


An evaluation of the inter-rater reliability in a clinical skills objective structured clinical examination

V de Beer,¹ 3rd-year MB ChB student; J Nel,¹ 3rd-year MB ChB student; F P Pieterse,¹ 3rd-year MB ChB student; A Snyman,¹ 3rd-year MB ChB student; G Joubert,² BA, MSc; M J Labuschagne,¹ MB ChB, MMed (Ophthalmology), PhD 

¹ Clinical Simulation and Skills Unit, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa

² Department of Biostatistics, Faculty of Health Sciences, University of the Free State, Bloemfontein, South Africa

Corresponding author: M J Labuschagne (labuschagnemj@ufs.ac.za)

Background. An objective structured clinical examination (OSCE) is a performance-based examination used to assess health sciences students and is a well-recognised tool to assess clinical skills with or without using real patients.

Objectives. To determine the inter-rater reliability of experienced and novice assessors from different clinical backgrounds on the final mark allocations during assessment of third-year medical students' final OSCE at the University of the Free State.

Methods. This cross-sectional analytical study included 24 assessors and 145 students. After training and written instructions, two assessors per station (urology history taking, respiratory examination and gynaecology skills assessment) each independently assessed the same student for the same skill by completing their individual checklists. At each station, assessors could also give a global rating mark (from 1 to 5) as an overall impression.

Results. The urology history-taking station had the lowest mean score (53.4%) and the gynaecology skills station the highest (71.1%). Seven (58.3%) of the 12 assessor pairs differed by >5% regarding the final mark, with differences ranging from 5.2% to 12.2%. For two pairs the entire confidence interval (CI) was within the 5% range, whereas for five pairs the entire CI was outside the 5% range. Only one pair achieved substantial agreement (weighted kappa statistic 0.74 – urology history taking). There was no consistency within or across stations regarding whether the experienced or novice assessor gave higher marks. For the respiratory examination and gynaecology skills stations, all pairs differed for the majority of students regarding the global rating mark. Weighted kappa statistics indicated that no pair achieved substantial agreement regarding this mark.

Conclusion. Despite previous experience, written instructions and training in the use of the checklists, differences between assessors were found in most cases.

Afr J Health Professions Educ 2023;15(2):e1574. <https://doi.org/10.7196/AJHPE.2023.v15i2.1574>

The current trend in health professions education is to design systems of assessment, where the triangulation of multiple assessment methods and tools is used before decisions on high-stakes assessments are made.^[1,2] Performance-based evaluation is a form of assessment where a student must demonstrate a specific clinical skill to an assessor or evaluator who observes the person performing it.^[1,3] The objective structured clinical examination (OSCE) is one of the performance-based examination methods used to assess students in medicine and other health professions and is a valid and reliable tool for assessing clinical interactions and clinical performance with or without real patients. OSCEs are widely used in undergraduate and postgraduate medical programme performance evaluations.^[4] OSCEs test the 'shows how' aspect of Miller's pyramid^[5] when clinical and communication skills are assessed in an examination set-up.

Scoring is done with an assessment tool with a specific checklist or combination of a checklist and rating scale and contributes to the objectivity and thoroughness of the assessment of a specific skill. However, it is possible that mark allocation (for similar execution of a specific task) can differ.^[6,7] Training of raters aims to minimise rater disagreement and to make assessments more consistent.^[1,6] Nonetheless, assessor factors such as rater cognition, bias and observations remain one of the most important contributors to assessment error.^[1,7,8]

Mazor *et al.*^[9] looked at the assignment of professionalism ratings by assessors during OSCE encounters between medical students and patients. They reported that at each station, at least one of the assessors made a positive evaluative comment and another a negative evaluative comment on the same topic regarding the OSCE that was taking place. In other studies,^[7,8,10,11] it was found that different assessors would each form their own impressions regarding the student being assessed, even though they received the same information and training.

