# Evaluating Feedback Comments in Entrustable Professional Activities: A Cross-Sectional Study

Journal of Medical Education and Curricular Development
Volume 11: 1–7

DOI: 10.1177/23821205241275810

Vasiliki Andreou[1], Sanne Peters[1,2], Jan Eggermont[3] and Birgitte Schoenmakers[1]

[1]Department of Public Health and Primary Care, Academic Centre for General Practice, KU Leuven, Leuven, Belgium. [2]School of Health Sciences, Faculty of Medicine, Dentistry and Health Sciences, The University of Melbourne, Melbourne, Australia. [3]Department of Cellular and Molecular Medicine, KU Leuven, Leuven, Belgium.

## ABSTRACT

**INTRODUCTION:** Competency-based medical education (CBME) has transformed postgraduate medical training, prioritizing competency acquisition over traditional time-based curricula. Integral to CBME are Entrustable Professional Activities (EPAs), that aim to provide high-quality feedback for trainee development. Despite its importance, the quality of feedback within EPAs remains underexplored.

**METHODS:** We employed a cross-sectional study to explore feedback quality within EPAs, and to examine factors influencing length of written comments and their relationship to quality. We collected and analyzed 1163 written feedback comments using the Quality of Assessment for Learning (QuAL) score. The QuAL aims to evaluate written feedback from low-stakes workplace assessments, based on 3 quality criteria (evidence, suggestion, connection). Afterwards, we performed correlation and regression analyses to examine factors influencing feedback length and quality.

**RESULTS:** EPAs facilitated high-quality written feedback, with a significant proportion of comments meeting quality criteria. Task-oriented and actionable feedback was prevalent, enhancing value of low-stakes workplace assessments. From the statistical analyses, the type of assessment tool significantly influenced feedback length and quality, implicating that direct and video observations can yield superior feedback in comparison to case-based discussions. However, no correlation between entrustment scores and feedback quality was found, suggesting potential discrepancies between the feedback and the score on the entrustability scale.

**CONCLUSION:** This study indicates the role of the EPAs to foster high-quality feedback within CBME. It also highlights the multifaceted feedback dynamics, suggesting the influence of factors such as feedback length and assessment tool on feedback quality. Future research should further explore contextual factors for enhancing medical education practices.

**KEYWORDS:** general practice, competency-based medical education, entrustable professional activities, medical education, postgraduate medical education, feedback, assessment

**RECEIVED:** March 28, 2024. **ACCEPTED:** July 4, 2024.

**TYPE:** Original Research Article

**FUNDING:** The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Foundation Flanders (FWO) under Grant [S003219N]-SBO SCAFFOLD.

**DECLARATION OF CONFLICTING INTERESTS:** The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

**CORRESPONDING AUTHOR:** Vasiliki Andreou, Department of Public Health and Primary Care, Academic Centre for General Practice, KU Leuven, Kapucijnenvoer 7 -Box 7001, 3000 Leuven, Belgium.
Email: vasiliki.andreou@kuleuven.be

## Introduction

Competency-based medical education (CBME) has brought a profound transformation in postgraduate medical education.[1,2] Departing from traditional time-based curriculum approaches, CBME emphasizes trainees' progression based on attaining competencies.[3,4] The successful implementation of CBME requires incorporating new methods of assessment that emphasize assessment continuity, enabling the connection between assessment and learning activities and ensuring competency of new physicians.[5] These new assessment methods should also provide enough meaningful information stimulating trainees' further development, and empower trainees to actively engage in their assessment process.[4,6,7]

To operationalize CBME, the concept of Entrustable Professional Activities (EPAs) has been introduced as a novel assessment method.[8] EPAs are concrete tasks or responsibilities at the heart of a medical profession, where trainees should demonstrate competence in order to be entrusted.[9,10] Unlike competencies, EPAs reflect real-world clinical activities and offer a more intuitive framework to assessors and trainees. By delineating concrete tasks and allowing observation of performance in clinical settings, EPAs could facilitate meaningful feedback.[11,12] Typically, EPAs consist of numerical entrustment scales and written feedback to support further development.[13] Through written feedback, trainees receive necessary information about their performance, identifying strengths and areas for improvement, not only within the scope of EPAs but across their broader development.[14] This iterative process is meant to enhance quality of feedback and facilitate its documentation.[9,11]

To promote further learning, written feedback of high-quality should be aligned with the EPA task, contain enough information on performance, and indicate areas for improvement.[11,14] High-quality written feedback comments can

provide valuable context to entrustment scales and performance that numerical scores alone cannot capture.[15,16] Despite the indisputable significance of written feedback in learning, there is little evidence regarding feedback within EPAs.[17] Therefore, there is a need for examining whether EPAs influence the quality of feedback. This study aims to investigate the characteristics and factors that influence the quality of feedback provided within EPAs.

