's Summer Student Grant.

Acknowledgments: We thank the trainees at the University of Saskatchewan and McMaster University for allowing us to conduct this quality improvement study. We hope that our work will further improve your training experiences in the future.

Author Contributions (based on ICMJE criteria): RW drafted outline of the paper. All of the authors contributed to collecting and analyzing the data. All of the authors contributed to content development. All of the authors contributed to writing and gave final approval to the manuscript.

QuAL Validation Collaborators: Omar Anjum (University of Ottawa, Ottawa, ON, Canada), Nick Bouchard (University of Saskatchewan, Saskatoon, SK, Canada), Savanna Boutin (University of Saskatchewan, Saskatoon, SK Canada), Rob Carey (University of Saskatchewan, Saskatoon, SK Canada) Alexander Chorley (McMaster University, Hamilton, ON, Canada), Nick Costain (University of Ottawa, Ottawa, ON, Canada), Sebastien Dewhirst (University of Ottawa, Ottawa, ON, Canada), Cody Dunne (University of Calgary, Calgary, AB, Canada), Kamini Erker (University of Saskatchewan, Saskatoon, SK Canada), Janet Ferguson (University of Saskatchewan, Saskatoon, SK Canada), Mark Francis (University of Calgary, Calgary, AB, Canada), Mark Hewitt (McMaster University, Hamilton, ON, Canada), Hasheem Kareemi (University of Ottawa, Ottawa, ON, Canada), Jeff Landreville (University of Ottawa, Ottawa, ON, Canada), Lynsey Martin, (University of Saskatchewan, Saskatoon, SK Canada) Shawn Mondoux (McMaster University, Hamilton, ON, Canada); Anjali Pandya (University of Calgary, Calgary, AB, Canada); Alim Pardhan (McMaster University, Hamilton, ON, Canada); Zoe Polsky (University of Calgary, Calgary, AB, Canada), Ian Rigby (University of Calgary, Calgary, AB, Canada); Spencer Sample (McMaster University, Hamilton, ON, Canada); Kari Sampsel (University of Ottawa, Ottawa, ON, Canada); Kelly Van Diepen (McMaster University, Hamilton, ON, Canada), Fareen Zaver (University of Calgary, Calgary, AB, Canada)

Previous related presentations: 2020 Competency-Based Medical Education Program Evaluation Summit, 2021 Canadian Association of Emergency Physicians Conference, 2020 Saskatchewan Health Research Showcase, 2020 University of Saskatchewan Dean's Project Presentations, 2021 Ottawa Student Emergency Medicine Conference.

Data: No dataset from outside academic or hospital-based institutions were used.

