European Journal of Education and Pedagogy
Vol 6 | Issue 5 | September 2025
ISSN 2736-4534

RESEARCH ARTICLE

A Comparative Analysis of Objective Structured Clinical Examination (OSCE) Observed Scores and Global Rating Scores using a Novel Approach

Akram Alsahafi[1,2,*], John Newell[3], Micheál Newell[1], and Thomas Kropmans[1]

ABSTRACT

Background: **Objective Structured Clinical Examinations (OSCEs) are a cornerstone of medical education. Despite their widespread use, the relationship between observed scores and global rating scores in OSCEs remains a topic of debate. This study aimed to identify potential scoring discrepancies between the observed scores and global rating scores of OSCEs.**

Method: **This retrospective observational study analyzed anonymized OSCE data from 1,571 undergraduate medical students in the 4 MB program at a single medical school over nine years. The data collected from randomly selected OSCE sessions included observed scores calculated as percentages for individual stations and global rating scores (GRS) assigned holistically at the station level. A key change made midway through the study refined the GRS, splitting the "Borderline" category into "Borderline Pass" and "Borderline Fail." The Data were analyzed using raincloud plots, ordinal regression modelling, and tree-based approaches to identify and visualize discrepancies between the two assessment measures.**

Results: **The analysis identified discrepancies between observed scores and global rating scores, revealing that a single observed score often corresponded to multiple global rating categories. Ordinal regression and tree-based models highlighted substantial variability, particularly within mid-range categories (GRS bands 2, 3, and 4), making evaluations of these ranges more subjective and uncertain. The conditional inference tree further illustrated that the mid-range observed scores lacked clear alignment with specific global rating categories, underscoring the inconsistency and variability in examiner assessments.**

Discussion: **This study emphasizes the need for consistent and comprehensive assessment tools. The findings align with previous research, highlighting the challenges in aligning the observed scores and global rating scores in OSCEs. The identified discrepancies emphasize the necessity of adopting a feedback system that integrates both qualitative and quantitative aspects.**

Conclusion: **This research highlights the importance of structured feedback in bridging the gaps between the two scoring methodologies and in enhancing student learning, professional development, and faculty advancement.**

Keywords: Assessments, global rating scores (GRS), medical education, objective structured clinical examinations (OSCEs).

Submitted: February 07, 2025

Published: September 01, 2025

10.24018/ejedu.2025.6.5.939

[1] *University of Galway, Galway; College of Medicine, Nursing and Health Sciences–School of Medicine, Ireland.*
[2] *Taif University, Department of Medical Education, College of Medicine, Saudi Arabia.*
[3] *University of Galway, Galway; School of Mathematical and Statistical Sciences, Ireland.*

* ***Corresponding Author:***
e-mail: a.alsahafi1@unviversityofgalway.ie

1. Introduction

Objective Structured Clinical Examinations (OSCEs) are used extensively as an essential component of medical education to evaluate students' clinical competence and practical skills in a standardized and objective manner (Harden *et al.*, 1975). Typically, observed scores are based on a student's capacity to perform particular clinical duties and adhere to standardized procedures (Patrício *et al.*, 2013). Moreover, global rating scores (GRS) are necessary assessment tools for evaluating objective structured clinical examinations (OSCEs)

(Hodges & McIlroy, 2003). They provide a holistic view of a student's performance beyond detailed evaluations of objective checklists (observed scores) (Hodges *et al*., 1999). The GRS effectively captures the examiner's overall professional impression, blending technical and clinical skills with essential competencies such as communication and professional behaviour, which are vital for a healthcare professional's comprehensive development.

Integrated as a single item at the end of the OSCE score, the GRS transitions from specific task evaluations to a broader judgment, assigning students to the following categories: fail, borderline, pass, good, or excellent (Malau-Aduli *et al*., 2012). This approach highlights the significance of the GRS in encapsulating a student's performance, offering a summative view that mirrors the examiner's overall professional assessment beyond the specifics captured in checklists (Hodges *et al*., 1999). Despite their pervasive application, the assessment methods used in OSCEs, particularly the relationship between observed and global rating scores, have been the subject of ongoing debate and investigation (Khan *et al*., 2013).

This study aimed to identify potential scoring discrepancies between observed and global rating scores in Objective Structured Clinical Examinations (OSCEs). This research provides valuable insights into the reliability and validity of OSCE scoring systems and enhances the assessment of future healthcare professionals.

2. Methodology

2.1. Study Design, Participants and Setting

This retrospective observational study analyzed anonymized OSCE data from 1,571 undergraduate medical students in a 4 MB program at a single medical school over nine cohort years. Focusing on the same year group ensured consistency in the comparisons over time. The OSCEs evaluated competencies such as communication, procedural skills, and clinical reasoning. Ethical approval for this study was obtained from the ethics committee of the university.

2.2. OSCE Structure

Each OSCE comprises up to 15 stations and is designed to assess key clinical competencies such as communication, procedural skills, and clinical reasoning. While the format and purpose of stations have remained broadly consistent over the years, minor variations in case presentations and checklist criteria may have occurred to align with curriculum updates.