## Methods

### *Organizational Setting*

To establish a standardized curriculum across Flanders, 4 universities (KU Leuven, Ghent University, University of Antwerp, and Free University Brussels) have collaboratively designed a General Practitioner's (GP) Training. The GP Training lasts for 3 years and is structured into 3 phases. The first phase entails a 12-month traineeship in a GP practice, followed by a 6-month hospital traineeship in the second phase, and concluding with an 18-month traineeship in a GP practice for the third phase. Complementing clinical traineeships, trainees are required to attend classes at their respective home universities and actively participate in peer-learning groups facilitated by university tutors.

The practical coordination and decision-making responsibilities regarding this curriculum were entrusted to the Interuniversity Centre for GP Training (ICGPT). Among its responsibilities, the ICGPT is accountable for tasks such as allocating clinical traineeships, administering examinations, preparing trainers for their role, safeguarding quality of training in clinical practice, and managing trainees' learning e-portfolios, where their competencies are assessed and recorded.

### *Educational Setting and Participants*

In 2018, the ICGPT initiated a transition to CBME, by incorporating the Canadian Medical Education Directives for Specialists (CanMEDS) roles into the curriculum. This integration prompted stipulation of assessment guidelines for the clinical traineeships. Trainees were required to have 3 high-stakes evaluations based on the CanMEDS roles on yearly basis, with their workplace trainer and their university-tutor, respectively. Additionally, to facilitate low-stakes workplace-based assessments, trainees were expected to document 5 video-consultations, subsequently evaluated by their trainers. An institutional web-based portfolio was developed to streamline the assessment process.

In 2022, in collaboration with the ICGPT, we introduced an EPA based framework to enhance workplace-based assessment.[18] In total, we developed 60 EPAs covering different care contexts specific for Primary Care (ie, short-term care, chronic care, emergency care, palliative care, elderly care, care for children, mental healthcare, prevention, gender related care, and practice management). For more complicated EPAs, we defined behavior anchors that trainees should demonstrate to be entrusted with a given EPA. An example of such an EPA under short-term care is "Clarify the need for assistance (Ideas, Concerns, Expectations), take a medical history (focused on the diagnostic landscape), conduct a targeted physical examination, and arrive at a diagnosis or a 'working hypothesis'." The accompanying behavior anchors of this EPA are (1) Take a medical history, focused on the tract or part thereof related to the need for assistance or complaint. (2) Conduct a physical examination. (3) Based on the above: make a diagnosis or arrive at a refined 'working hypothesis'. (4) Decide whether additional diagnostics are needed or if treatment and policy should be continued.

Concretely, EPAs were operationalized through a form available in the e-portfolio.[18] EPAs could be assessed through direct observation, video-observation, and/or case-based discussions. Both trainees and trainers could initiate an EPA assessment, by filling in the form choosing which EPAs needed to be assessed. Then, they had to choose an entrustability level based on the Ottawa Surgical Competency Operating Room Evaluation (O-SCORE), ranging from "I had to intervene" to "Supervising others during this EPA."[19] Moreover, 2 feedback fields were available encouraging positive ("What goes well") and negative feedback ("What needs improvement"). Once trainees completed an EPA form, trainers were notified in their e-portfolio accounts that an EPA assessment needs approval. We used simple random sampling to recruit our participants.[20] We recruited on voluntary basis dyads of trainees who were in the first phase of the GP Training, along with their trainers. Trainees and trainers were asked in their e-portfolio through a pop-up window whether they wanted to participate in the study.