A study by Gingerich *et al.*^[12] attempted to determine if psychological differences could be the reason for poor inter-rater reliability. Their findings suggest that the personality and mood of the assessor may be a significant factor to take into consideration regarding inter-rater reliability. Mortsiefer *et al.*^[8] identified clinical experience, context and gender of the examiners as factors influencing the inter-rater reliability. These factors could pose a problem for those being assessed and have an impact on the marks and outcome of the assessment.^[7]

Clinical educators used as assessors in an OSCE often have different clinical backgrounds and various levels of experience. The aim of this study was to determine the inter-rater reliability of the experienced and novice assessors' final mark allocations during assessments of the third-year University of the Free State (UFS) medical students' OSCE after assessor training and provision of written instructions.

Methods

Study design

This was a cross-sectional analytical study.

Study population

The study population comprised 24 assessors who rated 145 third-year medical students' formal OSCE. Assessors were family physicians, general practitioners and nurse educators. No specialists in the field of the station were used. The assessors were selected by the Clinical Simulation and Skills Unit management and were trained on the correct use of the checklists for mark allocation. The training involved explaining the assessment sheet and detail on the assessment of the particular skill at each of the individual stations to all the assessors involved in the station. Assessors also received a sheet with instructions regarding the assessments 2 days before the OSCE. These assessors were placed in pairs at different stations, one with >5 years' experience in clinical education and who had assessed in >3 OSCEs, and a novice doing an OSCE assessment in the medical programme for the first or second time. Each assessor independently assessed and scored the same student for the same skill by completing their individual assessment checklists.

Method of data collection

Half of the class was evaluated during a morning evaluation session, and the other half during an identical evaluation session in the afternoon. Students were split into two circuits at each evaluation session. The students moved in turn in their assigned circuits, from station to station, where they had to demonstrate a specific competency at each station. Of the 15 stations, 9 had assessors, whereas the remaining 6 were unmanned. Three stations with 2 assessors per circuit, i.e. a urology history-taking station, a gynaecology skills station and a respiratory examination station, were included in this study. The researchers deliberately selected a history-taking station, a technical skill station and a physical examination station to evaluate the inter-rater reliability, as the nature of these stations is so different. The circuit and station were clearly indicated on the assessor checklists. The checklists differed only in paper colour to distinguish between different assessors' assessment tools at different stations. This set-up ensured that the assessors' checklists were paired correctly when the researchers received the checklists from the Clinical Simulation and Skills Unit management team. Assessors did not discuss students' mark allocation with one another, and they were not informed whose mark would be used in the assessment.

At the respiratory examination station, students were evaluated on 9 items on a scale of 0 or 1 and 12 items on a scale of 0, 1 or 2, giving a maximum final mark of 33. The gynaecology skills evaluation consisted of 13 items on a scale of 0 or 1 and 10 items on a scale of 0, 1 or 2, giving a maximum final mark of 33. The urology history-taking evaluation consisted of 14 items on a scale of 0 or 1 and 8 items on a scale of 0, 1 or 2, giving a maximum final mark of 30. All final marks were expressed as a percentage of the maximum achievable. At each station, assessors were requested to also give a global rating mark as an overall impression on a scale of 1 - 5. Not every student received a global rating mark from the assessors. The global rating does not contribute to the student's final mark.

After completion of the OSCE, the checklists evaluating the students' skills were collected by the head of the simulation unit. After the checklists were processed by the unit's staff, the names of the assessors and students were redacted and replaced by specific study numbers.

Statistical analysis

Statistical analysis was done by the Department of Biostatistics, Faculty of Health Sciences, UFS. The following were determined: mean scores of the three stations; inter-rater differences in the mean final mark allocations (with 95% confidence intervals (CIs)) and differences in the allocated global rating marks. Weighted kappa statistics were also calculated to determine the agreement between assessors' categorisation regarding pass/fail/distinction on the final mark and between assessors' global rating mark.