References

1. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Med Teach*. 2010 Aug;32(8):638-45. https://doi.org/10.3109/0142159X.2010.501190
2. Lockyer J, Carraccio C, Chan MK, et al. Core principles of assessment in competency-based medical education. *Med Teach*. 2017 Jun 3;39(6):609-16. https://doi.org/10.1080/0142159X.2017.1315082
3. Cheung WJ, Patey AM, Frank JR, Mackay M, Boet S. Barriers and enablers to direct observation of trainees' clinical performance: a qualitative study using the theoretical domains framework. *Acad Med*. 2019 Jan;94(1):101-14. https://doi.org/10.1097/ACM.0000000000002396
4. 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(1):175-88. https://doi.org/10.1007/s10459-015-9622-0
5. Scarff CE, Bearman M, Chiavaroli N, Trumble S. Keeping mum in clinical supervision: private thoughts and public judgements. *Med Educ.* 2019 Feb;53(2):133-42. https://doi.org/10.1111/medu.13728
6. Sherbino J, Bandiera G, Doyle K, et al. The competency-based medical education evolution of Canadian emergency medicine specialist training. *CJEM*. 2020 Jan;22(1):95-102. https://doi.org/10.1017/cem.2019.417
7. Thoma B, Hall AK, Clark K, et al. evaluation of a national competency-based assessment system in emergency medicine: a CanDREAM study. *J Grad Med Educ.* 2020 Aug;12(4):425-34. https://doi.org/10.4300/JGME-D-19-00803.1
8. Chan TM, Sherbino J, Mercuri M. Nuance and noise: lessons learned from longitudinal aggregated assessment data. *J Grad Med Educ.* 2017 Dec;9(6):724-9. https://doi.org/10.4300/JGME-D-17-00086.1
9. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: validity evidence for qualitative educational assessments. *Acad Med*. 2016 Oct;91(10):1359-69. https://doi.org/10.1097/ACM.0000000000001175
10. Ginsburg S, van der Vleuten CPM, Eva KW. The hidden value of narrative comments for assessment: a quantitative reliability analysis of qualitative data. *Acad Med*. 2017 Nov;92(11):1617-21. https://doi.org/10.1097/ACM.0000000000001669
11. Thoma B, Caretta-Weyer H, et al. Becoming a deliberately developmental organization: using competency based assessment data for organizational development. *Med Teach*. 2021 Jul 3;43(7):801-9. https://doi.org/10.1080/0142159X.2021.1925100
12. 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 May 26;32(3):319-29. https://doi.org/10.1080/10401334.2019.1708365
13. 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
14. Tri-Council Policy Statement 2018. Available from: https://ethics.gc.ca/eng/tcps2-eptc2_2018_chapter2-chapitre2.html
15. Bismil R, Dudek NL, Wood TJ. In-training evaluations: developing an automated screening tool to measure report quality. Med Educ. 2014 Jul;48(7):724-32. https://doi.org/10.1111/medu.12490
16. Monteiro S, Sullivan GM, Chan TM. Generalizability theory made simple(r): an introductory primer to g-studies. *J Grad Med Educ*. 2019 Aug 1;11(4):365-70. https://doi.org/10.4300/JGME-D-19-00464.1
17. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. *Anesth Analg*. 2018 May;126(5):1763-8. https://doi.org/10.1213/ANE.0000000000002864
18. 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
19. Vleuten CPM, Norman GR, Graaff E. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ.* 1991 Mar;25(2):110-8. https://doi.org/10.1111/j.1365-2923.1991.tb00036.x
20. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. Med Educ. 2019 Jan;53(1):76-85. https://doi.org/10.1111/medu.13645
21. Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: residents' interpretations of written assessment comments. *Med Educ*. 2017 Apr;51(4):401-10. https://doi.org/10.1111/medu.13158
22. Sebok-Syer SS, Klinger DA, Sherbino J, Chan TM. Mixed messages or miscommunication? investigating the relationship between assessors' workplace-based assessment scores and written comments: *Acad Med*. 2017 Dec;92(12):1774-9. https://doi.org/10.1097/ACM.0000000000001743
23. 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
24. Cheung WJ, Chan TM, Hauer KE, et al. CAEP 2019 Academic Symposium: got competence? best practices in trainee progress decisions. *CJEM*. 2020 Mar;22(2):187-93. https://doi.org/10.1017/cem.2019.480
25. Hodges B. Assessment in the post-psychometric era: Learning to love the subjective and collective. *Med Teach*. 2013 Jul;35(7):564-8. https://doi.org/10.3109/0142159X.2013.789134
26. Chan TM, Paterson QS, Hall AK, et al. Outcomes in the age of competency-based medical education: Recommendations for emergency medicine training in Canada from the 2019 symposium of academic emergency physicians. *CJEM*. 2020 Mar;22(2):204-14. https://doi.org/10.1017/cem.2019.491
27. Chan T, Sebok-Syer S, Thoma B, Wise A, Sherbino J, Pusic M. Learning analytics in medical education assessment: the past, the present, and the future. Promes S, editor. *AEM Educ Train*. 2018 Apr;2(2):178-87. https://doi.org/10.1002/aet2.10087
28. Ginsburg S, Gingerich A, Kogan JR, Watling CJ, Eva KW. Idiosyncrasy in assessment comments: do faculty have distinct

writing styles when completing in-training evaluation reports? *Acad Med*. 2020 Nov;95(11S):S81-8. https://doi.org/10.1097/ACM.0000000000003643

