BMC Medical Education

https://doi.org/10.1186/s12909-026-08809-4

Article in Press

Agreement and reliability of global rating versus checklist scores in a high-stakes undergraduate OSCE in Rwanda

Received: 16 January 2026

Accepted: 10 February 2026

Published online: 14 February 2026

Cite this article as: Ibrahim O.R., McCall N., Bekele A. *et al.* Agreement and reliability of global rating versus checklist scores in a high-stakes undergraduate OSCE in Rwanda. *BMC Med Educ* (2026). https://doi.org/10.1186/s12909-026-08809-4

Olayinka Rasheed Ibrahim, Natalie McCall, Abebe Bekele, Biniam Ewnte Zelelew, Oluwaseun Ojomo, Anteneh Gadisa Belachew, Equlinet Misganaw Amare, Zelalem Mengistu Gashaw, Birhanu Abera Ayana & Ariane Nina Ndayikeje

Title: Agreement and Reliability of Global Rating versus Checklist Scores in a High-Stakes Undergraduate OSCE in Rwanda

Authors

Olayinka Rasheed Ibrahim[1*], Natalie McCall[1], Abebe Bekele[2], Biniam Ewnte Zelelew[2], Oluwaseun Ojomo[3], Anteneh Gadisa Belachew[2], Equlinet Misganaw[4], Zelalem Mengistu Gashaw[5], Birhanu Abera Ayana[5], Ariane Nina Ndayikeje[1]

1. Department of Pediatrics, Division of Clinical Medicine, School of Medicine, University of Global Health Equity, Kigali, Rwanda
2. Department of Surgery, Division of Clinical Medicine, School of Medicine, University of Global Health Equity, Kigali, Rwanda
3. School of Medicine, University of Global Health Equity, Kigali, Rwanda
4. Education Development and Quality Centre, Division of Academic and Research Affairs, University of Global Health Equity, Kigali, Rwanda.
5. Department of Obstetrics and Gynecology, Division of Clinical Medicine, School of Medicine, University of Global Health Equity, Kigali, Rwanda

*Corresponding Author

Olayinka Rasheed Ibrahim

Email address: ibroplus@gmail.com; iolayinka@ughe.org

Abstract

Background

Despite the objective structured clinical examinations (OSCE) being widely used in the assessment of clinical competency, the optimal scoring systems remains debatable with limited data from sub-Saharan Africa. This study compared the performance, reliability, and agreement between global rating scales (GRS) and checklist scores at a comprehensive exit examination for undergraduate medical students in Rwanda.

Methods

This cross-sectional descriptive study was conducted during the final 'Exit Exams' of undergraduate medical students at the University of Global Health Equity, Rwanda. The OSCE included 15 stations spread across major clinical specialties and subspecialties. Each station had a checklist with a total score of 20 marks and a three-level GRS (failed, borderline, or passed)

Results

A total of 36 students took part in the OSCE examinations. The mean (standard deviation) checklist score was 84.3 (5.3) %, which was lower than the mean GRS score, 94.3 (6.9) %, $p < 0.001$. All students achieved overall scores above the standard-setting pass mark of 64.4% [set from modified Angoff method], on both scoring systems. While no student failed any station

using checklist scores, the GRS identified failures in stations 5 and 6 (one student each), and stations 7 and 13 (two students each). Overall internal consistency (Cronbach's Alpha) across all the stations was 0.760 [ranged from -0.216 (station 15) to 0.746 (station 9)]. Pearson correlation demonstrated a very strong positive correlation between the checklist and GRS (r= 0.924, p<0.001). Bland-Altman plot showed a mean (standard deviation) of difference of 10.01 (2.8) in favor of GRS, with a lower and upper limit of agreement of 4.44 to 15.58 respectively.

Conclusion

Checklist and GRS scores in the OSCEs demonstrated strong positive correlation, but GRS showed a higher discriminatory ability, identifying performance differences that checklists did not capture. Incorporating GRS alongside checklists may enhance the robustness of high-stakes clinical assessment.

Key words: Checklist, Global rating scale, High-stake examinations, Undergraduates, Sub-Saharan, Rwanda

Introduction

The transition of undergraduate medical students into medical doctors often involves a process of certification through a standard final high stake examination often referred to as the 'Exit Examinations.[1] The exit examinations ensure that undergraduate medical students are assessed for the desired outcomes and competencies and ultimately meet the need of the community, where they are expected to perform clinical practice.[2] In many countries and universities, medical graduates may either proceed directly into independent practice or enter internships with varying levels of supervision. Because of this, exit examinations are considered high-stakes assessments and its play a critical role in ensuring that students possess the knowledge, skills, and professional standards required to practice safely. Determining the appropriate pass mark is therefore both crucial and challenging, as it directly affects patient safety and the credibility of training

programs. Accordingly, in many places, exit exams are designed and conducted after the completion of all clinical clerkships and rotations, and undergraduate students passing the examinations or otherwise will determine whether they are licensed to practice medicine or not.