Ethical considerations

Permission was obtained from the Health Sciences Research Ethics Committee, UFS (ref. no. HSREC-S 10/2016), the dean of the Faculty of Health Sciences, the head of the School of Medicine and the vice-rector of Research.

Informed consent was obtained from all the assessors. The redaction of the names of the assessors and students ensured anonymity to the research team. The marks given by the most senior or experienced assessor were used for the student's academic record.

Results

Twenty-four assessors rated 145 students in this study. A total of 870 assessments were evaluated, but as the global rating was not compulsory for the students' final mark, 51 global rating scores were left incomplete (94.1% response rate). Four final marks were not included because of technical reasons at the respiratory examination station (99.5% response rate).

The mean values of the allocated final marks were: urology history-taking station - 53.4%; respiratory examination station - 60.4%; and gynaecology skills station - 71.1%.

Table 1 summarises the final marks and the differences in the final mark between assessors allocated to a specific circuit of students. The first assessor number per pair denotes the more experienced assessor. There was no consistency within or across stations regarding whether the more experienced or less experienced assessor gave higher marks. In one respiratory examination station, the marks of both assessors had large coefficients of variation. Seven (58.3%) of the 12 pairs of assessors differed by >5% regarding the final mark, ranging from 5.2% to 12.2%. When considering the 95% CIs, for only two pairs of assessors the entire CI was within the 5% range, whereas for five pairs of assessors the entire CI was outside the 5% range. Only two pairs of assessors differed by >10%, and the CIs for these differences indicated that the largest differences expected were -13.2 and -13.8. According to both assessors, the urology history-taking station marks were generally low, and differences between assessors regarding fail/pass marks frequently occurred. In one respiratory examination and three gynaecology skills stations, one assessor's average mark was well below 70%, whereas the other assessor's average mark was close to a distinction, and differences regarding pass/distinction frequently occurred. Weighted kappa statistics measuring agreement regarding fail/pass/distinction indicated that only in one pair could the agreement be considered substantial (kappa >0.60). This was a urology history-taking station. One of the other urology history-taking stations had the lowest kappa value (0.03).

Table 2 shows the percentage of times a pair of assessors agreed (gave the same rating) or disagreed (gave different ratings) on the allocated global ratings. For the respiratory examination and gynaecology skills stations, all assessor pairs differed for the majority of students, whereas for three of

Table 1. Inter-rater differences in the final mark allocation

Station	Compared assessors*	Assessor mean score (coefficient of variation) of final mark, %	Mean difference (95% CI)	Differences regarding fail/pass, pass/distinction, and fail/distinction of students assessed, %			Weighted kappa statistics†
				Fail/pass	Pass/distinction	Fail/distinction	
Urology history taking							
Circuit A, round 1 (n=37)	1	57.7 (19.1)	-0.5 (-2.4; 1.3)‡	2.7	2.7	0	0.74
	2	58.2 (17.7)					
Circuit A, round 2 (n=37)	3	45.5 (23.8)	-8.7 (-10.7; -6.8)§	35.1	2.7	0	0.03
	4	54.2 (18.9)					
Circuit B, round 1 (n=35)	5	57.5 (16.4)	5.6 (3.0; 8.3)	20.0	0	0	0.47
	6	51.9 (18.2)					
Circuit B, round 2 (n=36)	7	55.0 (16.3)	7.4 (5.7; 9.1)§	38.9	0	0	0.31
	8	47.6 (17.8)					
Respiratory examination							
Circuit A, round 1 (n=37)	9	63.3 (16.4)	3.4 (1.6; 5.3)	16.2	8.1	0	0.51
	10	59.9 (18.6)					
Circuit A, round 2 (n=37)	11	54.8 (17.5)	2.5 (0; 5.1)	24.3	0	0	0.52
	12	52.3 (21.6)					
Circuit B, round 1 (n=32)	13	61.7 (17.3)	-12.2 (-13.8; -8.5)§	6.3	37.5	0	0.17
	14	73.9 (12.9)					
Circuit B, round 2 (n=36)	15	61.7 (24.4)	-3.9 (-0.3; 8.0)	8.3	25.0	2.8	0.41
	16	57.8 (23.2)					
Gynaecology skills							
Circuit A, round 1 (n=37)	17	72.6 (16.3)	0.3 (-2.6; 3.3)‡	5.4	32.4	0	0.37
	18	72.3 (15.8)					
Circuit A, round 2 (n=37)	19	65.6 (16.7)	-8.5 (-11.3; -5.7)§	8.1	35.1	2.7	0.26
	20	74.1 (11.0)					
Circuit B, round 1 (n=35)	21	72.2 (15.0)	5.2 (1.8; 8.6)	8.6	20.0	2.9	0.38
	22	67.0 (14.1)					
Circuit B, round 2 (n=36)	23	67.0 (18.2)	-10.9 (-13.2; -8.3)§	5.6	47.2	5.6	0.15
	24	77.9 (11.9)					