### *Study Design*

We employed a cross-sectional design, using quantitative content analysis to indicate quality of feedback from EPA assessments, and a correlation analysis and a multiple regression analysis.[21,22] We used the Quality of Assessment for Learning (QuAL) score to analyze and code the feedback comments, along with characters number as a quality proxy.[23] Prior studies have established validity evidence of the QuAL score for evaluating short feedback comments produced during workplace-based assessments.[23,24] The QuAL score comprises 3 key criteria as illustrated in Figure 1. The first component pertains to whether and to what extent the feedback provides evidence about the trainees' performance. The second component focuses on the presence of suggestions for improvement, while the third component asks whether these suggestions are linked to the described behavior.[23]

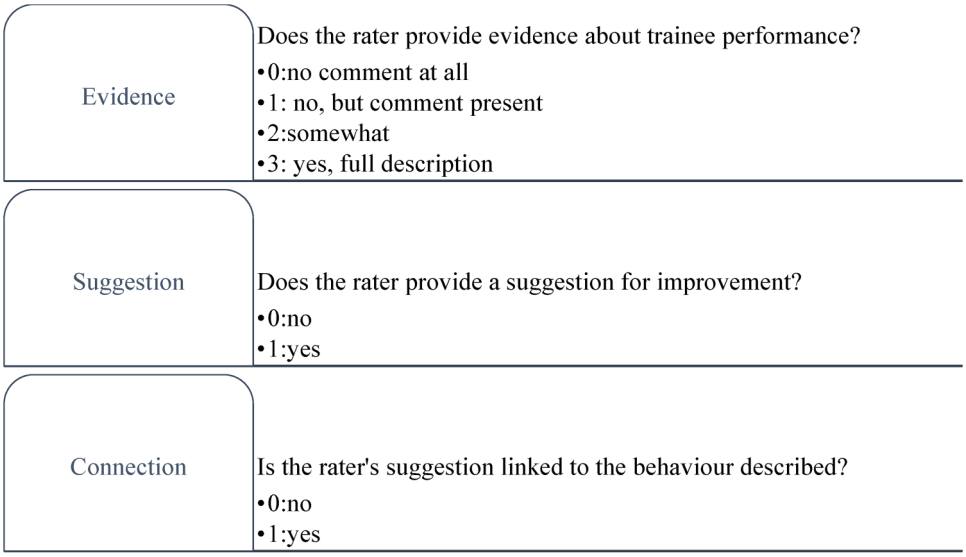


**Figure 1.** Structure of quality of assessment for learning (QuAL) Score.[23]

### *Data Collection and Analysis*

In collaboration with the developers of the e-portfolio, we collected feedback comments registered in trainees' e-portfolios, after completing EPA assessments. The EPAs were assessed by workplace trainers, spanning the period of December 2022 until September 2023. All feedback comments were anonymous to limit potential bias. Initially, we examined all feedback comments and familiarized ourselves with the content and the QuAL scoring scale. We analyzed the data per EPA assessment for each QuAL component. Two raters independently scored a pilot of 120 comments. Throughout the pilot scoring process, we maintained documentation of decisions to enhance methodological rigor. Afterwards, we compared the results, discussed and reviewed ambiguities.[25] After completion of the coding, we aggregated the scores of each QuAL component to decide overall quality. Overall quality ranged from 0 (no quality) to 5 (most quality). We also calculated inter-rater reliability based on intraclass correlation coefficient (ICC) to ensure consistency between the 2 raters.[26] We used Microsoft Excel to code and analyze the feedback comments.

Furthermore, we used non-parametric Spearman's correlation coefficients ($r_s$), since our data violated the assumption of normality. We investigated the relationship between the total quality score on QuAL, number of characters that each feedback comment contained, the entrustment level, and the type of assessment tool used.[27,28] By doing so, we could discover possible relationships and patterns in the data. We recoded the type of assessment tool into a dummy variable (1: Video observation, 2: Direct observation, 3: Case-based discussion). Based on the results of the Spearman's correlations, we performed a hierarchical multiple regression analysis to investigate which parameters influence the number of characters in feedback comments. Variables demonstrating stronger correlations with the outcome variable, namely the number of characters, were prioritized, specifically, total quality score and type of assessment tool. We opted not to include entrustment level on EPAs, because of its weak correlation with the outcome variable. We used number of characters as the outcome variable and the identified predictors based on Spearman's correlations. All statistical analyses were performed in IBM SPSS Statistics (Version 29). The level of significance was set at $P < .05$. The reporting of this study conforms to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement[29] (Supplemental File 1).