2.3. Scoring Process and Feedback Provision

Observed scores were calculated as percentages for each station, allowing for a granular analysis of task-level performance. The total scores across the stations were not aggregated, thus ensuring a station-specific focus. The examiners independently recorded these scores on standardized checklists without access to cumulative totals.

In addition to the checklist-based scores, examiners assigned a Global Rating Score (GRS) to each station based on their professional judgment and experience. Initially, the GRS included five levels: fail, borderline, pass, good, and excellent. In 2014, the "borderline" category was further refined into "Borderline Pass" and "Borderline Fail".

Feedback during the nine-year study period was optional and inconsistently provided. Although some examiners offered written comments, many left the feedback section blank. Moreover, the feedback provided often lacked specificity and actionable insights.

2.4. Data Collection

Data were obtained from the official records of the medical school, encompassing comprehensive information on students' observed and global rating scores for each OSCE station. The observed scores were derived from objective checklists that evaluated students' performance on specific clinical tasks. Global rating scores were assigned based on the examiners' holistic evaluation of the students' overall clinical competence, communication skills, and professional conduct.

2.5. Statistical Analysis

This study utilized various statistical techniques to summarize and interpret the relationship between the observed scores and global rating scores. A raincloud plot was used to visualize the differences between the scoring systems and within each global score category. Furthermore, plots of the predicted Global Rating Score as an ordinal response, using Observed Scores as an explanatory variable were created to highlight the uncertainty associated with predicting global rating scores from observed scores at the student level.

Tree-based approaches, using conditional inference, are an informative tools for visualizing the relationship between a response and an explanatory variable by visualizing the binary splits recursively of the explanatory variable that best predicts the response. Two approaches were used to better understand the relationship between the two response variables. In the first tree the Global Rating Score was used as the response with the Observed Score as the predictor whereas in the second tree, the roles of each variable were reversed.

This exploration elucidates the variability intrinsic to each global rating category, demarcating a spectrum of plausible values for global rating scores based on observed scores, thereby enriching our understanding of the interplay between these two scoring metrics.

3. Results

This study identifies the detailed relationship between students' observed scores and their corresponding global rating scores in the Objective Structured Clinical Examination (OSCE) assessment. The analysis was based on performance data from 1571 undergraduate medical students spanning nine cohorts, all from a single medical school. Given the pivotal role of OSCEs in measuring clinical competencies, understanding this association is of paramount importance. Rigorous exploration focused on examiners' assessments from various randomly selected OSCE sessions, ensuring a comprehensive and meticulous examination of the data.

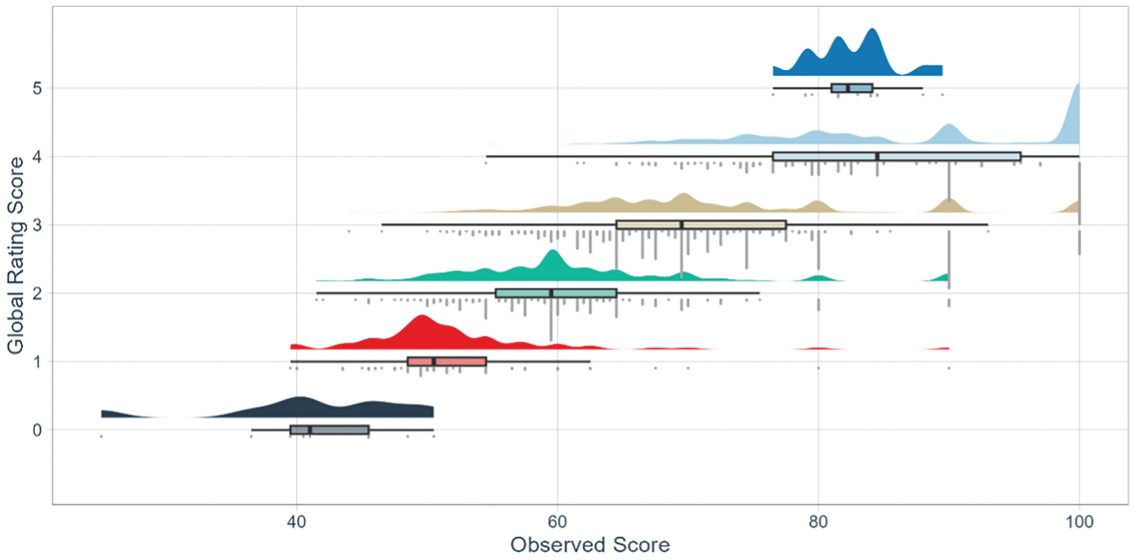

Fig. 1. Raincloud plot of global rating score and observed score.

TABLE I: Expanded Global Rating Score Categories for OSCE Performance Evaluation [2014–2019]

Global rating scores	Categories
Fail	0
Borderline fail	1
Borderline pass	2
Pass	3
Good	4
Excellent	5

The visual and analytical approaches taken provided insight into the alignment of students' overall performance (as indicated by observed scores) with expert reviews expressed in the global rating scores.