CI = confidence interval.

*The first assessor number per pair denotes the more experienced assessor.

†Weighted kappa statistics take into account how large the discrepancy in rating is, not only whether there is a discrepancy or not.

‡Entire CI falls within the 5% range.

§Entire CI falls outside the 5% range.

the four urology history-taking station assessor pairs, the assessors agreed for the majority of the students regarding the global rating mark. Two gynaecology skills stations had the largest discrepancies, with 16.7% and 24.3% of students' scores differing between the assessors ≥ 2 units on the 5-point scale. The weighted kappa statistics indicate that no pair of assessors achieved substantial agreement on the global rating mark. As with the final mark, there was no consistency within or across stations regarding whether the more experienced or less experienced assessor gave the higher mark.

Discussion

The authors deliberately selected a history station, examination station and clinical skills station because the difference in the nature of these stations could play a role in the inter-rater reliability. The typical competencies tested in an OSCE include history taking, physical examination, communication skills, practical/technical skills and clinical reasoning.^[1]

History-taking and interviewing skills are needed to gather essential and accurate information from patients. The urology history-taking station was the selected history station in this study. The poor performance of

the students in this station could be due to urology history taking being a difficult skill to master, and especially junior students struggle with history-taking skills and clinical reasoning overall.^[13,14] When assessing a history-taking station, the assessor must listen carefully and maintain a high level of concentration. Assessor distraction, students' language proficiency or assessor fatigue could play a role in the poor inter-rater reliability of this station. Given the generally low marks of ~50%, the differences regarding whether the mark reflects a pass or fail are to be expected.

The data indicate an overall poor inter-rater reliability, as for the majority of the pairs of assessors the mean final mark allocations fell outside the 5% difference margin, and differences regarding fail/pass/distinction frequently occurred with only one assessor pair achieving substantial agreement regarding this classification. According to Jönsson and Svingby,^[15] the reliable scoring of performance assessments can be enhanced using checklists, especially if they are analytical, topic specific and complemented with exemplars and/or assessor training. However, the gynaecological skills stations had poor inter-rater reliability despite the assessed procedures having a structured methodology and checklist for

Table 2. Agreement of global rating marks allocated

Station	Compared assessors*	Agree (the same mark allocated), %	Disagree (different marks allocated), %	Weighted kappa statistics†	Students rated by both assessors, <i>n</i>
Urology history taking					
Circuit A, round 1	1 and 2	38.2	61.8	0.49	34
Circuit A, round 2	3 and 4	57.1	42.9	0.41	35
Circuit B, round 1	5 and 6	67.7	32.4	0.29	34
Circuit B, round 2	7 and 8	77.1	22.9	0.58	35
Respiratory examination					
Circuit A, round 1	9 and 10	38.9	61.1	0.31	36
Circuit A, round 2	11 and 12	34.3	65.7	0.30	35
Circuit B, round 1	13 and 14	-	-	-	0
Circuit B, round 2	15 and 16	43.8	56.3	0.27	32
Gynaecology skills					
Circuit A, round 1	17 and 18	48.7	51.4	0.28	37
Circuit A, round 2	19 and 20	32.4	67.6	0.07	37
Circuit B, round 1	21 and 22	41.2	58.8	0.25	34
Circuit B, round 2	23 and 24	45.7	54.3	0.22	35

*The first assessor number per pair denotes the more experienced assessor.