29. Dudek NL, Marks MB, Bandiera G, White J, Wood TJ. Quality in-training evaluation reports-does feedback drive faculty performance? *Acad Med.* 2013 Aug;88(8):1129-34. https://doi.org/10.1097/ACM.0b013e318299394c
30. Zhang R. Automated assessment of medical training evaluation text. *AMIA Annu Symp Proc*. 2012;1459-68.
31. Ötleş E, Kendrick D, Solano QP, et al. Using natural language processing to automatically assess feedback quality: findings from three surgical residencies. *Acad Med*. 2021 May 4; Publish Ahead of Print. Available from: https://journals.lww.com/10.1097/ACM.0000000000004153 [Accessed May 31, 2021].
32. Ross S, Hamza D, Zulla R, Stasiuk S, Nichols D. Development of and preliminary validity evidence for the EFeCT feedback scoring tool*. J Grad Med Educ.* 2022 Feb 1;14(1):71-9. https://doi.org/10.4300/JGME-D-21-00602.1
33. ten Cate O, Regehr G. the power of subjectivity in the assessment of medical trainees: *Acad Med*. 2019 Mar;94(3):333-7. https://doi.org/10.1097/ACM.0000000000002495
34. Gomez-Garibello C, Young M. Emotions and assessment: considerations for rater-based judgements of entrustment. *Med Educ*. 2018 Mar;52(3):254-62. https://doi.org/10.1111/medu.13476
35. Watling C, LaDonna KA, Lingard L, Voyer S, Hatala R. 'Sometimes the work just needs to be done': socio-cultural influences on direct observation in medical training. *Med Educ.* 2016 Oct;50(10):1054-64. https://doi.org/10.1111/medu.13062
36. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: rethinking the etiology of rater errors*: Acad Med.* 2011 Oct;86:S1-7. https://doi.org/10.1097/ACM.0b013e31822a6cf8
37. Ginsburg S, Kogan JR, Gingerich A, Lynch M, Watling CJ. taken out of context: hazards in the interpretation of written assessment comments. *Acad Med*. 2020 Jul;95(7):1082-8. https://doi.org/10.1097/ACM.0000000000003047

Appendices

Appendix A. Example of Entrustable Professional Activity (EPA) entry

Resident Stage of Training	Foundations
Title of EPA	F2: Diagnosis and Management of Uncomplicated Presentations
Description of EPA	The focus of this EPA is the assessment and emergency department management of simple or uncomplicated urgent and non-urgent presentations. These presentations are not complicated by co-existing clinical conditions (e.g. concurrent illness or medical conditions) or patient factors (e.g. communication barriers, access to care etc.) or ED environmental factors (e.g. availability of clinical resources, excessive ED patient volumes etc.) Examples of these types of presentations include, but are not limited to: ● Cough or wheeze ● Musculoskeletal injuries or pain ● Eye complaints ● ENT complaints ● Headache
EPA Context	Emergency Department, Complex Clinical Characteristic, Adult
EPA Score Given	4 - I had to be there just in case.
Narrative Comment	“Trainee managed patient well”

Appendix A1

Resident Stage of Training	Foundations
Title of EPA	F1: Initiating and assisting in resuscitation of critically ill patients
Description of EPA	The focus of this EPA is on early stages of resuscitation based on symptom management of patients, including but not limited to those experiencing cardiorespiratory arrest, dysrhythmias, shock, respiratory distress, or altered mental status. Initial management plans for oxygenation and ventilation, management of blood pressure, and management of critical dysrhythmias are part of this EPA. More complex resuscitation and management after initial threats to life have been addressed is not part of this EPA.
EPA Context	Emergency Department, Complex Clinical Characteristic, Adult
EPA Score Given	4 - I had to be there just in case.
Narrative Comment*	The trainee saw a 44 year old male who was triaged as rectal bleeding Patient had had a hemorrhoidectomy 1 week ago and then presented after many hours of heavy lower go bleeding. The trainee did an initial assessment and found patient to be pale, tachy and slow to respond to questions. The trainee alerted staff to move patient to resus and started IV fluids and called me. The trainee ensured multiple IV's were started, bolused the patient, gave a dose of TXA and spoke with surgery on call. I needed to be there just in case as the trainee wanted to run their initial resuscitation passed me. Feedback going forward - when you had initially called me you stated I need you to come this patient is unwell. That was very appropriate, but helpful if you can provide a bit more of a succinct summary - can you come, this patient is tachy and hypotensive from a significant lower go bleed - I have started fluids and am moving them to resus. They were able to get this done when I asked them to walk we through why the patient was unwell and what they were doing, but helpful to do this when you initially make the call, especially as you transition to more acute care rotations (cardio, ICU where you will need to call fellow/staff).