The assessments carried out during exit examinations utilized various methods depending on the domains, competencies, and learning outcomes being assessed.[3] The methods of assessment during exit examinations range from multiple-choice exams to objective structured clinical examinations [OSCE].[3] Objective structured clinical examination, first proposed in the early 1970s by Ronald Harden and colleagues, has become the standard for clinical-related assessment of competencies and learning outcomes.[4] It consists of a series of timed stations and, as the acronym implies, it is structured, objective, and designed to minimize examiner bias. Studies have demonstrated and confirmed its reliability, reproducibility, and fairness in the assessments of students when compared with old traditional methods of clinical assessment.[5, 6] Whereas OSCE has remained a tool for clinical assessment of students, the scoring systems [use of checklists vs. global rating scales] have remained a subject of debate [7]. Traditionally, checklists are often used; however, the checklist systems have been reported to have some limitations. Some of the limitations of the checklists include less discrimination between students, lack of construct validity, the potential for students to master the checklists over time without detailed knowledge, and failure to reflect real-life management of clinical scenarios or to adequately

probe clinical reasoning.[8] Nonetheless, checklists remain simple to use, can be modified to improve discrimination such as the use of three grade levels: well done, fairly done, and not done, and are easier for junior examiners to implement.[9] On the other hand, the Global rating scale [GRS], usually based on the overall impression of the examiners, has been shown to have better construct validity, better reliability compared with checklists and complement students' competency assessment better. Despite the advantages, global rating score pose challenge for young examiners as they depend largely on experience of the examiners, with better reliability among experienced examiners and, may be subjected to some bias and halo effects.[10, 11]

Studies comparing the two scoring systems for OSCE have been conducted mainly in postgraduate settings, high-income countries and limited to subspecialties.[12, 13] In addition, the few studies among undergraduate students were carried out in specialty-based areas involving fewer stations, and without comprehensive assessment of overall medical and sub-disciplines, expected at an exit examination. [14, 15] While OSCE has been adopted in most medical schools in Sub-Saharan Africa for the clinical assessment of students, there is a lack of comprehensive data evaluating the performance of checklists compared to GRS among undergraduate students at professional exit examinations. Hence, this study aimed to compare the performance, reliability, and agreement of checklist and global rating scales in scoring a comprehensive exit OSCE conducted at the 2025 exit

examinations of undergraduate medical students at the University of Global Health Equity [UGHE], Rwanda.

Methods

Study design and settings

This was a cross-sectional descriptive study conducted during the final 'Exit Exams' of undergraduate medical students at the University of Global Health Equity, Rwanda. The university is a non-profit institution that is anchored on an integrated competency-based curriculum for training of its medical students. It is located in the northern part of Rwanda, where its primary affiliated teaching hospital, Butaro Level 2 Teaching Hospital, is also situated.

Sample size

A convenience sample size that comprised all 36 students who took part in the exit examinations was used for this study.

Description of the OSCE

The OSCE comprised 15 stations distributed across the major clinical specialties and subspecialties of medicine, with additional rest stations included to facilitate smooth student flow within each circuit. The examination was developed through a structured blueprinting process aligned with the expected competencies of graduating undergraduate medical students preparing for general practice.

Using a list of common medical conditions, clinical scenarios, and essential clinical skills that graduates are expected to demonstrate upon entry into internship, 15 integrated subject and topic areas were identified [Supplementary Figure 1]. These topic areas were mapped to the desired competencies and learning outcomes using a two-dimensional blueprint, with competencies represented on the Y-axis and subject areas on the X-axis.

The blueprinting and OSCE development process was conducted over a period of three months and involved a multidisciplinary team of faculty representing various clinical departments. The Educational Development and Quality Center was engaged at every stage of the process, providing oversight and guidance throughout blueprint development, station design, and review. In keeping with the university's integrated, competency-based curriculum, OSCE stations were designed to assess integrated clinical skills rather than isolated disciplines. Initial stations drafts were developed within departments, followed by iterative internal departmental review, joint faculty review, and external peer review. Feedback from each stage was discussed and incorporated prior to finalization of the examination.