## Results

### *Descriptive Statistics*

In total, 199 trainees and trainers agreed to participate. We collected 1163 feedback comments from 1163 EPA assessment forms. Case-based discussions were used 419 times as an assessment tool, observation of performance via video was used 410 times, while 334 direct observations took place. The average number of characters was 400 characters per comment, the average total quality score was 3.63, and the average entrustment score was 3.62 out of 5 (Table 1).

### *Quality of Feedback Comments*

During the coding stage, the feedback comments were coded and assessed according to the QuAL criteria: evidence, suggestion, and connection. Intraclass correlation coefficients fell within an acceptable range (0.5-0.8), indicating consistency and agreement between the 2 raters (Table 2).[26] Because ICC for connection was close to the lower acceptable range,

we discussed it afterwards to solve discrepancies. Most feedback comments provided scored high on evidence about trainees' performance (n = 555), included suggestions for improvements (n = 790) and linked these suggestions to the behavior described (n = 713) (Figure 2).

The feedback comments were categorized into 3 distinct quality levels based on their total quality scores: low, moderate, and high. Low-quality comments received scores ranging from 0 (n = 21) to 1 (n = 96), moderate-quality comments scored between 2 (n = 163) and 3 (n = 155), while high-quality comments scored between 4 (n = 308) and 5 (n = 420).

Feedback comments of low quality received a score of either 0 or 1 in evidence, 0 in suggestion, and 0 in connection. They either lacked any comments or contained evidence that was insufficiently clear or relevant to the assessed EPA. The following example illustrates a low quality feedback comment given on the EPA "Clarifies the help request (Ideas, Concerns, Expectations), takes medical history (focused on diagnostic landscape), performs a targeted physical examination, reaches at a diagnosis or 'working hypothesis'": "*communication, structure of medical history*" (comment 29).

Moderate quality feedback comments received a score of 2 in evidence, or a combined score of 1 to 2 in evidence and 1 in suggestion. However, they did not link the provided suggestion to the behavior being assessed. This is an example of a feedback comment given for the same EPA as above:

> "[Anonymized] conducts a comprehensive medical history taking and fills in correctly (patient's) Electronic Medical Record. [Anonymized] sometimes still doubts when it comes to musculoskeletal pathology. However, [anonymized] has looked up and studied more in the meanwhile. I already notice an improvement." (comment 65)

High-quality feedback comments received a combined score of 2 to 3 in evidence, 1 in suggestion and 1 in connection. They included evidence on trainees' performance, had suggestions for improvements, and linked those to the assessed EPA behavior. The following feedback comment had a score of 5 on the same EPA as above:

> "Clear and comprehensive medical history taking with open and closed questions. Understanding of the complaint. Clear explanation about the clinical findings. Clear communication of the diagnosis. Correct and well-communicated treatment plan. Presentation of alarm symptoms and discussion of further course. Overall, very good and thorough consultation with a lot of understanding, empathy, and patience for the patient. (Improvement) Mixing open and closed questions at times. Asking useless questions with no added value for diagnosis or differential diagnosis (possibly influenced by filming oneself and thus appearing somewhat artificial?). Information (given to the patient) sometimes overly detailed, leading to the loss of the main message by the end of the consultation." (comment 7)

**Table 1.** Descriptive statistics.

| | MEAN | STANDARD DEVIATION | VARIANCE |
|---|---|---|---|
| Number of characters | 400.14 | 396.658 | 157 337.821 |
| Total quality score on QuAL | 3.63 | 1.399 | 1.958 |
| Entrustment score on O-SCORE | 3.62 | .966 | .933 |

**Table 2.** Intraclass correlation coefficients (ICC) measuring inter-rater agreement.

| QuAL CRITERION | ICC |
|---|---|
| Evidence | .649 |
| Suggestion | .791 |
| Connection | .550 |

## *Correlations*

Number of characters was significantly and positively related to total quality score ($r_s = .344$). The correlation between number of characters and entrustment score given for an EPA was also