The raincloud plot in Fig. 1 visualizes the relationship between the observed scores from Objective Structured Clinical Examinations (OSCEs) and the corresponding global rating scores. Global rating scores were categorized into six distinct levels (Table I): fail (category 0), borderline Fail (category 1), borderline pass (category 2), pass (category 3), good (category 4) and excellent (category 5). The raincloud plot in Fig. 1 visualizes the relationship between the observed scores from Objective Structured Clinical Examinations (OSCEs) and the corresponding global rating scores. Each raincloud consists of a density plot (upper part), a box plot (middle part), and individual data points (lower part), representing the distribution of observed scores across different global rating score categories.

The raincloud plot in Fig. 1 illustrates that the observed scores for global rating categories 0 (Fail) and 5 (Excellent) are distinct and clear, aiding examiners' decisions. For instance, GRS 0 primarily falls between observed scores of 30 and 45, while GRS 5 predominantly ranges from 80 to 100. However, categories 2 (Borderline Pass), 3 (pass), and 4 [good] showed a significant overlap, with observed scores generally between 50 and 80. This overlap suggests that there is a discrepancy between the observed scores and global rating scores, making evaluations in these ranges more uncertain and subjective.

We observed distinct differences in the mean observed scores across various global rating score categories over different cohort periods. For the cohorts from 2010 to 2013, students categorized as "Clear Pass" (GRS 2) had a mean observed score of 67.4 with a standard deviation of 11.4. In contrast, those rated as "Excellent" (GRS 4) achieved a mean score of 91.6 with a standard deviation of 8.39 (Table II). Similarly, in the 2014 to 2019 cohorts, the mean observed scores for "Borderline Pass" (GRS 2) and "Good" (GRS 4) were 58.7 and 76.6, respectively, highlighting a significant progression in the observed scores for higher global ratings (Table III). When combining the data from 2010 to 2019, the trends remained consistent, with the "Fail" category (GRS 0) having a mean observed score of 41.4, while the "Excellent" category (GRS 5) maintained a mean score of 82.6 (Table IV). These observations highlight the variability and progression of observed scores across different global rating score categories and time periods, further illustrating inherent discrepancies and overlaps, particularly in mid-range global rating scores.

Fig. 2 has three charts of boxplots, The boxplot 2010–13 (a) in Fig. 2 illustrates an overall upward trend in global rating scores as observed scores increase, in addition to considerable overlap in scores across categories. The middle chart (b) in the same figure follows this trend for the cohort of 2014–19. The last chart (c) 2010–19 in Fig. 2 aggregates the data, presenting a cumulative view that reinforces the observed patterns of the individual cohorts.

To formally explore the relationship between the observed and global rating scores, an ordinal regression model Fig. 3 was fitted and the estimated probabilities of being in each global rating category as a function of a student's observed score were calculated and visualized. This is quite clear from the predicted probabilities and the uncertainty in each estimate (using 95% Confidence Intervals) of the overlap in the global rating categories awarded across the range of observed scores, particularly for categories 2,3 and 4 (Fig. 3).

For instance, if a student had an observed score of 65, it would look vertically at 65 on the X-axis across all panels.

TABLE II: Summary Statistics of Observed Scores by Global Rating Scores for Cohorts (2010–2013)

Global rating scores	Mean observed score	Standard deviation	Sample size (n)
0 = Fail	–	–	0
1 = Borderline	56.6	14.1	16
2 = Clear pass	67.4	11.4	89
3 = Good pass	81.9	11.1	242
4 = Excellent	91.6	| 8.39	190

TABLE III: Summary Statistics of Observed Scores by Global Rating Scores for Cohorts (2014–2019)

Global rating scores	Mean observed score	Standard deviation	Sample size (n)
0 = Fail	41.4	7.62	9
1 = Borderline fail	50.8	4.41	61
2 = Borderline pass	58.7	6.13	283
3 = Pass	66.50	6.70	522
4 = Good	76.6	7.27	143
5 = Excellent	82.6	3.34	16

TABLE IV: Summary Statistics of Observed Scores by Global Rating Scores for Combined Cohorts

Global rating scores	Mean observed score	Standard deviation	Sample size (n)
0	41.4	7.62	9
1	52.0	7.75	77
2	60.8	8.54	372
3	71.4	11.0	764
4	85.2	10.8	333
5	82.6	3.34	16

The borderline pass (GRS2) panel shows a predicted probability of 0.4 with a confidence interval from 0.3 to 0.5. In the GRS 3 (pass) panel, the predicted probability could be 0.5 with a confidence interval from 0.4 to 0.6. In the GRS 4 [Good] panel, the predicted probability might be 0.3 with a confidence interval from 0.2 to 0.4. This example in the plot indicates that a student with an observed score of 65 has overlapping probabilities of being rated as Borderline Pass, Pass, or Good, with the highest probability of being rated as pass (Fig. 3).