Standard setting for OSCE

The OSCE standard setting was done using modified Angoff method. Two faculty teams (5-7 members each) evaluated each station: one team for the first seven stations and another team for the remaining eight stations. For each station, faculty projected the likelihood of a borderline, minimally competent student passing each item in the checklist. The total weighted

average determined the station pass mark. Where there was a difference of more than 0.2 between faculty' projected scores for a borderline candidate, a consensus approach was used to resolve the differences. The overall OSCE cut-off was then calculated as the average across all 15 stations, yielding a final pass mark of 64.4%.

Scoring systems—checklist and GRS

For the scoring system for the OSCE, each station had a checklist with a total score of 20 marks and GRS categorized as failed, borderline, or passed. The checklists were structured with components weighted to ensure the overall score was 20 marks for each station, while the GRS was scored based on the overall impression of a student at the end of the encounter.

Examination delivery-training of the examiners, exam circuit, and students' rotation.

A day prior to the examinations, the examiners had training on the overall exams, scoring system, examination delivery platform [SpeedwellÒ], and expectations. The examiners' preparation consisted of a structured six-hour training and calibration program conducted on the same day. The training was divided into a three-hour morning session and a three-hour afternoon session. During the morning session, examiners were oriented to the OSCE blueprint, standard setting principles, and core concepts of assessment, including strategies to minimize bias and ensure fairness and consistency in grading. This session also clarified examiners' roles and expectations during the examination. The afternoon session focused on practical aspects of

examination delivery. This included a demonstration of the examination software, covering examiner login procedures, student selection, checklist completion, and submission of scores. This part lasted 90 minutes and was followed by a 30-minute hands-on session during which all examiners logged into their individual portals and practiced scoring using test cases. The interactive training included a pilot with test students, where the examiners were asked to score both the checklists and GRS as practice. The “test students” were anonymous, system-generated profiles labeled Test 1 to Test 15 and entered into the examination platform solely for examiners’ training purposes. Any technical or procedural issues identified were addressed in real time. The final hour of the training was dedicated to examiners’ calibration. During this session, examiners assigned to corresponding stations reviewed checklist items together and agreed on scoring criteria to promote scoring consistency across circuits. The OSCE stations were designed by faculty members from the relevant departments, many of whom also served as examiners. However, all examiners—regardless of involvement in station development—participated in the same standardized training and calibration process.

The OSCE was conducted using two parallel circuits (A and B), each comprising 15 stations. Each station in each circuit was staffed by a single trained examiner; therefore, a total of 30 examiners participated in the examination (15 per circuit). As illustrated in Supplementary Figures 2 and 3, Circuits A and B ran concurrently, ensuring that all students were assessed

simultaneously throughout the examination period. To accommodate all 36 students, each circuit included three additional rest stations, resulting in a total of 18 stations per circuit. Each station was allocated 12 minutes, followed by a 3-minute transition period to allow students to move between stations and examiners to complete scoring within the examination delivery platform (Speedwell®). The examination was administered using Speedwell®, an online examination management system that enables real-time scoring and grading with integrated audit trails.

During the examination, examiners completed the checklist based on students' performance and subsequently assigned a global rating scale (GRS) judgment of pass, borderline, or fail. To minimize examiner bias, checklist scores were not visible or summable during the examination; final aggregated scores were accessible only to examination administrators via the back-end system after completion of the examination.

Data analysis

The examination data from Speedwell were exported as a comma-separated values (CSV) file and subsequently imported into Statistical Package for the Social Sciences (SPSS) version 30 for analysis. The scores from the checklists [maximum of 20 points] were transformed into percentages prior to the analysis. The 3-point scale was used for the GRS, with corresponding scores of 1, 3, and 5 for failed, borderline, and passed, respectively. The GRS scores were also transformed into percentages prior to the analysis. Both transformed checklists and GRS were normally distributed. Paired t-tests

were used to compare the mean scores between the methods. The Cronbach's Alpha was used to assess the internal consistency of the OSCE stations, while intraclass correlations were used to assess the correlations between the two methods by each examiner. The Pearson correlation coefficient was used to assess the overall relationship between the scores by both methods, while a linear regression was used to express the relationship between GRS and checklists. The agreement between the two methods was further assessed using the Bland-Altman plot. For all levels of statistical significance, the p value was set at <0.05.

Ethical approval

This study was conducted in accordance with the Declaration of Helsinki. The University of Global Health Equity (UGHE) Institutional Review Board (IRB) gave ethical approval for this study (UGHE-IRB2025/453). The UGHE-IRB also waived the informed consent, as the data were anonymously extracted without an identified link to the study participants. The data were also secured in a password-encrypted computer and maintained with absolute confidentiality.

