Original Article

OSCE grading: A cross sectional study on inter-observer variability in assessment

Tehzeeb Zehra, Sahira Aaraj, Humera Naeem, Humera Naz Altaf, Rahila Aamir, Mahnoor Aitazaz

Shifa International Hospital, Shifa Tameer-e-Millat University, Islamabad, Pakistan

Objective: To determine the inter observer variability in OSCE grading of students.

Methodology: This observational study was conducted at Shifa College of Medicine and included students of 3, 4 and 5 year of medicine, gynae, surgery and pediatric clerkships undergoing their end of clerkship OSCE. We enrolled 217 students, 7 dropped out and 210 were analyzed. Each department conducted it on one station of each OSCE with four preceptors assessed the same student on that station. Scores were recorded in two sets; set 1 had preceptor 1 with standardized checklist and preceptor2 without any check list; set 2 had both preceptors given standardized checklists. SPSS 25 was used for data analysis. ANOVA was applied to see variance of grading, figures and tables gathered showing p values.

Results: In this study, 21.9%, 41%, 37.1% students belonged to year 3, 4 and 5, respectively. 22.9%, 37.1%, 11.4 %, 28.6% belonged to gynae, medicine, surgery and pediatric clerkship respectively. 51.4% were male and 48.6% were female. Inter-observer variability in set 1 was statistically significant ($p = 0.000$), and in set 2 was also significant ($p = 0.000$) in all groups regarding subject and year.However, gender of students did not affect the grading variability.

Conclusion: The inter-observer variability among preceptors was high whatever the assessment method they used, concluding the role of some other factors in OSCE grading then the method alone.

Keywords: OSCE grading, assessment methods, preceptors.

INTRODUCTION

Assessment of students is essential to enable the faculty to know how much they are picking up from the learning objectives set for them. There are many different ways to assess clinical learning in medical students. As described by George Miller, assessment of skills in students fall under a framework.[1] At the base level, students should have knowledge of what they have been taught and this can be tested by objective test methods. At the next level, they should know how to analyze and use that knowledge; this can be tested through MCQs and mini vivas. These levels of learning can be tested through Objective Structured Clinical Examinations (OSCE).

Many studies have described the setup of an OSCE.[2,3] It consists of multiple stations, in which students have 4 – 5 minutes to perform tasks after which they continue to move on to the next station till they have gone through them all. Student's performance is graded by faculty at each station, and these grades determine the student's credibility in their respective clerkships. However, there are many bias that need to be overcome in the OSCE format in stations where teacher student encounters are present.[4-6] In stations where the student examiner encounters were more pleasurable, a "halo effect" was present due to which students were graded higher and the level of training of the examiner also proved to present as a bias.[7] In situations where the examiners were less trained, their scoring was more inflated as compared to trained examiners.

In another study by Stroud et al, it was shown that if examiners knew the student beforehand with a positive familiarity, they had a greater tendency to score them higher as compared to students they did not know or had negative familiarity with.[8] Such biases can be overcome by the use of keys to make the results more standardized.[9] These keys can be in the form of checklists or in the form of global scales like Likert scales to assess their performance, as described by Tavakol and Pinner.[10] The aim of this study was to bring awareness about more transparent, just and less biased OSCE to ultimately eradicate guess methods in marking.

METHODOLGY

This descriptive observational study was conducted in four departments of Shifa College of Medicine. The protocol was approved by Shifa Tameer-e-Millat University's Research Board and a written informed consent was taken from all participants. Medical students from 3rd, 4th and 5th year in, medicine, gynecology, surgery and pediatric clerkships undergoing their regular OSCE at the end of clerkship were included

in the study.The data was gathered from back-to-back OSCE from all departments within 6 weeks from June to July 2019.

We enrolled 217 students using WHO sample size calculator.The sample size was calculated by taking $\alpha = 0.05$ at 95% confidence level, the absolute precision was set at 5%. With drop out of 7 cases, 210 students were analyzed for results. Score of students by each of these four preceptors recorded in two sets; set 1 had preceptor 1 with standardized checklist and preceptor 2 without any key so to mark according to guess; set 2 had both preceptors given standardized checklist. So, we got four results for eachstudent on that particular station.

Statistical Analysis: Data were analyzed with SPSS version 25. One way ANOVA was applied to see variance among grades of set 1 and set 2 separately. Univariate analysis performed subject wise to see that how much each subject is causing variance in marking in set 1 and set 2. $p < 0.05$ was considered significant.

RESULTS

Out of 210 students, 21.9% belonged to year 3, 41% to year 4, and 37.1% to year 5. Regarding subject wise distribution, 22.9% students belonged to gynae rotation, 37.1% medicine, 11.4% surgery, and 28.6% to pediatrics. 51.4% students were male and 48.6% students were female.Marking variability in set 1 examiners (one using standardized check list and other doing guess marking) is statistically significant ($p = 0.000$) while it is also significant statistically among examiners of set 2 ($p = 0.000$) (both using standardized check list).