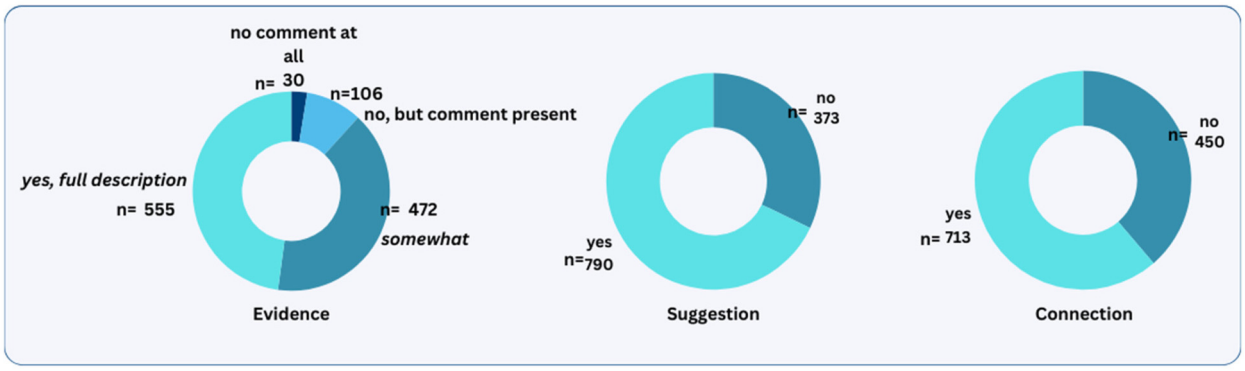


**Figure 2.** Distribution of quality ratings of feedback comments across the QuAL scale.

significant, but weak ($r_s = .079$). There was also a significant correlation between number of characters and type of assessment tool used for the EPA ($r_s = -.237$). In Table 3, the findings related to Spearman's correlations are displayed and the most important findings are discussed. Moreover, entrustment score was significantly related to the type of assessment tool used for the EPAs ($r_s = -.082$). There was no significant correlation between total quality score and given entrustment score, $r_s = -.02$, $P = .503$.

### *Multiple Regression Model*

From the regression analysis, a $R^2$ value of .155 obtained suggesting that the predictors included in the model collectively explained approximately 15.5% of the variability observed in the lengths of written feedback (Table 4). Moreover, the overall regression model was statistically significant, $F = 106.369$, $P < .001$, indicating that the predictors significantly contributed to the prediction of the number of characters in feedback comments. For every one unit increase in quality score, the dependent variable increased by 90.278 ($P < .001$), suggesting that feedback quality has a positive effect on feedback length. Additionally, the regression coefficients for the type of assessment tool was −90.523 ($P < .001$), indicating a significant effect of the assessment tool used on number of characters.

## Discussion

In this study, we analyzed a set of feedback comments derived from EPAs, with the dual aim of evaluating and examining their quality, and investigating factors influencing the length of the feedback provided. Our findings suggest that EPAs can generate high-quality formative feedback, contributing to the literature about EPA assessments and providing evidence on feedback quality. Remarkably, a substantial proportion (approximately 63%) of the feedback comments had high-quality scores, fulfilling all 3 quality criteria (evidence, suggestion, and connection).

Consistent with prior research, our results indicated that most feedback comments were task-oriented and actionable.[30,31] Notably, around 62% of these comments included concrete suggestions for improving the assessed behavior. By enhancing actionable feedback, EPAs can play a pivotal role in promoting learning growth, particularly in the context of low-stakes assessments in the workplace.[32] Furthermore, our findings suggest that EPAs can serve as valuable tools in supporting implementation of CBME, by linking learning activities to assessment.[6,7]

Nevertheless, we did not find a significant correlation between entrustment score and feedback quality. A potential explanation is that quality of feedback does not directly reflect trainees' performance. Registered feedback might be limited

**Table 3.** Nonparametric correlations (Spearman's correlation coefficients).[a]

| | TOTAL QUALITY SCORE ON QuAL | ENTRUSTMENT SCORE ON O-SCORE | ASSESSMENT TOOL | NUMBER OF CHARACTERS |
|---|---|---|---|---|
| Total quality score on QuAL | 1 | −.02 [−0.79, .035] | −.138[b] [−.196, −.078] | .344[b] [.286, .406] |
| Entrustment score on O-SCORE | | 1 | −.082[b] [−.135, −.025] | .079[b] [0.34, .120] |
| Assessment tool | | | 1 | −.237[b] [−.281, −.191] |

[a]BCa bootstrap 95% confidence intervals reported in brackets.
[b]$P < .01$.