Results

General characteristics of the scoring systems

A total of 36 students took part in the OSCE examinations. The mean score using the checklist was 84.3 (5.3) %, which was lower than the mean score using GRS, 94.3 (6.9) %, $p < 0.001$. Using the standard setting cut-off pass

mark of 64.4%, all students' scores were above it, with a range of 69.8 to 91.8% for the checklist and 73.3 to 100.0% for GRS.

For each station, the scores from GRS were significantly higher than the checklists [Table 1]. Based the scoring scale, the minimum mean score for checklist was obtained in station 6 [73.3%], while the minimum score for GRS was obtained in station 1 [87.8%]. Based on students' performance using checklists, no student failed any station using the 64.4% pass mark set from standard settings. Based on students' performance using GRS, stations 5 and 6 had one student fail each, while Stations 7 and 13 had two students fail each of them.

Table 1: Distribution of the scores of OSCE stations using the checklists and GRS

Station	Checklists - Mean (SD)	GRS Mean (SD)	GRS Passed	GRS Borderline	GRS Failed	P*
1	79.1 (8.5)	87.8 (18.7)	25 (69.4)	11 (30.6)	0 (0.0)	<0.001
2	89.9 (10.2)	97.8 (9.3)	34 (94.4)	2 (5.6)	0 (0.0)	<0.001
3	82.9 (13.3)	94.4 (14.0)	31 (86.1)	5 (13.9)	0 (0.0)	<0.001
4	76.0 (13.6)	86.7 (19.1)	24 (66.7)	12 (33.3)	0 (0.0)	<0.001
5	80.4 (13.5)	92.2 (18.7)	30 (83.3)	5 (13.9)	1 (2.8)	<0.001
6	73.3 (10.7)	88.9 (20.5)	27 (75.0)	8 (22.2)	1 (2.8)	<0.001
7	86.5 (12.2)	92.2 (21.0)	31 (86.1)	3 (8.3)	2 (5.6)	0.007
8	94.6 (5.4)	98.9 (6.7)	35 (97.2)	1 (2.8)	0 (0.0)	<0.001
9	85.4 (13.7)	97.8 (9.3)	34 (94.4)	2 (5.6)	0 (0.0)	<0.001

10	85.1 (7.8)	97.8 (9.3)	34 (94.4)	2 (5.6)	0 (0.0)	<0.001
11	84.0 (13.4)	100.0 (0.0)	36 (100.0)	0 (0.0)	0 (0.0)	<0.001
12	86.3 (10.2)	92.2 (16.1)	29 (80.6)	7 (19.4)	0 (0.0)	0.027
13	85.4 (11.7)	91.1 (21.6)	30 (83.3)	4 (11.1)	2 (5.6)	0.027
14	88.1 (7.7)	96.7 (11.2)	33 (91.7)	3 (8.3)	0 (0.0)	<0.001
15	87.2 (6.4)	100.0 (0.0)	36 (100.0)	0 (0.0)	0 (0.0)	<0.001

*Paired T-test

Relationship between checklist and GRS

The Cronbach's Alpha [internal consistency] for all the stations showed a value of 0.760 with variability among the individual stations that ranged from -0.216 in station 15 to 0.746 in station 3 [Table 2].

The intra-class correlation (ICC) that assessed the correlation within individual examiner scores for the checklist and GRS showed an overall coefficient of 0.943, $p<0.001$. Based on the individual station performance, the ICC coefficient ranged from <-0.1 in station 15 to 0.839 in station 7 [Table 2].

Pearson correlation also showed overall correlations of 0.924 between the checklist and GRS. Also, most the stations show varying degree of statistically significant positive correlation except stations 8 and 10. [Table 2]

Figure 1 shows the scatter plot of both GRS [dependent variable] and checklist [independent variable] shows correlation coefficient of 0.924,

p<0.001 with a linear regression equation of GRS =checklist * 0.71 +17.7 [R square=0.854, p<0.001].

Using the Bland-Altman plot, the agreement between the methods showed a mean difference (standard deviation) of 10.010 (2.8) in favor of GRS with an upper limit of agreement of 15.583 and a lower limit of agreement of 4.437 [Figure 2].

Table 2: Correlation between the checklist and GRS [n=36]

Station	Cronbach's Alpha	Intra-class correlation coefficient	p	Pearson Correlation coefficient	P
1	0.257	0.691	0.0004	0.699	<0.001
2	0.734	0.689	0.044	0.694	<0.001
3	0.746	0.621	0.051	0.609	<0.001
4	0.580	0.581	0.017	0.516	0.001
5	0.525	0.619,	0.026	0.591	<0.001
6	0.559	0.573	0.075	0.705	<0.001
7	0.742	0.839,	<0.001	0.871	<0.001
8	0.461	0.389,	0.071	0.305	0.071
9	0.688	0.413,	0.119	0.433	0.008
10	0.309	0.236	0.226	0.281	0.097
11	0.630	<0.100,	0.500	-*	
12	0.665	0.478,	0.024	0.374	0.025
13	0.472	0.758,	<0.001	0.762	<0.001
14	0.594	0.471,	0.074	0.456	0.005
15	-0.216	<-0.1	0.5	-*	
Total	0.760	0.943	<0.001	0.924,	<0.001

*Not computed as all students had 100% in the GRS.