This plot provides a nuanced view of how observed scores translate into global rating scores, incorporating uncertainty. It shows the likelihood of each global rating category as a function of observed scores, with shaded areas representing the 95% confidence intervals, which indicate the level of uncertainty in the estimates.

Following this a tree-based approach, conditional inference (Hothorn *et al*., 2006; Zeileis *et al*., 2008) was used with i) the Global Rating Score as the Response and Observed Score as the explanatory variable and ii) the Observed Score as the response and Global Rating Score as the explanatory variable. The depth of each tree was determined by the significance of the splitting rule, expressed at each split as a p-value, using the Bonferroni approach to adjust for multiple testing.

Once again, it is clear that the regression tree in Fig. 4 shows a significant overlap in the boxplots between observed score nodes 6, 7, and 8 for the global rating score categories 2 (borderline pass), 3 (pass), and 4 [good], respectively. These overlaps indicate variability and subjectivity in mid-range evaluations. In contrast, clear distinctions are observed in nodes 5 and 9, corresponding to Fail and Excellent categories, which aid examiners' decisions. Significant splits at each global rating level help differentiate between lower and higher performance levels, illustrating the decision logic applied by examiners during the rating process.

The classification tree in Fig. 5 illustrates how observed score thresholds are associated with different global rating score categories. Significant splits at observed scores of 73 and 90 differentiated the lower and higher GRS categories, respectively. Clear distinctions are observed in nodes 4 and 15, which represent the Fail and Excellent categories, aiding examiners' decisions. In contrast, overlaps in mid-range nodes (e.g., 5, 8, 9, 11, 12, and 14) highlight the variability and subjectivity in examiner evaluations for the borderline fail, borderline pass, and pass categories.

4. Discussion

The study's retrospective data analysis revealed insights consistent with prior research (Allen *et al*., 1998; Pell *et al*., 2015; Read *et al*., 2015), examining the discrepancy between observed scores and global rating scores in the OSCE assessment. This study highlights the multifaceted relationship between observed scores and corresponding global rating scores, which is further complicated by the potential for a single observed score to map onto multiple global rating scores.

In their seminal 2015 study published in Medical Teachers, Pell *et al*. explored the complexities of OSCE assessment. Their detailed analysis uncovered discrepancies between assessors' checklist-based (observed scores) evaluations and their overarching 'predictions' or global rating scores (Pell *et al*., 2015). This study spanned a single