**Narrative comments are de-identified but otherwise real comments from supervisors. Spelling mistakes and shorthand have been maintained for realism.*

Appendix B. Entrustable Professional Activity randomization categories

School	Stage of Training	Gender Trainee	Gender Rater	Word Count	Randomization
McMaster	TTD	Female	Male	< 21	31
McMaster	TTD	Female	Female	< 21	43
McMaster	TTD	Male	Female	< 21	7
McMaster	TTD	Male	Male	< 21	14
McMaster	Foundations	Female	Male	< 21	47
McMaster	Foundations	Female	Female	< 21	34
McMaster	Foundations	Male	Female	< 21	4
McMaster	Foundations	Male	Male	< 21	6
McMaster	Core	Female	Male	< 21	15
McMaster	Core	Female	Female	< 21	5
McMaster	Core	Female	Male	< 21	17
McMaster	Core	Female	Male	< 21	47
McMaster	Core	Female	Female	< 21	15
McMaster	Core	Female	Male	< 21	10
McMaster	Core	Female	Female	< 21	19
McMaster	Core	Male	Female	< 21	3
McMaster	Core	Male	Male	< 21	8
McMaster	Core	Male	Male	< 21	33
McMaster	Core	Male	Female	>20	15
McMaster	Core	Male	Male	>20	3
McMaster	Core	Male	Female	>20	19
McMaster	Core	Male	Male	>20	5
McMaster	TTP	Female	Male	< 21	18
McMaster	TTP	Female	Female	< 21	11
McMaster	TTP	Male	Female	>20	13
McMaster	TTP	Male	Male	>20	8
Saskatchewan	TTD	Female	Female	>20	6
Saskatchewan	TTD	Female	Male	>20	10
Saskatchewan	TTD	Male	Female	>20	8

Saskatchewan	TTD	Male	Male	>20	1
Saskatchewan	Foundations	Female	Female	>20	14
Saskatchewan	Foundations	Female	Male	>20	15
Saskatchewan	Foundations	Male	Female	>20	2
Saskatchewan	Foundations	Male	Male	>20	2
Saskatchewan	Core	Female	Female	>20	16
Saskatchewan	Core	Female	Male	>20	2
Saskatchewan	Core	Female	Female	< 21	25
Saskatchewan	Core	Female	Male	< 21	12
Saskatchewan	Core	Female	Female	< 21	8
Saskatchewan	Core	Female	Male	< 21	3
Saskatchewan	Core	Female	Female	< 21	6
Saskatchewan	Core	Male	Female	>20	7
Saskatchewan	Core	Male	Male	>20	19
Saskatchewan	Core	Male	Female	>20	11
Saskatchewan	Core	Male	Male	>20	5
Saskatchewan	Core	Male	Female	>20	17
Saskatchewan	Core	Male	Male	>20	16
Saskatchewan	Core	Male	Female	>20	18
Saskatchewan	TTP	Female	Male	< 21	3
Saskatchewan	TTP	Female	Female	< 21	17
Saskatchewan	TTP	Male	Male	>20	13
Saskatchewan	TTP	Male	Female	>20	5

TTD = Transition to Discipline, TTP = Transition to Practice

Appendix C. Breakdown of EPAs extracted

EPA Category	EPA	Number of EPAs (% of total)	Number of EPAs in Category (% of total)
Transition to Discipline	TD1	2 (4%)	8 (15%)
	TD2	6 (12%)	
Foundations	F1	2 (4%)	8 (15%)
	F2	2 (4%)	
	F3	1 (2%)	
	F4	3 (6%)	
Core	C1	2 (4%)	28 (54%)
	C2	2 (4%)	
	C3	5 (10%)	
	C4	1 (2%)	
	C5	5 (10%)	
	C6	3 (6%)	
	C8	2 (4%)	
	C9	3 (6%)	
	C13	2 (4%)	
	C14	1 (2%)	
	C17	2 (4%)	
Transition to Practice	TP1	2 (4%)	8 (15%)
	TP2	2 (4%)	
	TP5	3 (6%)	
	TP6	1 (2%)	