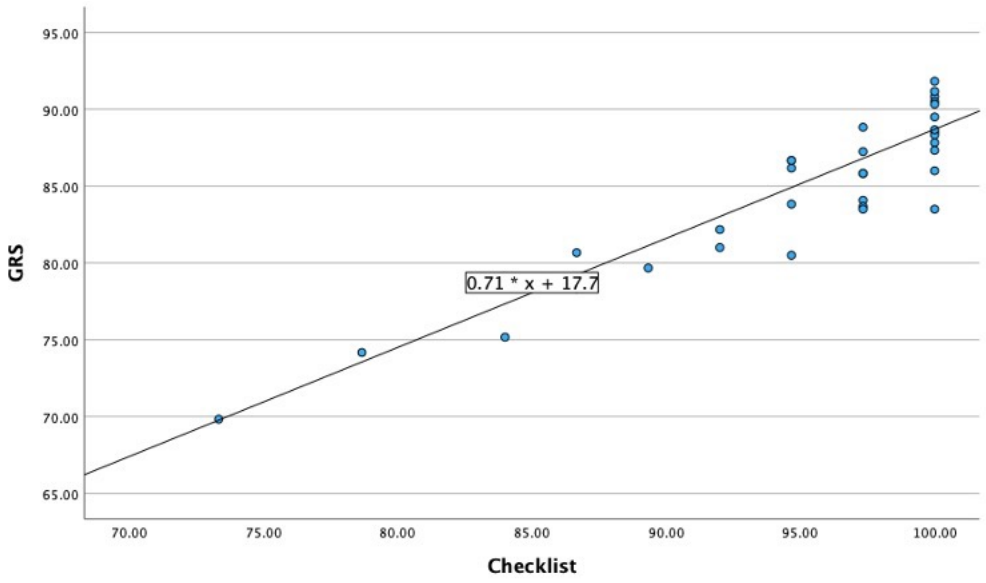

Figure 1: Scatter plot of Global rating scale (GRS) and checklist scores

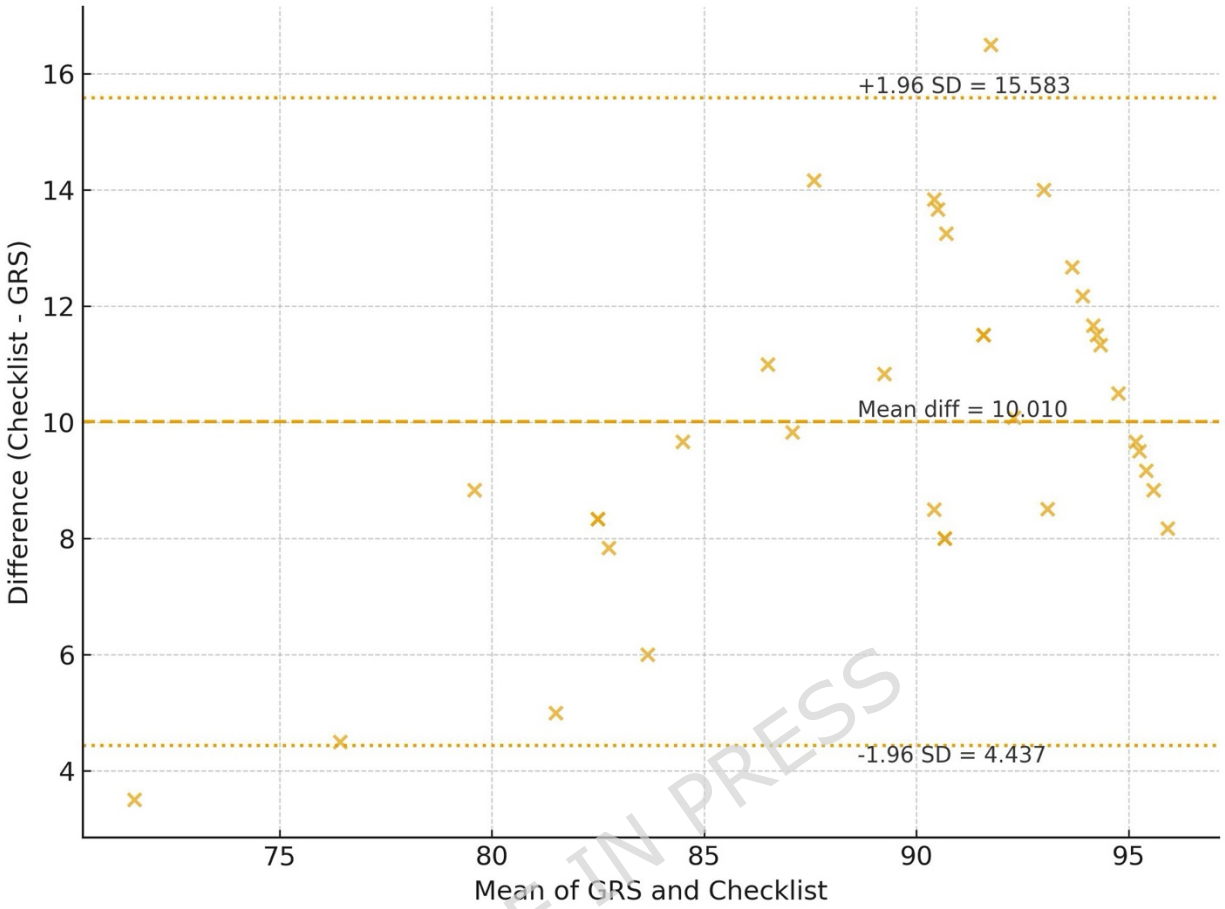

Figure 2: Bland-Altman plot of Global rating scale and checklist scores

Discussion

GRS and checklists are two common methods for scoring OSCEs, with continued debate about the roles of both. The overall mean performance using GRS was higher than the mean performance using the checklist scoring scale. These findings are similar to the observations in Pakistan [16], and Canada [17], where the mean score of GRS was higher than the mean score using the checklists. The high scores obtained from the GRS may be due to well-known attributes of the GRS, including better construct validity and

internal consistency.[8] Besides, it gives overall impression at the end of an encounter with the students rather than scores derived from the subset of actions that constitute a checklist. Thus, this finding further supports the possible better performance of GRS compared with the checklist.

Although we observed that all students scored above the pass mark established during standard setting when using checklists at all stations, the GRS identified a few students who failed one or two stations. Previous studies that used standard settings for the pass mark also identified students that failed some checklists and GRS.[9][14] This finding indicates that checklists, although useful for documenting observable actions, may lack the discriminatory power needed to identify marginal and underperformers. The inability of the checklists to detect students that may not perform to expectation at a station brought forward the issue of construct validity and reinforced that absolute reliability on checklists as a way of scoring students may not provide sufficient discrimination. This observation further supports the roles for continuous use of GRS as an adjunct to checklists to ensure a more accurate reflection of student competency, especially at high-stakes examinations such as exit examinations. [11]