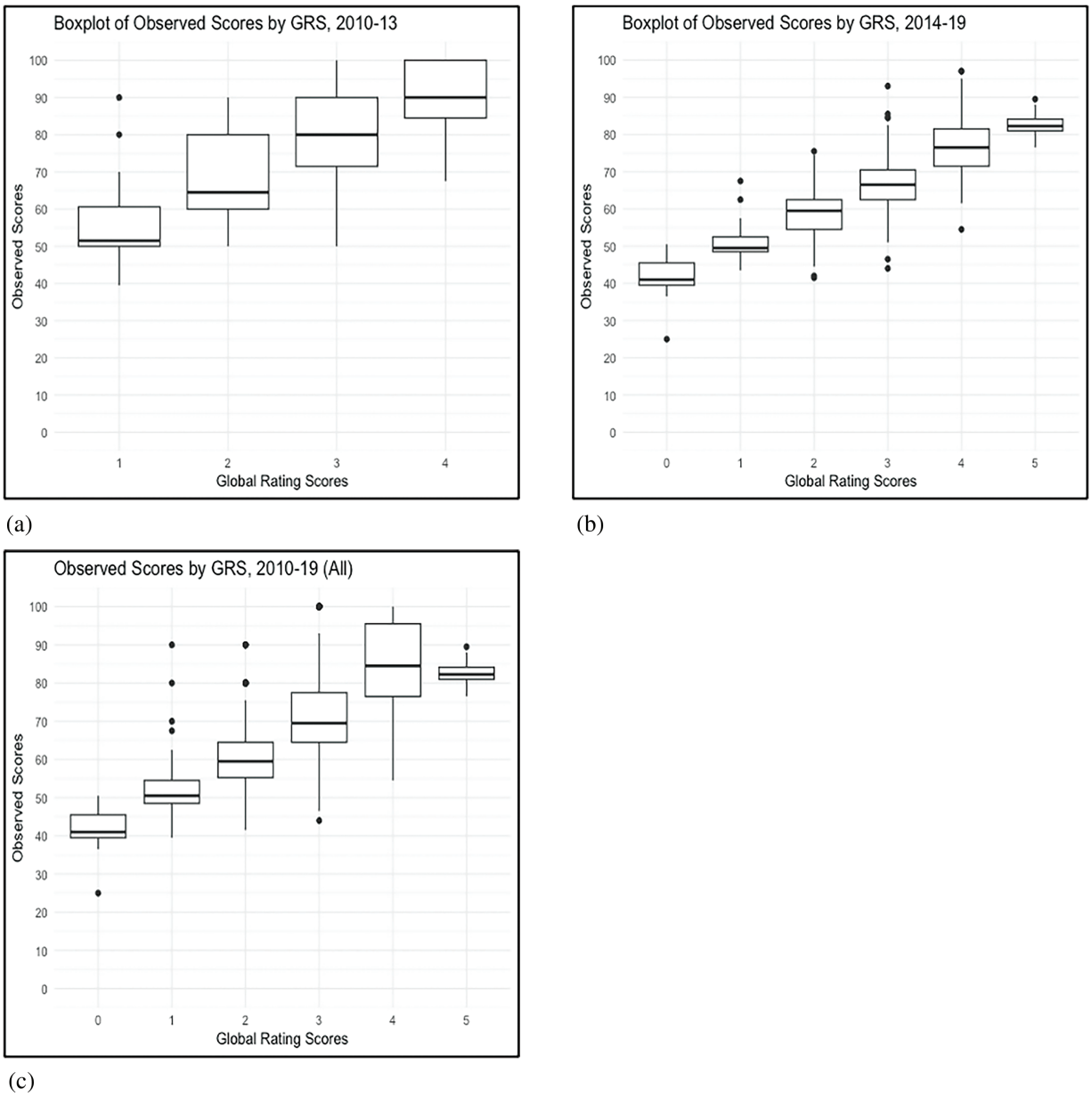

Fig. 2. Boxplots of global rating scores across observed score levels in a single academic year across nine cohorts (2010–2019).

academic year across nine cohorts, offering a comprehensive view of assessment patterns in large-scale OSCEs. Such discrepancies in grading, as their work and this study suggest, are not merely statistical irregularities. They mirror a prevalent challenge in OSCE assessments across boards (Tavakol & Pinner, 2018).

While this study focused on a single institution, its nine-year duration provides an opportunity to evaluate trends and patterns in OSCE assessments. Importantly, the data analysis was conducted at the station level, with both observed and global rating scores calculated separately for each station rather than aggregated across the exam. This station-specific approach allowed for a more granular exploration of the relationship between observed scores and global ratings, thus addressing task-specific performance variability. Variability in scoring may also stem from differences in station design, case complexity, and examiner interpretation, all of which could influence the relationship between the scores.

Changes in the 2014 Global Rating Scale further shaped the scoring trends. The shift from a single "Borderline" category to "Borderline Pass" and "Borderline Fail" was implemented to improve the granularity of mid-range assessments. However, this change introduces new challenges including increased examiner judgment and calibration variability. Future studies could investigate examiners' perceptions and training regarding the use of these categories to better understand their impact on scoring consistency.

The educational ramifications of this discrepancy are profound (Downing, 2005). When there is a disconnect between observed and global rating scores, it raises questions about the validity of the assessments. This can directly impact the quality of the feedback students receive, the instructional methods employed, and the overall efficacy of the OSCE as a tool for measuring clinical competencies. Addressing these discrepancies is not just a matter of refining assessment metrics but is central to

