THE ASIA-PACIFIC SCHOLAR
MEDICAL AND HEALTH PROFESSIONS EDUCATION

ORIGINAL ARTICLE

Submitted: 2 September 2023
Accepted: 9 January 2024
Published online: 2 July, TAPS 2024, 9(3), 15-21
https://doi.org/10.29060/TAPS.2024-9-3/OA3129

Standard setting OSCE: A comparison of arbitrary and Hofstee methods in a low stake OSCE

Uzma Khan

Department of Clinical Sciences, College of Medicine, Al Rayan Colleges, Saudi Arabia

Abstract

Objectives: To compare the cut scores and pass/fail rates achieved by arbitrary 60% method and Hofstee method in an undergraduate year 4 end semester objective structured clinical examination (OSCE) and check the possibility of using Hofstee method of standard setting in future exams.

Method: 102 medical students of year 4 underwent a 10 station OSCE exam conducted in a state of art simulation lab in 3 cycles. The cut scores were calculated using arbitrary method aiming at 60% of task achievement and by Hofstee method. The student's obtained scores were compared for cut scores and pass rates for individual stations and the entire exam.

Results: The arbitrary and Hofstee methods of standard setting leads to different cut scores. For the individual stations it was 60% vs 65-70% and for the overall score it was 60% vs 70%. The percentage of students failing the exam is 13.7% based on arbitrary scores and is 29.4% when Hofstee cut score is applied.

Conclusions: The two methods lead to different cut scores and students' failure rates. Overall, Hofstee method is more appropriate for assessing competencies in an OSCE exam in medical schools as it leads to calculation of cut scores based on the difficulty level of the station/exam and the examiners expected level of performance by the students.

Keywords: *Objective Structured Clinical Examination, Standard Setting, Hofstee Method, Arbitrary Method*

Practice Highlights

- Standard settings of OSCE identifies objective, reliable and valid cutoff scores.
- Arbitrary method scrutinises the test content and nominates the percentage of items to be answered correctly.
- Hofstee method is calculative and avoids illogical very high and low scores.
- A retrospective descriptive study design assessing applicability of Hofstee method in low stake exam in a private medical school in Saudi Arabia.
- Students' failure rate increased with applying Hofstee standard settings in OSCE.

I. INTRODUCTION

Objective structured clinical examination OSCE is invented in 1975 by Harden (Harden & Gleeson, 1979) for the assessment of learners' clinical competences and behaviors by using actors and choreographed storylines (Hodges, 2003). He succeeded in controlling the classic variables, the patient and the examiner, that enabled him to establish a comprehensive and objective assessment (Khan et al., 2013) of competence by defining clearly what skills, attitudes, problem solving capabilities and factual knowledge are to be measured (Harden et al., 1975).

As quoted by Harden "Competency is the compound of cognitive, psychomotor and affective skills as appropriate, while competence is an attribute of a person" (Khan et al., 2013).

During an OSCE, candidates are supposed to execute different clinical tasks in a simulated setting (Khan et al., 2013). As a rule, the students rotate through several time limited stations in which they are expected to interact with a standardised patient (SP), mannequins or simulation models and perform a specified task meanwhile they are being assessed by examiners using standardised rating instruments (Pugh & Smee, 2013). OSCE gets rid of many detrimental concepts in examining students, faced previously with other exam methods, by compelling them all to go through the same scope and criteria for assessment (Zayyan, 2011). This has made it a valuable evaluative tool in medical practice, so it has been adopted in countries all over the world, in all the high stakes examinations in USA (Dwivedi et al., 2020), Canada (Pugh & Smee, 2013), and the UK (Gormley, 2011).

Reliability and validity of the OSCE exam is directly related to how it is implemented (Harden & Gleeson, 1979) and can be maximised by several ways, the first and the foremost is the designing of structured reconcilable mark schemes for several stations observed by different trained assessors which will eliminate the individual assessor bias (Gormley, 2011). Competence assessment will be more reliable by arranging a variety of patient presentations for different cases and skills and standardising patients' performance (Dent et al., 2021; Khan et al., 2013).

As a prerequisite of a good test, a process called standard setting must be set that if followed will lead to a fair decision (Boulet et al., 2003). The inferences derived from a test result matter a lot to the examiners, examinees and the institutes (Norcini et al., 2011). Cusimano in his review paper defines standard setting as a process that determines "what is good enough" for assessing competence, which itself is continuously changing, and leads to separation of a competent student from an incompetent (Cusimano, 1996). According to Harden the standard is the score that decides pass fail status of the students, also known as pass fail point. It provides an answer to the question" how much or what is good enough to know?" (Dent et al., 2021). He has defined the standard setting as the process of translating a description of characteristics denoting the desired level of performance into a number that applies to a particular test" (Dent et al., 2021).