The overall internal consistency as measured by Cronbach's alpha for all the OSCE stations was 0.76, which demonstrated acceptable reliability.[18] This obtained value falls within the range reported in a systematic review of 23 OSCE studies (0.43–0.79), but is higher than the 0.54 observed among final-year students in Canada.[6][17] The overall acceptable internal consistency

of the OSCE stations in this study may be due to the rigorous process of blueprinting with internal and external reviews of the stations, which improved their quality. Though the overall internal consistency is acceptable, there are significant variability across the stations, which is not unexpected given the large number of stations (15); an observation that is consistent with other large-scale OSCE studies.[6] The findings of variability in the internal across the stations emphasizes the significance of post-examination analyses, as stations with poor reliability (stations 2 and 15, with alphas of 0.257 and –0.216, respectively) would require revision or retirement to maintain examination quality.[19]

Overall intraclass correlations that assessed the correlations between the scoring systems by each examiner show very high correlation coefficients between the two methods' scoring scales, which reaffirm their convergent validity. Also, overall, there were very high correlations between the scores for the methods, which is consistent with findings from previous studies. The very high correlation coefficient found in this study is higher than values reported in some literature—0.62 to 0.88 in Malaysia [20] , 0.63 to 0.88 in Sudan [21], and 0.86 in Saudi Arabia.[15] The high correlation between the methods supports their continuous use as valid approaches for scoring student performance during OSCE. However, a few stations (8, 10, 11, and 15) exhibited poor correlations, which may reflect issues with station quality or intra-examiner variability.[19] This suggests the need for post-exam analysis to identify and address potential sources of discrepancy, which could

be either from the examination quality, or examination delivery, including the examiners' inherent limitations and factors. The post-examination analysis of stations with poor correlates will strengthen the reliability of future assessments of the stations and serve as part of quality improvement.