Appendix D. Entrustment ratings of extracted comments compared to datasets

Entrustment Score	Items Extracted (*n* = 52)	McMaster Overall Entrustment Distribution in 2018-2019 (*n* = 2507)	University of Saskatchewan Overall Entrustment Distribution in 2018-2019 (*n* = 2103)
I didn't need to be there (5)	48.1% (*n* = 25)	47.0% (*n* = 1179)	50.6% (*n* = 1064)
I needed to be there just in case (4)	38.5% (*n* = 20)	40.9% (*n* = 1026)	31.8% (*n* = 669)
I needed to prompt (3)	9.6% (*n* = 5)	10.4% (*n* = 260)	13.7% (*n* = 288)
I had to talk them through (2)	3.8% (*n* = 2)	1.5% (*n* = 38)	3.2% (*n* = 67)
I had to do (1)	0% (*n* = 0)	0.2% (*n* = 4)	0.7% (*n* = 15)

Appendix E. Survey distribution

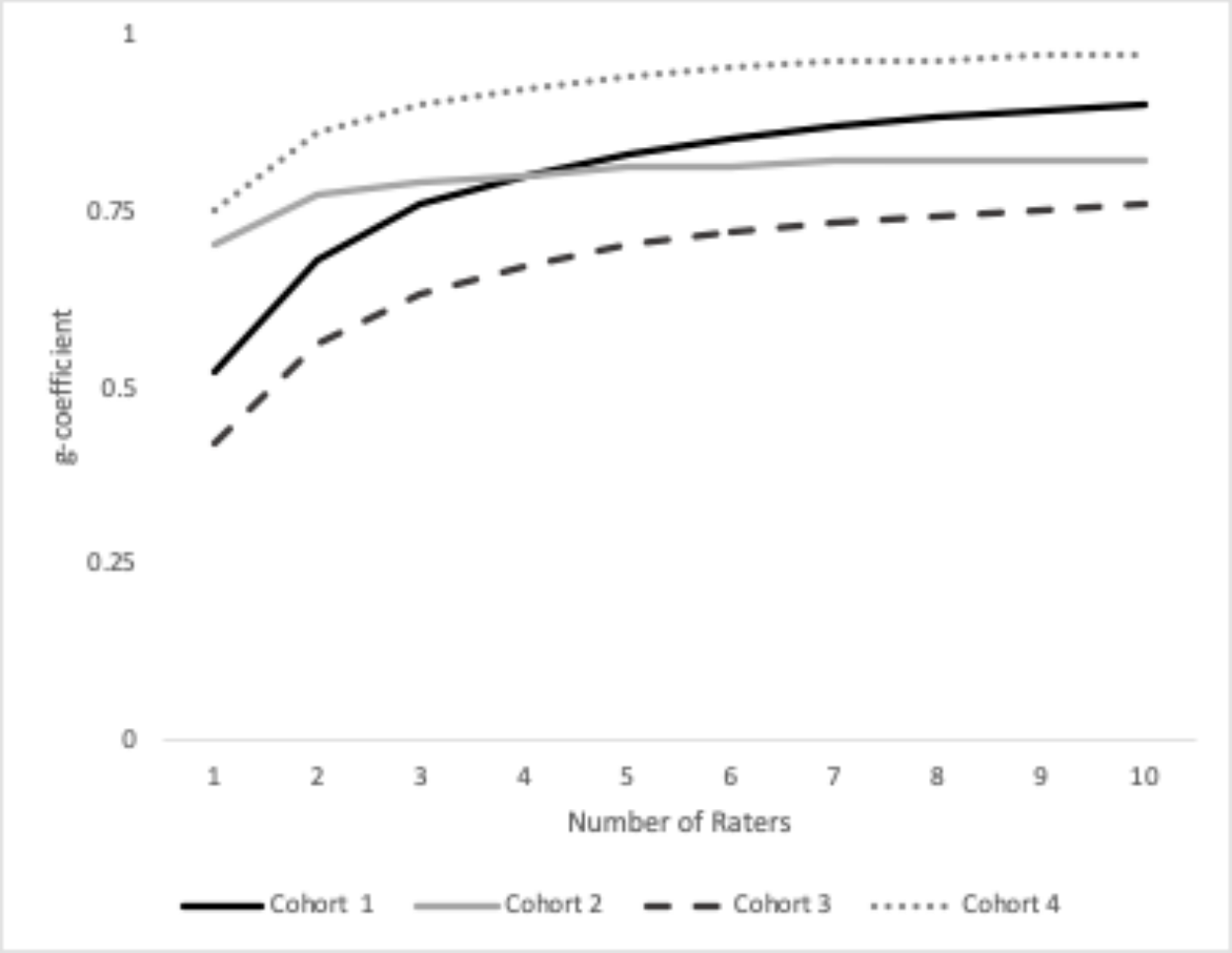
Rater Cohort	EPA comments rated for utility	EPA comments rated using QuAL	Participant Rater Sub-Groups		
			Residents	Academic Advisors	Competence Committee Members
1	1-26	1-13	2 (Ottawa, McMaster) [2 males]	2 (Ottawa, Saskatchewan) [1 male, 1 female]	2 (McMaster, Calgary) [1 male, 1 female]
2		14-26	2 (Saskatchewan, Calgary) [2 males]	2 (McMaster, Calgary) [1 male, 1 female]	2 (Saskatchewan, Ottawa) [1 male, 1 female]
3	27-52	27-39	2 (McMaster, Calgary) [1 male, 1 female]	2 (McMaster, Calgary) [1 male, 1 female]	2 (Calgary, McMaster) [1 male, 1 female]
4		40-52	2 (Saskatchewan, Ottawa) [1 male, 1 female]	2 (Saskatchewan, Ottawa) [1 male, 1 female]	2 (Saskatchewan, Ottawa) [2 males]