**Table 4.** Regression Results Using Total Score on QuAL Score as the Criterion.[a]

| PREDICTOR | *B* | *b* 95% CI [LL, UL] | β | *P* | FIT |
|---|---|---|---|---|---|
| (Intercept) | 254.677[b] | [23.663, 260.115] | | <.001[c] | |
| Total quality score on QuAL | 90.278[b] | [75.082, 105.475] | .318 | <.001[c] | |
| Assessment tool | −90.523[b] | [−115.643, −65.404] | −.193 | <.001[c] | |
| | | | | | $R^2 = .155$[b] |

[a]b represents unstandardized regression weights, beta indicates the standardized regression weights, LL and UL indicate the lower and upper limits of a confidence interval, respectively, *P* represents the *P* value.
[b]$P < .01$.
[c]$P < .05$.

due to other competing needs and increased demands of clinical work.[33] Another explanation for this finding is the impact of using numerical scales for entrustment. Trainers might have chosen a higher entrustment level than they should have, in order not to disappoint their trainees.[34] In the context of low-stakes assessment, using numbers might hamper the learning purpose of assessment, and blur the stakes for the stakeholders involved.[35] This finding suggests that the use of numbers might not be beneficial for low-stakes assessments and aligns with rising concerns about the sustainability of entrustment scales.[34,36]

Also, our findings illustrate the multifaceted nature of feedback quality, indicating an interplay of various influential factors. Despite our regression model explaining only 15.5% of variability, it shed light on crucial feedback dynamics. First, the length of feedback comments is influenced by the quality, suggesting that longer and more extensive comments tend to be of higher quality. Interestingly, our results indicate that the tool used in EPAs is also relevant for both the length and quality of feedback. Specifically, video or direct observations offer higher quality of feedback in comparison to case-based discussions. This echoes the importance of observing trainees during clinical trainings and strengthens observations as a key strategy in CBME implementation.[37–39]

Finally, future research should focus on comparing the quality of feedback provided by EPAs with feedback provided by other forms of workplace-based assessments. Such a comparison could help to further contextualize the strengths and weaknesses of EPAs compared to other assessment methods, providing a more comprehensive understanding of how different assessment tools impact feedback quality within a workplace context. Additionally, this could identify best practices and areas for improvement across various assessment methodologies.

### *Limitations*

While our study had significant strengths in terms of sampling adequacy and rigorous analysis, certain limitations need to be acknowledged. Due to the fact that we relied on a third party for data collection, we might have missed some variables relevant to feedback quality. For instance, other contextual factors, such as frequency of EPAs or trainers' experience, could also play a significant role on feedback quality. Future research should seek to explore how these variables relate to feedback quality. One more potential limitation is the voluntary basis for recruitment, which may have introduced selection bias. Given that participation was voluntary, it is possible that the dyads that chose to participate were more motivated to use the EPA framework, potentially leading to a higher proportion of high-quality feedback comments. Furthermore, although the QuAL scale comprises 3 key criteria, it might not capture all dimensions of feedback quality comprehensively. Nonetheless, we selected this scale due to its established validity and structured criteria.[40,41] Also, the QuAL scale is constructed to evaluate feedback in workplace-based assessments, aligning with our study setting.[23]

## Conclusion

Given the importance of EPAs within CBME, this study provides evidence on the potential of the EPAs to foster high-quality feedback. In this study, EPA feedback comments contained evidence of performance and improvements linked to the assessed behavior, highlighting the potential of EPAs to foster formative feedback leading to learning growth. Furthermore, this study shed light onto the multifaceted feedback dynamics, with findings indicating the influence of factors such as feedback length and type of assessment tool. Consequently, this study contributes evidence to the literature on EPAs and their role in low-stakes assessments.

To further investigate feedback challenges, future research should explore the impact of contextual factors on feedback quality, building upon the results of this study to enhance medical education practices. Also, future research should seek to replicate these findings in another context and, eventually, include a wider scope of influential factors beyond those explored in this study. Such factors could be frequency or complexity of EPAs, trainers' experiences, and the relationship between trainee and trainer. Also, further studies could explore the discrepancy between entrustment score and feedback quality. It is necessary to understand how the use of numerical scales for entrustment can affect feedback quality and explore whether there are alternative methods. Finally, the use of video observations in EPAs is worth exploring further. Technological advancements could offer possible solutions and advancements in medical education.