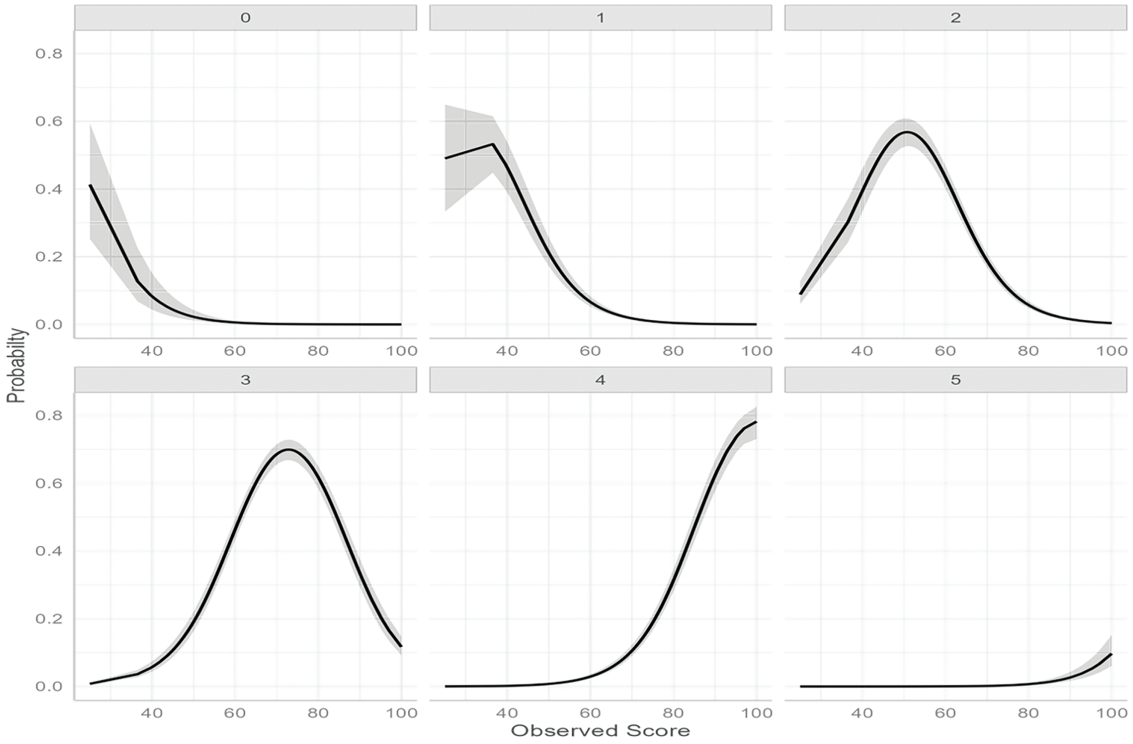

Fig. 3. Plot of the predicted probability of each global rating score by observed score with 95% confidence bands.

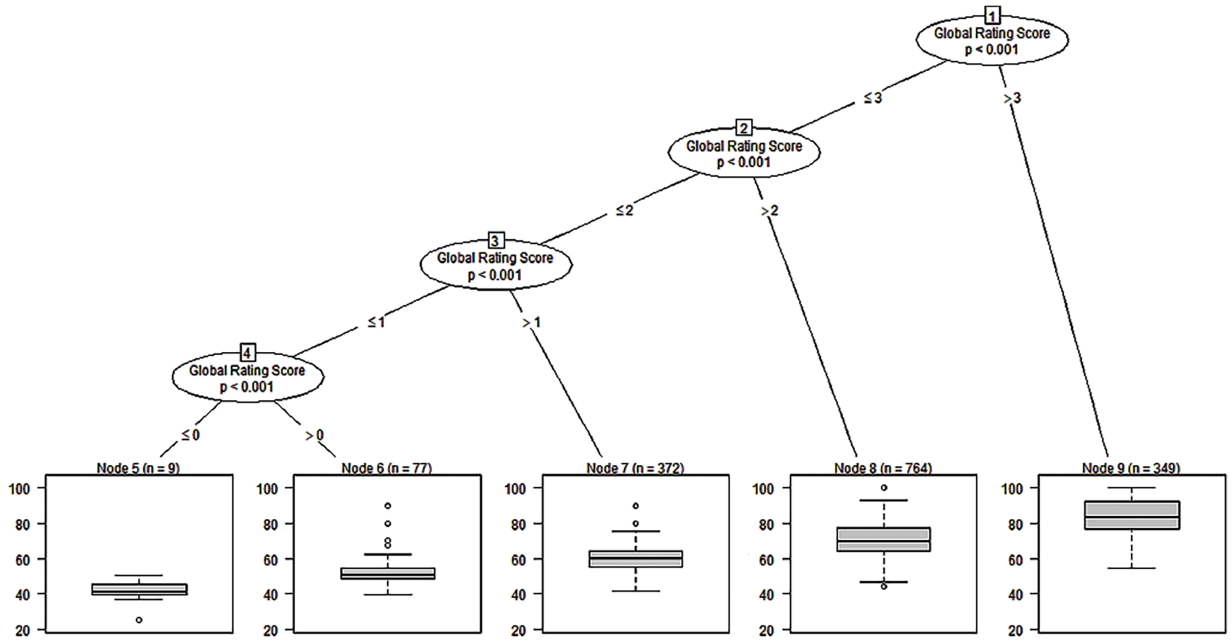

Fig. 4. Plot of the regression tree for observed score.

ensuring students' holistic and meaningful educational experiences.

Feedback, particularly in its qualitative form, is an indispensable pillar of the educational process (Jurs & Špehte, 2020; Schartel, 2012). Given the discrepancies observed between global rating scores and observed scores, written feedback has emerged as a vital instrument for bridging this gap (Ngim *et al.*, 2021; Sterz *et al.*, 2021). Currently, the feedback provided during the study period was optional and inconsistent. While some examiners offered written comments, many left the feedback section blank; when feedback was given, it was often nonspecific and lacked actionable insights. Integrating structured written feedback tailored to specific GRS elements (e.g., clinical competence, communication, and professionalism) could significantly enhance the educational value of the OSCE.

Moreover, feedback plays a pivotal role in student enhancement (Carless *et al.*, 2011). For educators and examiners, feedback is a reflective medium for promoting assessment methods and teaching strategies (Branch & Paranjape, 2002). The feedback cycle extends and loops back, ensuring that students are not merely quantified entities but are recognized for their comprehensive clinical