Regression analysis showed that checklist scores significantly predicted GRS scores (R^2 = 85.4%), demonstrating a robust relationship between the two scoring systems. This level of predictive strength in this study is higher than the reported value in core clinical clerkships in Michigan (R^2 = 0.49).[12] The higher values obtained in this study compared with Michigan could be due differences in the study method, as the present study is a high-stake exams compared with Michigan study that involved all regular core clinical clerkships. Additionally, the number of students in this study was 36, while Michigan had 524 students from 2018 to 2021.

The Bland–Altman analysis showed limits of agreement from +4.437 to +15.583, indicating that GRS consistently produced higher scores than the checklist across all students. The fact that the entire interval lies above zero suggests a positive systematic bias in favor of GRS rather than random variation.[22] While the limits of agreement remain relatively narrow, this consistent upward shift underscores that GRS may systematically inflate performance compared with checklist-based scoring, reinforcing its tendency to capture broader impressions of competence

Study Limitations

This study has some limitations. First, the role of the examiners' experience on scoring was not explored. Second, the same examiners assessed students using both scoring methods, which may have introduced potential bias, as checklist-based assessment was completed prior to assignment of the global rating scale (GRS). Third, students did not encounter every examiner, potentially limiting the uniformity of assessment. Finally, this study's single-institution data and may limit it generalizability to the sub-region.

Practice points

Global Rating Scales (GRS) detect performance variations that checklists overlook, identifying station-level failures even when checklist scores suggest students passed.

Checklist and GRS scores show strong correlation, confirming that both methods assess overlapping dimensions of competence while offering unique contributions.

A combined scoring approach improves assessment robustness, leveraging the objectivity of checklists and the expert judgment embedded in GRS to enhance the validity of high-stakes OSCEs.

Quality of individual OSCE stations varies, highlighting the need for continuous examiner training, standardization, and station refinement to maintain reliability.

Conclusion

This study shows a strong correlation between checklist and GRS scoring methods in OSCEs, while also revealing a better performance and

discriminatory capacity of GRS. Given its stronger construct validity and ability to identify underperforming students, GRS should be considered an essential component of high-stakes examinations, complementing rather than replacing checklists. We recommend that high-stakes OSCEs employ a model that combined both scoring system with using checklists to ensure coverage of critical steps and GRS to capture overall clinical competence and reasoning, with the final score being a weighted composite of both.

List of abbreviations

OSCE- Objective structured clinical examination

GRS- Global rating scale

Figure legends

Figure 1: Scatter plot of Global rating scale (GRS) and checklist scores

Figure 2: Bland-Altman plot of Global rating scale and checklist scores

Supplementary Files

Figure 1: Figure 1: OSCE Stations based on the Blueprint for the Exit Examination

Figure 2: Examiner station allocation with student rotation for circuit A

Figure 3: Examiner station allocation with student rotation for circuit C

Declarations

Ethics approval and consent to participate

This study was conducted in accordance with the Declaration of Helsinki. The University of Global Health Equity (UGHE) Institutional Review Board (IRB) gave ethical approval for this study (UGHE-IRB2025/453). The UGHE-IRB also waived the informed consent, as the data were anonymously extracted without an identified link to the study participants. The data were also secured in a password-encrypted computer and maintained with absolute confidentiality

Consent for publication

Not applicable

Availability of data and materials

The data and material used in this study is available from the corresponding author upon reasonable request.

Competing interests

The authors have none to declare

Funding

The author(s) reported there is no funding associated with the work featured in this article.

This was work did not receive any fundings

Authors' contributions

ORI, NM, AB, BE, OO, AGB, and ZMG contributed to the conceptualization, design, data curation and analysis, draft and review of the final version. EM,

BA, and ANN contributed to design, data analysis, draft, and review of the final version. All co-authors approved the final version.

Acknowledgements

We thank the examiners and students who participated in this study

Reference

1. Dehury RK, Samal J. “Exit exams” for medical graduates: a guarantee of quality? Indian J Med Ethics. 2017;2 2 (3 (NS)):190–3. https://doi.org/10.20529/IJME.2017.037.

2. Nath R, Prerna P, Rathi V, Ish P. What’s national exit test for medical students: Merits, challenges, and the way forward. Indian Journal of Medical Specialities. 2023;14(4):252–4. https://doi.org/10.4103/injms.injms_88_23.

3. Al-Wardy NM. Assessment methods in undergraduate medical education. Sultan Qaboos University Medical Journal. 2010;10(2):203–9. https://doi.org/10.18295/2075-0528.1187.