At the time of setting the standards, the purpose of the exam must be considered (Kamal et al., 2020) along with the consequences of letting an incompetent examinee get through the exams and acquire medical licensure that could be devastating (Gormley, 2011).

Standard setting methods are designated into norm-referenced, criterion-referenced methods and a third category of combination or compromise methods (Dwivedi et al., 2020; Kamal et al., 2020). In absolute or criterion referenced standards a benchmark is set based on certain predefined criteria and the candidate performance is tested according to that standard competency or mastery. Whereas Norm-referenced, also called relative methods, are based on identification of the cut-off score relative to performance of the group or top scoring students taking the examination, which results in loss of motivation for progressing and improving in top scoring students (Dwivedi et al., 2020).

For assessing the quality of OSCE exam, the determinants are dictated by the method of standard setting. The AMEE Guide 85 describes a number of standard setting methods of which Cohen, Angoff, Borderline Regression, Borderline Group, Hofstee Method, and the fixed arbitrary 60% method are some of the commonly used (McKinley & Norcini, 2014).

Cohen method is the best form of the norm-referenced standard setting methods extensively used in low stakes exams. The best performing students' mark is used as a reference point to define the difficulty of the exam. The remaining students' scores are arranged from the lowest to the highest scores; the mean value of the top 5% of the scores is calculated, and finally, 60% of the total mean score is considered as the standard/passing score (Kamal et al., 2020).

Angoff method is entirely based on test/examination items (Pell et al., 2010). In this method the pass mark is statistically calculated on item or station characteristics, and it varies according to the difficulty level of the station defined by the items on checklist, but the students' performance is not taken into consideration. (Dwivedi et al., 2020; Impara & Plake, 1997).

The borderline methods are reasonable and defensible as they are based on candidates' performance (Kaufman et al., 2000; Pell et al., 2010). So borderline regression and modified borderline Group methods are also known as "Examinee centered" methods (Dwivedi et al., 2020). Borderline group methods necessitate the examiner be able to identify what is considered as minimally acceptable performance. The mean or median score of minimally acceptable performances is declared as cutoff score (Cusimano, 1996; Humphrey-Murto & MacFadyen, 2002). Apart from checklist scores, a global grade is also awarded which provides insights into quality of assessment (Pell et al., 2010; Smee et al., 2022).

Hofstee method aims to achieve a balance between the norm and criterion reference judgements and is a combination/compromise method (Dwivedi et al., 2020). In this method the examiners specify 4 values before the exam: the maximum and the minimum percentage correct, and the maximum and minimum acceptable percentage of failures (Smee et al., 2022). This method is more calculative, but it avoids illogical very high and very low scores (Cusimano, 1996; Kamal et al., 2020).

The arbitrary 60% method uses faculty wide standard of passing score of 60% in OSCE exam and is the easiest to implement (Humphrey-Murto & MacFadyen, 2002; Kamal et al., 2020; Kaufman et al., 2000).

Until August 2022, the clinical science department at Al Rayyan college of medicine was applying an arbitrary cut off score of 60% as a passing score for OSCE. This decision had always been based on tradition, without taking test content or students' performance into consideration. The need for a process to differentiate well between a student with adequate competencies from those having inadequate competencies had always been observed (Khan et al., 2013). The examinee centered Hofstee method can help us to adjust cut scores for a station according to its difficulty level and accepted number of students unable to pass such a station. (Downing et al., 2006; Dudas & Barone 2014; Hofstee, 1983).

The purpose of this study is to compare the pass /fail rates of students achieved by applying arbitrary and the Hofstee methods and to assess if Hofstee method can provide us satisfactory results.

II. METHODS

The current study is a descriptive study design conducted at Al Rayyan college of medicine department of clinical science. Al Rayan college of medicine, Al Rayan national colleges is a newly established private institute based in Al-Medina Al-Munawara, under Ministry of education at Kingdom of Saudi Arabia. Having been established in 2017, the first batch of graduates have completed MBBS and have joined the local and international health sector. Currently there are 700 students enrolled and studying in 6 academic years.

This study includes a total number of 102 year 4 students undertaking the final OSCEs in the general practice 1 course with foundation to general practice, Cardiovascular system (CVS) and endocrine and breast modules during semester 1 of academic year 2022-23. The project was approved by the Research Ethical Committee (REC) of Al Rayyan colleges. All the students consented to the use of their data for research and quality control purposes with the agreement that any reports would only use aggregate data with all personal identifiers removed.