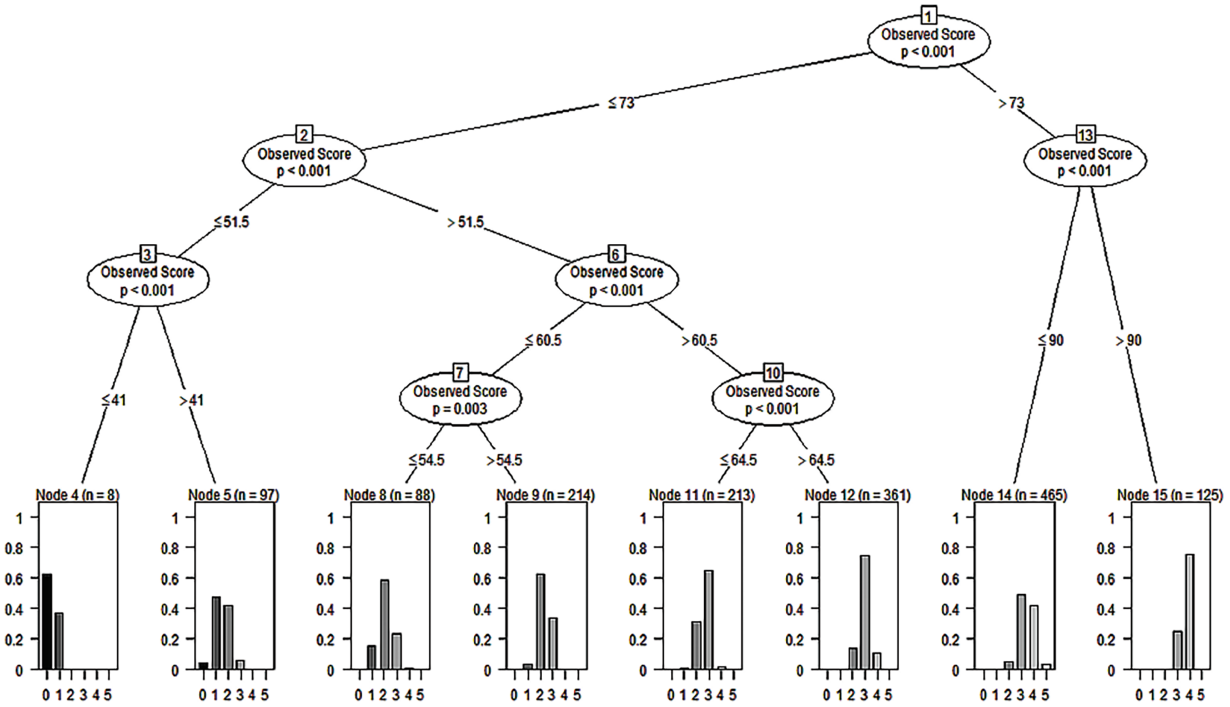

Fig. 5. Plot of the classification tree for global rating score.

competencies. At the same time, it fosters an environment for faculty development, pushing towards a more harmonized, consistent, and accurate student assessment (Sargeant *et al*., 2009). In the overarching framework of education, particularly in the context of our study, prioritizing feedback means championing a dynamic, continuous learning ethos that benefits both students and educators.

The results of this study pave the way for several promising directions in future research. Since the study was limited to one institution, further research incorporating data from multiple institutions would provide a broader and more diverse perspective on the relationship between the observed and global rating scores. Further investigations should also consider examiner-specific factors such as examiner training, experience, and individual scoring tendencies, as they can significantly influence scoring patterns. Additionally, qualitative methods such as examiner interviews or focus group discussions could provide nuanced insights into the sources of variability and the nature of discrepancies observed (Watling & Lingard, 2012).

This study has several strengths, including a large sample size, the use of comprehensive data, and the application of rigorous statistical methods, all of which lend credence to the robustness and reliability of the findings. Furthermore, it addresses a crucial aspect of medical education that has received limited attention in the existing literature, thereby contributing to a better understanding of the dynamics of medical student assessments (Norcini & McKinley, 2007).

However, this study had several limitations. The single-institution design may limit the generalizability of the findings, and the retrospective nature of the analysis means that potential changes in the OSCE station content, examiner roles, or scoring criteria over time were not systematically tracked. Although efforts were made to maintain consistency, variability in case presentation or checklist design could have influenced the results. Future research should explore these factors to provide a more comprehensive understanding of the complexities of the OSCE assessments.

5. Conclusion

In this study's detailed examination of OSCEs, pronounced ambiguity was observed in the relationship between the observed scores and global rating scores. Notably, a single observed score can correspond to multiple potential global rating scores, emphasizing the inherent uncertainty in directly mapping one score onto the other. While observed (numeric) scores offer invaluable insights into student performance, they may need to reflect a student's clinical proficiency comprehensively. This observed overlap and uncertainty between the two metrics are not exclusive to the current institution, but instead echo a broader, prevailing trend in global medical education assessments.

These findings emphasize the need to develop a better system to deliver refined, specific, and meaningful written feedback. Rather than merely appending qualitative feedback to quantitative scores, it is imperative to seamlessly integrate the two.

Future research could investigate the potential of using a composite score that combines both observed and global rating scores, either by adding or multiplying them, to reduce discrepancies and provide a more holistic evaluation of student performance. This approach may help to reconcile the differences between the two scoring systems and offer a more accurate reflection of clinical competence.

Data Availability

These data are not been publicly available because they are owned by a university. The corresponding author was contacted for information regarding the data from the study.

Funding

This study did not receive a specific grant from any funding agency in the public, commercial, or not-for-profit sector.