The result difference between set 1 and set2 was statistically significant ($p = .000$) (Table 1). Univariate analysis performed subject wise to see that how much each subject is causing variance in marking in set 1 and set 2. All four subjects exhibited separate significant variance in results of both set 1 and set 2 (p value in all 8 sets is 0.000) (Table 2).

Table 1: One way ANOVA indicating variability contributed by all three groups in set 1.

		Sum of Squares	df	Mean Square	F	Sig.
Gender	Between Groups	7.240	29	.250	.994	.481
	Within Groups	45.217	180	.251		
	Total	52.457	210			
Subject	Between Groups	149.054	29	5.140	7.771	.000
	Within Groups	119.060	180	.661		
	Total	268.114	210			
Class year	Between Groups	25.962	29	.895	1.730	.017
	Within Groups	93.162	180	.518		
	Total	119.124	210			

Table 2: One way ANOVA indicating variability contributed by all three groups in set 2.

		Sum of Squares	df	Mean Square	F	Sig.
Gender	Between Groups	8.676	38	.228	.892	.652
	Within Groups	43.782	171	.256		
	Total	52.457	210			
Subject	Between Groups	160.583	38	4.226	6.720	.000
	Within Groups	107.532	171	.629		
	Total	268.114	210			
Class year	Between Groups	34.793	38	.916	1.857	.004
	Within Groups	84.330	171	.493		
	Total	119.124	210			

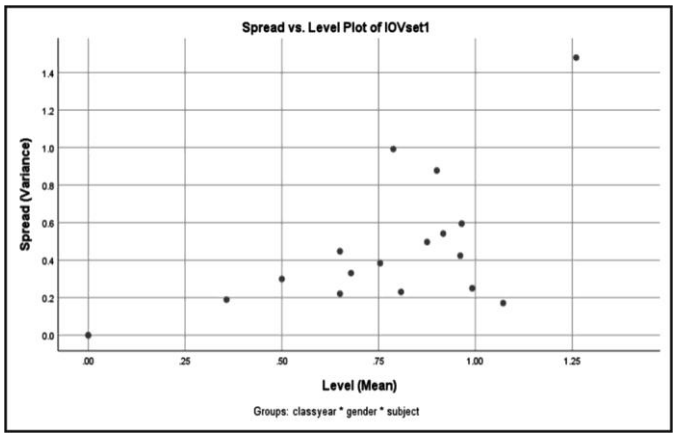

Fig. 1: Spread vs. level of plot in set 1.

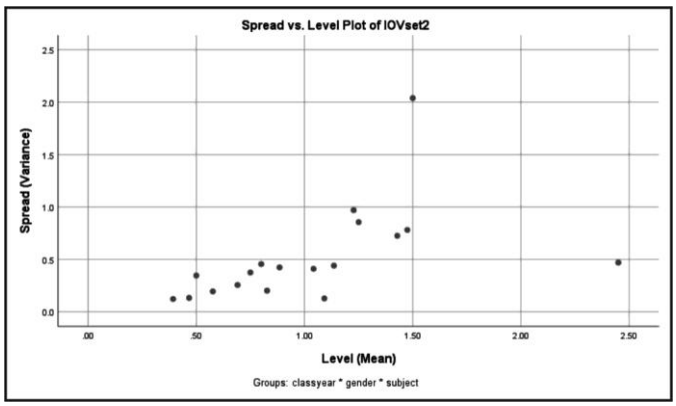

Fig. 2: Spread vs. level of plot in set 2.

Surgery subject was effecting the overall variance the most in set 1 (SD = .82505) and in set 2 (SD = 1.20611). Univariate analysis year wise indicated that none of the year had significant variability effect in set 1 but year 4 in set 2 (SD = .90710). Multivariate analysis (manova) used to see the effect of different variables on inter-observer variability. Wilks lambda's exact stats confirmed that subject wise interobserver variability is most effected (value = 0.000).Overall variance, among set 1 examiners and set 2 examiners shown in (Fig. 1 and 2, respectively).