## Acknowledgments

The authors would like to thank Mr Guy Gielis, Mrs. An Stockmans, Mrs. Fran Timmers, and Mrs Karolina Bystram from the Interuniversity Center for GP Training that facilitated this study. The authors would also like to thank Mr Karel Verbert and Mrs. Cindy Rossi (Imengine-Medbook) for facilitating data collection through the Medbook e-portfolio. Finally, the authors would like to thank and acknowledge Prof d.Martin Valcke and Dr Mieke Embo for facilitating this study through the SBO SCAFFOLD project (www.sbo-scaffold.com).

## Author Contributions

All authors contributed to designing the study. VA wrote the protocol, led the data analysis, and wrote this manuscript. BS participated in the data analysis as a second coder to achieve investigator triangulation. SP, BS, and JE contributed to the critical revision of the paper. All authors have read and approved the manuscript.

## Data Availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

## Ethical Approval

This study was approved by the KU Leuven Social and Societal Ethics Committee G-2022-5615-R2(MIN), and all participants signed an informed consent prior to participation.

## ORCID iDs

Vasiliki Andreou https://orcid.org/0000-0002-6679-0514
Birgitte Schoenmakers https://orcid.org/0000-0003-1909-9613

## Supplemental Material

Supplemental material for this article is available online.