†Weighted kappa statistics take into account how large the discrepancy in rating is, not only whether there is a discrepancy or not.

assessment. Each assessor demonstrates a certain amount of variability due to certain characteristics, such as their personality traits and preconceived notions that may have a small or significant influence on their clinical judgement, according to Wood^[16] and Williams *et al.*^[17] In our study, no pattern could be found within or across stations regarding whether the more experienced or less experienced assessor gave higher marks.

Training the assessors to obtain stronger reliability may not be possible, as stated by several authors.^[8,17,18] They suggest that medical assessors may be impervious to training. In these studies, it is mentioned that training may benefit some assessors while it does not affect others. In the consensus statement following the 2020 Ottawa conference, it is recommended that assessor training should focus on conduct, behaviours and bias.^[19] Assessor factors remain the most important contribution to overall assessment error, even in a well-designed and valid OSCE station.^[6] Schleicher *et al.*^[7] identified gender-related bias and suggested that assessors be made aware of assessor bias.

The assessors who scored this OSCE are professionals in the healthcare field but not specialists, and it was assumed that there would be similarity in the global rating mark given because they have an idea of what a professional doctor should act like. However, some studies^[1,7,10,11] found that different assessors would each form their own impressions, and it was therefore not unexpected that the agreement among assessors was not strong on the global rating. The increased diversity in the student population, students' use of language and assessors' interpretation and assessor fatigue may also contribute to the marks allocated by various assessors.

Conclusion and recommendations

Despite instructions and training in the use of the checklists, differences between assessors were found in the majority of cases. We agree with Smee^[6] and Boursicot *et al.*,^[1] who state that the reliability of an OSCE can be improved by making use of more stations (at least 12), well-designed checklists, consistency in simulated patient portrayals and assessors who

score consistently and are unbiased. The validity depends on the alignment of the curriculum with the assessment and the quality of the assessment checklist reviews.

The authors propose that a possible solution to prevent poor inter-rater reliability from affecting the students' final marks could be to have more than one assessor evaluate the same student and that the average mark is then used as the final mark. This could, however, be very difficult to achieve in an environment with a limited number of available assessors. In the consensus statement after the 2020 Ottawa conference, a recommendation regarding OSCEs was to 'embrace examiner variability by ensuring sufficient numbers of examiners, rather than trying to standardise their judgements'.^[1] In the triangulation of multiple assessment instruments as proposed for programmatic assessment, decisions on students' competencies and progress are based on a combination of various assessment methods, resulting in compensation for the shortcoming of individual tools in an assessment.^[2]

Declaration. This was an undergraduate MB ChB student research project – the students were in their third year at the time of the study (2017).

Acknowledgements. We thank Ms W du Toit, senior assistant officer, Clinical Simulation and Skills Unit, for administrative support, Dr J E Raubenheimer for advice during the planning of the project and initial statistical analysis, and Ms T Mulder, medical editor/writer, Faculty of Health Sciences, for technical and editorial finalisation of the manuscript.

Author contributions. VdB, JN, FPP and AS developed the protocol, collated the data and did the initial write-up of this study. GJ assisted with the planning, performed data analysis and assisted with the interpretation and write-up of the article. MJ supervised the study, suggested the concept, assisted with the protocol development, data collection and interpretation of data and write-up of the manuscript.

Funding. None.

Conflicts of interest. None.