DISCUSSION

OSCE is a widely employed tool for measuring clinical competence.[12] Our study results bore that using checklist scores only, doesn't minimize the inter observer rating differences so the reliability of results does not increase. Checklist scores and global ratings correlated well for the station as a whole, as well as across the circuits as adjudicated in many studies.[18]

However, despite the increasing introduction of OSCE to assess clinical competencies, concerns of higher variability still exist.[13] Mostly, the "hawk-dove" effect is mentioned, which means that some examiners are consistently stringent, while others are consistently lenient. This effect is observed in many studies and cannot be easily eliminated.[14-16]

Delivering OSCE, examiner's training is a necessary yet challenging part of the OSCE process. A novel approach to implementing training for OSCE examiners was trialed by delivering large-group education sessions at major teaching hospitals.[17] Schleicher et al, underlines the imperative for regular evaluation and training of examiners as conclusion in his study.[13]

We also selected only one station in each OSCE, because it was not possible to arrange a big faculty to cover four corners of every station, however, if we apply the result to the whole OSCE for each students then results may sway more.Our study results showed that inter observer variability is most effected due to subject difference, one reason may be diverse distribution of sample, subject wise as the sample from surgery subject was quite small, moreover probably indicating norm bias which is also different in different specialties.

So, in the drive toward comprehensive assessment, OSCE stations and checklists may become increasingly complex.[12] To exclude norm bias, preceptor's related bias, environmental bias, and applying uniformity in all these aspects, we need to take OSCE in a bias free environment which is not humanly conceivable. However, may become possible by using artificial intelligence in OSCE in future which would be congruently helpful also in situations like covid-19.

CONCLUSION

The inter-observer variability among preceptors was high whatever the assessment method they used, concluding the role of some other factors in OSCE grading then the method alone.

Author Contributions:
Conception and design: Tehzeeb Zehra, Rahila Aamir, Mahnoor Aitazaz.
Collection and assembly of data: Tehzeeb Zehra, Sahira Aaraj, Humera Naeem, Humera Naz Altaf, Rahila Aamir, Mahnoor Aitazaz.
Analysis and interpretation of the data: Tehzeeb Zehra, Sahira Aaraj.
Drafting of the article: Tehzeeb Zehra, Mahnoor Aitazaz.
Critical revision of the article for important intellectual content: Tehzeeb Zehra, Humera Naeem.
Statistical expertise: Tehzeeb Zehra, Humera Naeem.
Final approval and guarantor of the article: Tehzeeb Zehra.
Corresponding author email: Tehzeeb Zehra: drteh.ali@gmail.com
Conflict of Interest: None declared.
Rec. Date: Mar 7, 2020 Revision Rec. Date: Jul 15, 2020 Accept Date: July 2, 2022.

REFERENCES

1. Miller GE. The assessment of clinical skills/competence/ performance. Acad Med 1990;65:63-7.

2. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. Ann Surg 1995;222:735-9.
3. Harden RT, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975;1:447-1.
4. Schleicher I, Leitner K, Juenger J, Moeltner A, Ruesseler M, Bender B, Sterz J, et al. Examiner effect on the objective structured clinical exam–a study at five medical schools. BMC Med Educ 2017;17:71-5.
5. Chong L, Taylor S, Haywood M, Adelstein BA, Shulruf B. The sights and insights of examiners in objective structured clinical examinations. J Educ Eval Health Prof 2017;14:34-9.
6. Chong L, Taylor S, Haywood M, Adelstein BA, Shulruf B. Examiner seniority and experience are associated with bias when scoring communication, but not examination, skills in objective structured clinical examinations in Australia. J Educ Eval Health Prof 2018;15:17-3.
7. Guraya S, Alzobydi A, Salman S. Objective structured clinical examination: Examiners' bias and recommendations to improve its reliability. J Med Sci 2010;1:269-7.
8. Stroud L, Herold J, Tomlinson G, Cavalcanti RB. Who you know or what you know. Effect of examiner familiarity with residents on OSCE scores. Acad Med 2011;86:8-3.
9. Read EK, Bell C, Rhind S, Hecker KG. The use of global rating scales for OSCEs in veterinary medicine. PloS One 2015;10:121-7.
10. Tavakol M, Pinner G. Enhancing Objective Structured Clinical Examinations through visualization of checklist scores and global rating scale. Int J Med Educ 2018;9:132-7.
11. Pell G, Homer M, Fuller R. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach 2015;37:1106-1.
12. Hurley KF, Giffin NA, Stewart SA, Bullock GB. Probing the effect of OSCE checklist length on inter-observer reliability and observer accuracy. Med Educ Online 2015;20:2924-2.
13. Schleicher I, Leitner K, Juenger J. Examiner effect on the objective structured clinical exam – a study at five medical schools. BMC Med Educ 2017;17:71-5.
14. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. BMC Med Educ 2006;6:42-6.
15. Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Adv Health Sci Educ Theory Pract 2008;13:617-3.
16. Inn Y, Cantillon P, Flaherty G. Exploration of a possible relationship between examiner stringency and personality factors in clinical assessments: a pilot study. BMC Med Educ 2014;14:105-2.
17. Reid K, Smallwood D, Collins M, Sutherland R, Dodds A. Making OSCE examiner training on the road: reaching the masses. J Med Educ 2016;323:31-7.
18. Abass MO, Ahmed MEM. 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. Eas J Humanit Cult Stud 2020;6:257-2.