## REFERENCES

1. Frank JR, Snell LS, Cate OT, et al. Competency-based medical education: theory to practice. *Med Teach*. 2010;32(8):638-645.
2. Frank JR, Snell L, Englander R, Holmboe ES. Implementing competency-based medical education: moving forward. *Med Teach*. 2017;39(6):568-573.
3. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting paradigms: from Flexner to competencies. *Acad Med*. 2002;77(5):361-367.
4. Carraccio CL, Englander R. From Flexner to competencies: reflections on a decade and the journey ahead. *Acad Med*. 2013;88(8):1067-1073.
5. Caverzagie KJ, Nousiainen MT, Ferguson PC, et al. Overarching challenges to the implementation of competency-based medical education. *Med Teach*. 2017;39(6):588-593.
6. Nousiainen MT, Caverzagie KJ, Ferguson PC, Frank JR. Implementing competency-based medical education: what changes in curricular structure and processes are needed? *Med Teach*. 2017;39(6):594-598.
7. Lockyer J, Carraccio C, Chan M-K, et al. Core principles of assessment in competency-based medical education. *Med Teach*. 2017;39(6):609-616.
8. Carraccio C, Martini A, Van Melle E, Schumacher DJ. Identifying core components of EPA implementation: a path to knowing if a complex intervention is being implemented as intended. *Acad Med*. 2021;96(9):1332-1336.
9. Ten Cate O, Chen HC, Hoff RG, Peters H, Bok H, van der Schaaf M. Curriculum development for the workplace using entrustable professional activities (EPAs): AMEE guide no. 99. *Med Teach*. 2015;37(11):983-1002.
10. Ten Cate O, Taylor DR. The recommended description of an entrustable professional activity: AMEE guide no. 140. *Med Teach*. 2021;43(10):1106-1114.
11. Peters H, Holzhausen Y, Boscardin C, Ten Cate O, Chen HC. Twelve tips for the implementation of EPAs for assessment and entrustment decisions. *Med Teach*. 2017;39(8):802-807.
12. Ferguson PC, Caverzagie KJ, Nousiainen MT, Snell L, collaborators I. Changing the culture of medical training: an important step toward the implementation of competency-based medical education. *Med Teach*. 2017;39(6):599-602.
13. 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;96(7S):S81-SS6.
14. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. *Perspect Med Educ*. 2015;4(6):284-299.
15. Watling C, Driessen E, van der Vleuten CP, Lingard L. Learning culture and feedback: an international study of medical athletes and musicians. *Med Educ*. 2014;48(7):713-723.
16. Ginsburg S, van der Vleuten CP, Eva KW, Lingard L. Cracking the code: residents' interpretations of written assessment comments. *Med Educ*. 2017;51(4):401-410.
17. Bing-You R, Hayes V, Varaklis K, Trowbridge R, Kemp H, McKelvy D. Feedback for learners in medical education: what is known? A scoping review. *Acad Med*. 2017;92(9):1346-1354.
18. Andreou V, Peters S, Eggermont J, Schoenmakers B. Co-designing entrustable professional activities in general practitioner's training: a participatory research study. *BMC Med Educ*. 2024;24(1):549.
19. Gofton WT, Dudek NL, Wood TJ, Balaa F, Hamstra SJ. The Ottawa surgical competency operating room evaluation (O-SCORE): a tool to assess surgical competence. *Acad Med*. 2012;87(10):1401-1407.
20. Taherdoost H. Sampling methods in research methodology; how to choose a sampling technique for research. How to choose a sampling technique for research (April 10, 2016). 2016.
21. Krippendorff K. *Content Analysis: An Introduction to its Methodology*. Sage Publications; 2018.
22. Savin-Baden MMCH. *Qualitative Research: The Essential Guide to Theory and Practice*. Routledge; 2013.
23. 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;32(3):319-329.
24. Woods R, Singh S, Thoma B, et al. Validity evidence for the quality of assessment for learning score: a quality metric for supervisor comments in competency based medical education. *Can Med Educ J*. 2022;13(6):19-35.
25. Carter N, Bryant-Lukosius D, DiCenso A, Blythe J, Neville AJ. The use of triangulation in qualitative research. *Oncol Nurs Forum*. 2014;41(5):545-547.
26. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163.
27. Cooper HE, Camic PM, Long DL, Panter A, Rindskopf DE, Sher KJ. *APA Handbook of Research Methods in Psychology, Vol 3: Data Analysis and Research Publication*. American Psychological Association; 2012.
28. Cohen L, Manion L, Morrison K. *Research Methods in Education*. Routledge; 2002.
29. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Br Med J*. 2007;335(7624):806-808.
30. Bonnie LHA, Nasori M, Visser MRM, Kramer AWM, van Dijk N. Feasibility, and validity aspects of entrustable professional activity (EPA)-based assessment in general practice training. *Educ Prim Care*. 2022;33(2):69-76.
31. Shorey S, Lau TC, Lau ST, Ang E. Entrustable professional activities in health care education: a scoping review. *Med Educ*. 2019;53(8):766-777.
32. Bonnie LHA, Visser MRM, Bont J, Kramer AWM, van Dijk N. Trainers' and trainees' expectations of entrustable professional activities (EPAs) in a primary care training programme. *Educ Prim Care*. 2019;30(1):13-21.
33. Lip A, Watling CJ, Ginsburg S. What does "timely" mean to residents? Challenging feedback assumptions in postgraduate education. *Perspect Med Educ*. 2023;12(1):218-227.
34. Watling CJ, Ginsburg S. Assessment, feedback and the alchemy of learning. *Med Educ*. 2019;53(1):76-85.
35. Schut S, Driessen E, van Tartwijk J, van der Vleuten C, Heeneman S. Stakes in the eye of the beholder: an international study of learners' perceptions within programmatic assessment. *Med Educ*. 2018;52(6):654-663.
36. Martin L, Sibbald M, Brandt Vegas D, Russell D, Govaerts M. The impact of entrustment assessments on feedback and learning: trainee perspectives. *Med Educ*. 2020;54(4):328-336.
37. Eeckhout T, Gerits M, Bouquillon D, Schoenmakers B. Video training with peer feedback in real-time consultation: acceptability and feasibility in a general-practice setting. *Postgrad Med J*. 2016;92(1090):431-435.
38. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: the do's, don'ts and don't knows of direct observation of clinical skills in medical education. *Perspect Med Educ*. 2017;6(5):286-305.
39. Hauer KE, Holmboe ES, Kogan JR. Twelve tips for implementing tools for direct observation of medical trainees' clinical skills during patient encounters. *Med Teach*. 2011;33(1):27-33.
40. Choo EK, Woods R, Walker ME, O'Brien JM, Chan TM. The quality of assessment for learning score for evaluating written feedback in anesthesiology postgraduate medical education: a generalizability and decision study. *Can Med Educ J*. 2023;14(6):78-85.
41. Woods R, Singh S, Thoma B, et al. Validity evidence for the quality of assessment for learning score: a quality metric for supervisor comments in competency based medical education. *Can Med Educ J*. 2022;13(6):19-35.