Conflict of Interest

The authors declare that no competing interests are associated with this study.

Ethics Approval

This study was conducted in accordance with ethical standards of the University of Galway. This study was approved by the University Ethics Committee on December 2nd, 2020 (Ethical Committee Application Reference Number 2020.12.019).

Generative Artificial Intelligence (AI)

ChatGPT (version 4.0) was used as the generative AI tool to support the proofreading process. The tool efficiently identified and corrected grammatical errors, enhanced stylistic consistency, and improved overall clarity and readability of the manuscript. Its use was limited to refining language and structure without influencing the substantive content or conclusions of the study.

References

Allen, R., Heard, J., & Savidge, M. (1998). Global ratings versus checklist scoring in an osce. *Academic Medicine*, *73*(5), 597–598.

Branch, W. T. Jr, & Paranjape, A. (2002). Feedback and reflection: Teaching methods for clinical settings. *Academic Medicine*, *77*(12), 1185–1188.

Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education*, *36*(4), 395–407.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*, 133–143. https://doi.org/10.1007/s10459-004-4019-5.

Harden, R. M., Downie, W. W., Stevenson, M., & Wilson, G. M. (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal*, *1*(5955), 447–451. https://doi.org/10.1136/bmj.1.5955.447.

Hodges, B., & McIlroy, J. H. (2003). Analytic global osce ratings are sensitive to level of training. *Medical Education*, *37*(11), 1012–1016. https://asmepublications.onlinelibrary.wiley.com/doi/pdfdirect/10.1046/j.1365-2923.2003.01674.x?download=true.

Hodges, B., Regehr, G., McNaughton, N., Tiberius, R., & Hanson, M. (1999). Osce checklists do not capture increasing levels of expertise. *Academic Medicine*, *74*(10), 1129–1134. https://doi.org/10.1097/00001888-199910000-00017.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Jurs, P., & Špehte, E. (2020). The value and topicality of feedback in improving the learning process. *Society. Integration. Education. Proceedings of the International Scientific Conference*.

Khan, K. Z., Gaunt, K., Ramachandran, S., & Pushkar, P. (2013). The objective structured clinical examination (osce): Amee guide no. 81. Part ii: Organisation & administration. *Medical Teacher*, *35*(9), e1447–e1463. https://doi.org/10.3109/0142159X.2013.818635.

Malau-Aduli, B. S., Mulcahy, S., Warnecke, E., Otahal, P., Teague, P. -A., Turner, R., & Van der Vleuten, C., (2012). Inter-rater reliability: Comparison of checklist and global scoring for osces. *Creative Education*, *3*(6), 937.

Ngim, C. F., Fullerton, P. D., Ratnasingam, V., Arasoo, V. J. T., Dominic, N. A., Niap, C. P. S., & Thurairajasingam, S. (2021). Feedback after osce: A comparison of face to face versus an enhanced written feedback. *BMC Medical Education*, *21*, 1–9.

Norcini, J. J., & McKinley, D. W. (2007). Assessment methods in medical education. *Teaching and Teacher Education*, *23*(3), 239–250. https://doi.org/10.1016/j.tate.2006.12.021.

Patrício, M. F., Julião, M., Fareleira, F., & Carneiro, A. V. (2013). Is the osce a feasible tool to assess competencies in undergraduate medical education? *Medical Teacher*, *35*(6), 503–514. https://doi.org/10.3109/0142159X.2013.774330.

Pell, G., Homer, M., & Fuller, R. (2015). Investigating disparity between global grades and checklist scores in osces. *Medical Teacher*, *37*(12), 1106–1113. https://doi.org/10.3109/0142159X.2015.1009425.

Read, E. K., Bell, C., Rhind, S., & Hecker, K. G. (2015). The use of global rating scales for osces in veterinary medicine. *PLoS One*, *10*(3), e0121000. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0121000&type=printable.

Sargeant, J. M., Mann, K. V., Van der Vleuten, C. P., & Metsemakers, J. F. (2009). Reflection: A link between receiving and using assessment feedback. *Advances in Health Sciences Education*, *14*, 399–410. https://link.springer.com/content/pdf/10.1007/s10459-008-9124-4.pdf.

Schartel, S. A. (2012). Giving feedback-an integral part of education. *Best Practice & Research Clinical Anaesthesiology*, *26*(1), 77–87.

Sterz, J., Linßen, S., Stefanescu, M., Schreckenbach, T., Seifert, L., & Ruesseler, M. (2021). Implementation of written structured feedback into a surgical osce. *BMC Medical Education*, *21*(1), 1–9.

Tavakol, M., & Pinner, G. (2018). Enhancing objective structured clinical examinations through visualisation of checklist scores and global rating scale. *International Journal of Medical Education*, *9*, 132. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5951780/pdf/ijme-9-132.pdf.

Watling, C. J., & Lingard, L. (2012). Grounded theory in medical education research: Amee guide no. 70. *Medical Teacher*, *34*(10), 850–861. https://doi.org/10.3109/0142159X.2012.704439.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.