The OSCE consisted of 10 stations that sampled common and important patient presentations. Examinees were required to complete each station within 07 min. Performance was scored using 10 predefined competencies related to general practice competencies aligned to course learning outcomes CLOs, designed under the umbrella of the competence specifications for Saudi medical graduates (Saudi Meds). Skill competency assessed were (1) history taking, (2) physical examination, (3) analysis and interpretation of findings, (4) communication, (5) suggestion of appropriate investigations, (6) listing relevant differential diagnoses, (7) management care plan development. For values assessment, there were three competencies: (1) ethical rules and confidentiality, (2) taking and maintaining consent, and (3) time management. Three to four of these competencies were assessed in each station except for clinical approach; management stations where only one competency is evaluated.

3 panels were laid down, each having 10 stations and 2 circuits of students. Students rotated through the stations completing a single circuit in an anticlockwise manner. Every student was examined by a single examiner at each station except for the station of data interpretation chest Xray which was just monitored by a silent invigilator and students were recording their answers on answer sheet.

Examiners were all trained faculty staff from department of clinical sciences, 12 examiners belonged to the college faculty, 17 joined from Taibah college of medicine, Taiba university. They received formal training sessions 2 hours ahead of the exam that began with information about the OSCE (fundamentals, competencies being assessed, rating guidelines and cases and question items were explained), followed by instructions on scoring through a google link. Four Hofstee questions were presented, discussed and answered by each examiner for each station and the mean percentage for each of the four questions across all examiners was computed. Meanwhile examiners were asked to answer the same four questions for the overall scores for the exam.

During the OSCE, examiners scored examinee performances within their assigned stations using the 20-26 items scale for each station except for interpretation; chest x ray station which was the only station having 5 item scale. Global ratings (overall assessment from 0 to 5) were also included.

The examiners decided that the cut score for minimally acceptable performance for the whole exam should be no lower than 57.5% and no higher than 76%. Similarly, they indicated that the failure rate should be at least 9% but no higher than 32%.

For cut score calculation, the student's obtained score is plotted with scores along X coordinate and the number of candidates along the Y coordinate. A line graph is drawn showing the score and the number of students obtaining that score. The finally calculated Hofstee limits of cut scores and failure rates are drawn on the graph, which resulted in generation of a rectangle, the cross diagonal from top left of the rectangle to bottom right is drawn. The place where it intercepts the plot of cumulative number of candidates is the cut score for the stations. The graph is shown in figure 1. The same graphs were drawn for the individual stations and their cut scores were calculated. The detail of each station is not mentioned to avoid complexity.

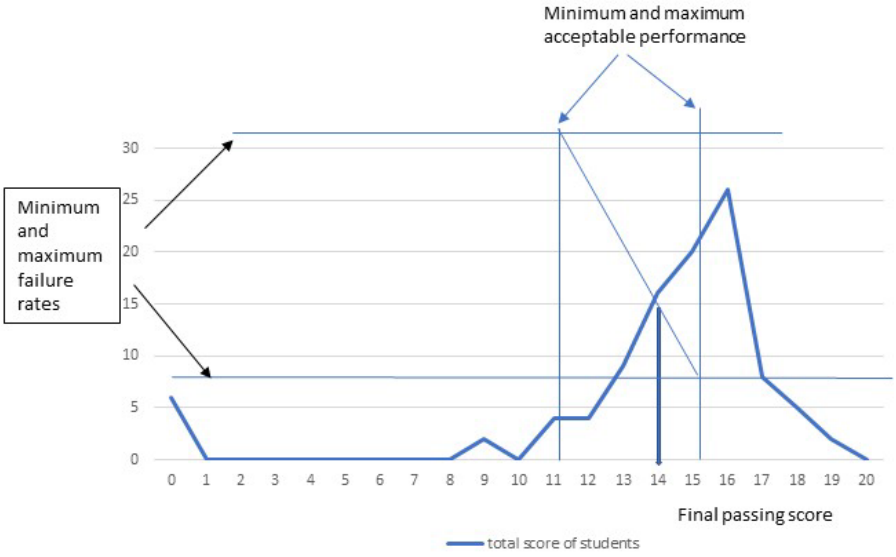

Figure 1. Calculation of final cut score based on examiners provided four Hofstee values

Arbitrary cut score of 60% is applied to students achieved scores and the pass/fail percentage is calculated and recorded in table 2.