4. Khan KZ, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE Guide no. 81. Part I: An historical and theoretical perspective. Med Teach. 2013;35(9):e1437–46. https://doi.org/10.3109/0142159X.2013.818634.

5. Ibrahim AG. Reliability & validity of the objective structured clinical examination (OSCE): A Meta-analysis (Master’s thesis, University of Calgary, Calgary, Canada). University of Calgary, Calgary, Canada; 2016. https://doi.org/http://dx.doi.org/10.11575/PRISM/27615.

6. Peng Q, Luo J, Wang C, Chen L, Tan S. Impact of station number and duration time per station on the reliability of objective structured clinical

examination (OSCE) scores: A systematic review and meta-analysis. BMC Med Educ. 2025;25(1):84. https://doi.org/10.1186/s12909-025-06691-0.

7. Salajegheh M, Razavi NS. Review of checklist and global rating form scoring methods in objective structured clinical examination stations: A narrative review. Strides in Development of Medical Education Journal. 2021;18(1):e1406. https://doi.org/10.22062/sdme.2021.195846.1046.

8. Regehr G, MacRae H, Reznick RK, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. Academic Medicine. 1998;73(9):993–7. https://doi.org/10.1097/00001888-199809000-00020.

9. Cade AE, Mueller N. Measuring the quality of an objective structured clinical examination in a chiropractic program: A review of metrics and recommendations. Journal of Chiropractic Education. 2024;38(1):9–16. https://doi.org/10.7899/JCE-22-29.

10. Farrell SE. Evaluation of student performance: Clinical and professional performance. Academic Emergency Medicine. 2005;12(4):302.e6-302.e10. https://doi.org/10.1197/j.aem.2004.05.037.

11. Daelmans HEM, Van der Hem-Stokroos HH, Hoogenboom RJI, Scherpbier AJJA, Stehouwer CDA, Van der Vleuten CPM. Global clinical performance rating, reliability and validity in an undergraduate clerkship. Neth J Med. 2005;63(7):279–84.

12. DeCloux K, Hoy G, Grum C, Kwakye G, Schiller J, Heidelbaugh J, et al. Assessing correlations between competency ratings and assessment-specific global rating scores across seven core clinical clerkships at the university of Michigan medical school. Academic Medicine. 2023;98(s3) Supplement_3:S198–9. https://doi.org/10.1097/ACM.0000000000005342.

13. Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague P-A, Turner R, et al. Inter-rater reliability: Comparison of checklist and global scoring for OSCEs. Creat Educ. 2012;3:937–42. https://doi.org/10.4236/ce.2012.326142.

14. Khan U, Khan YN. Correlation between task-based checklists and global rating scores in undergraduate objective structured clinical examinations in Saudi Arabia: a 1-year comparative study. J Educ Eval Health Prof. 2025;22:19. https://doi.org/10.3352/jeehp.2025.22.19.

15. Mahmoud A. A comparison of checklist and domain-based ratings in the assessment of objective structured clinical examination (OSCE) performance. Cureus. 2023;15(6):e40220. https://doi.org/10.7759/cureus.40220.

16. Saghir S, Farhana BS. Comparative analysis: Global rating scale vs. checklist in teaching and assessing skill competence. Pakistan Journal of Medicine and Dentistry. 2024;13(2):102–5. https://doi.org/10.36283/pjmd13-2/015.

17. Hodges B, Mcilroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 2003;37(11):1012–6.

18. Tavakol M, Dennick R. Making sense of Cronbach's alpha. Int J Med Educ. 2011;2:53–5. https://doi.org/10.5116/ijme.4dfb.8dfd.

19. Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach. 2015;37(12):1106–13. https://doi.org/10.3109/0142159X.2015.1009425.

20. Sim JH, Abdul Aziz YF, Vijayanantha A, Mansor A, Vadivelu J, Hassan H. A closer look at checklist scoring and global rating for four OSCE stations: Do the scores correlate well? Education in Medicine Journal. 2015;7(2):e39–44. https://doi.org/10.5959/eimj.v7i2.341.

21. Abass MO, Elimam M, Ahmed M. Comparison of a task-specific checklist and end exam global rating scale for scoring the objective structured clinical examination used to evaluate sixth year medical students in surgery at Shendi University, Sudan. East African Scholars Journal of Education, Humanities and Literature. 2020;2(6):257–62. https://doi.org/10.36349/EASJEHL.2020.v03i06.010.

22. Giavarina D. Understanding Bland Altman analysis. Biochem Med (Zagreb). 2015;25(2):141–51. https://doi.org/10.11613/BM.2015.015.