Appendix F - Demographics of survey participant-raters

Factor	Residents (*n* = 8)		Academic Advisors (*n* = 8)		Competence Committee members (*n* = 8)	
Age (years-of-age)	21-30	100% (*n* = 8)	21-30	0% (*n* = 0)	21-30	13% (*n* = 1)
	31-40	0%	31-40	75% (*n* = 6)	31-40	64% (*n* = 5
	41-50	0%	41-50	25% (*n* = 2)	41-50	25% (*n* = 2)
Gender	Female	25% (*n* = 2)	Female	50% (*n* = 4)	Female	38% (*n* = 3)
	Male	75% (*n* = 6)	Male	50% (*n* = 4)	Male	62% (*n* = 5)
Training Demographics	PGY1	25% (*n* = 2)	CCFP-EM	0% (*n* = 0)	CCFP-EM	13% (*n* = 1)
	PGY2	37.5% (*n* = 3)	FRCPC	100% (*n* = 8)	FRCPC	87% (*n* = 7)
	PGY3	37.5% (*n* = 3)	Other	0% (*n* = 0)	Other	0% (*n* = 0)
Participation in Program Leadership	Residency Training Committee Member	37.5% (*n* = 3)	Residency Training Committee Member	0% (*n* = 8)	Residency Training Committee Member	50% (*n* = 4)
	Competence Committee	25% (*n* = 2)	Program Director	13% (*n* = 1)	Program Director	13% (*n* = 1)
			Assistant Program Director	0% (*n* = 0)	Assistant Program Director	25% (*n* = 2)

FRCPC = Fellow of the Royal College of Physicians & Surgeons of Canada, CCFP-EM = Certificate of the College of Family Physicians – Emergency Medicine

Appendix G. Decision study analysis



A plot of the decision-study (D-study) analysis to show how reliability changes with the total number of raters, from 1 to 10. In this study, each cohort rated a different set of comments so each cohort is plotted separately.