1. Boursicot K, Kemp S, Wilkinson T, et al. Performance assessment: Consensus statement and recommendations from the 2020 Ottawa conference. *Med Teach* 2021;43(1):58-67. <https://doi.org/10.1080/0142159X.2020.1830052>
2. Schuwirth LW, van der Vleuten CP. Current assessment in medical education: Programmatic assessment. *J Appl Test Technol* 2019;20(S2):2-10.
3. Harden RM. Outcome-based education: AMEE Guide No. 14. Part I: An introduction to outcome-based education. *Med Teach* 2009;21(1):7-14. <https://doi.org/10.1080/01421599979969>
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-e1446. <https://doi.org/10.3109/0142159X.2013.818634>
5. Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65(9 Suppl):S63-S67. <https://doi.org/10.1097/00001888-199009000-00045>
6. Smee S. Skill based assessment. *BMJ* 2003;326(7391):703-706. <https://doi.org/10.1136/bmj.326.7391.703>
7. Schleicher I, Leitner K, Juenger J, et al. Examiner effect on the objective structured clinical exam - a study at five medical schools. *BMC Med Educ* 2017;17(1):71. <https://doi.org/10.1186/s12909-017-0908-1>
8. Mortsiefer A, Karger A, Rothhoff T, Raski B, Pentzek M. Examiner characteristics and interrater reliability in a communication OSCE. *Patient Educ Coun* 2017;100(6):1230-1234. <https://doi.org/10.1016/j.pec.2017.01.013>
9. Mazor KM, Zanetti ML, Alper EJ, et al. Assessing professionalism in the context of an objective structured clinical examination: An in-depth study of the rating process. *Med Educ* 2007;41(4):331-340. <https://doi.org/10.1111/j.1365-2929.2006.02692.x>
10. Kenny DA. PERSON: A general model of interpersonal perception. *Pers Soc Psychol Rev* 2004;8(3):265-280. https://doi.org/10.1207/s15327957pspr0803_3
11. Park B, DeKay ML, Kraus S. Aggregating social behavior into person models: Perceiver-induced consistency. *J Pers Soc Psychol* 1994;66(3):437-459. <https://doi.org/10.1037//0022-3514.66.3.437>
12. Gingerich A, Regehr G, Eva KW. Rater-based assessments as social judgments: Rethinking the etiology of rater errors. *Acad Med* 2011;86(10 Suppl):S1-S7. <https://doi.org/10.1097/ACM.0b013e31822a6cf8>
13. Seitz T, Raschauer B, Längle AS, Löffler-Stastka H. Competency in medical history taking - the training physicians' view. *Wien Klin Wochenschr* 2019;131(1-2):17-22. <https://doi.org/10.1007/s00508-018-1431-z>
14. McKenna L, Innes K, French J, Streitberg S, Gilmour C. Is history taking a dying skill? An exploration using a simulated learning environment. *Nurse Educ Pract* 2011;11(4):234-238. <https://doi.org/10.1016/j.nepr.2010.11.009>
15. Jönsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. *Educ Res Rev* 2007;2(2):130-144. <https://doi.org/10.1016/j.edurev.2007.05.002>
16. Wood TJ. Exploring the role of first impressions in rater-based assessments. *Adv Health Sci Educ Theory Pract* 2014;19(3):409-427. <https://doi.org/10.1007/s10459-013-9453-9>
17. Williams RG, Klamen DA, McGaghie WC. Cognitive, social and environmental sources of bias in clinical performance ratings. *Teach Learn Med* 2003;15(4):270-292. https://doi.org/10.1207/S15328015TLM1504_11
18. Cook DA, Dupras DM, Beckman TJ, Thomas KG, Pankratz VS. Effect of rater training on reliability and accuracy of mini-CEX scores: A randomised, controlled trial. *J Gen Intern Med* 2009;24(1):74-79. <https://doi.org/10.1007/s11606-008-0842-3>

Accepted 16 November 2022.