III. RESULTS

The OSCE went on without any significant issues. The data provided by the Exam and evaluation Unit (EEU) according to their software showed that the mean score was 75% with a standard deviation of 3.99% and an alpha coefficient of 1.03. Station wise descriptive results show a range in mean scores from 61 to 84%, illustrating a range in station difficulty. The detailed analysis of each station is shown in table 1.

Station	Average %	SD	Variance	Cronbach alpha if item deleted	Coefficient of determination R2	Inter-grade discrimination
Station 1	0.69	0.47	0.22	0.94	0.73	0.03
Station 2	0.84	0.23	0.05	0.74	0.62	0.03
Station 3	0.77	0.21	0.04	0.76	0.46	0.02
Station 4	0.73	0.33	0.11	0.72	0.60	0.03
Station 5	0.73	0.22	0.05	0.67	0.28	0.02
Station 6	0.61	0.56	0.31	0.68	0.62	0.03
Station 7	0.83	0.99	0.08	0.82	0.74	0.02
Station 8	0.78	0.32	0.10	0.83	0.46	0.02
Station 9	0.81	0.21	0.04	0.64	0.45	0.02
Station 10	0.71	0.38	0.14	0.81	0.67	0.03

Table 1. Stations Analysis

The descriptive results for the individual stations with their titles, maximum, minimum and average scores obtained are illustrated in Table 2. The cut scores calculated by arbitrary and hofstee methods are applied to the stations and accordingly pass percentages achieved are shown in Table 2.

Station	Mean %	Minimum marks obtained	Maximum Marks obtained	Cut score Hofstee method (%)	Pass rate according to Hofstee cut score (%)	Cut score arbitrary method (%)	Pass rate according to arbitrary cut score
1. History taking DKA patient	69.1%	16%	95.9%	65	63	60	64.7
2. Clinical examination of a breathless patient	83.31%	37%	97.3	70	85	60	92
3. Clinical examination breast	76.99%	45%	100	70	72	60	89
4. Clinical approach management of hypertension	73.31%	31%	100%	70	53	60	73
5. History taking of a febrile patient	73.36%	50%	92%	70	63	60	86
6. Data interpretation chest Xray	61.17%	11.1%	100%	65	54	60	54
7. Clinical examination abdomen	82.68%	41%	97.5%	65	84	60	86
8. Clinical approach obesity management	78.48%	34.3%	96.8%	65	75	60	81
9. History taking breathlessness	80.80%	45.7%	97.1%	70	80	60	90
10. Examination neck swelling	70.47%	0%	100%	70	52	60	70

Table 2. Station wise descriptive statistics, the two cut scores and students pass rates according to cut scores

The mean score for the station reflects its level of difficulty ranging from 61.17% to 83.3%. The cut score of the individual stations for the Hofstee method was higher than the cut score for the arbitrary method. So is the difference in pass rates, pass rates achieved with arbitrary cut scores are higher than with Hofstee method, as shown in table 3.

Method	Cut scores (%)	Number of students declared Pass	Pass percentage (%)
Arbitrary method	60	88	86
Hofstee method	70	72	70.5

Table 3. Comparison of overall cut scores and pass rates

Using Hofstee method and cut of score of 14 out of 20 passing rates achieved is 72 out of 102 which in percentage makes 70.5%. When compared with arbitrary method and cut score of 12 out of 20, students pass rate increased to 88 out of 102 leading to 86% overall. This study points out a higher pass rate for the students by arbitrary method, which creates a doubt on the competency of passing students.

IV. DISCUSSION

In this research, the results of end semester OSCE exam are compared by two methods, arbitrary fixed 60% standard setting method used at our college for the last 4 years and a compromise Hofstee method, which is applied for the very first time.

According to our study, the failure rate has increased from 13.7% to 29.4%, and has almost doubled. In fact, this increase is higher than what had been usually observed previously. This gives the impression that the students who have not yet achieved the required competency would have been allowed to pass. The same observation was made by Dudas et all who did apply Hofstee standard setting to a historic cohort of 116 Johns Hopkins University School of Medicine students from the academic year 2012–2013 to assess the potential impact on grade distributions (Dudas & Barone 2014).

According to the results of a study conducted by Doaa Kamal in Suez Canal University, Egypt in 2020 where four standard methods, the modified Cohen's, borderline regression, Hofstee methods, and the fixed 60% arbitrary method were compared in determining the passing score in ophthalmology OSCE exam, it was concluded that 60% fixed arbitrary method resulted in a marked difference in failure and pass rates among students and Hofstee method yielded low pass rates which is consistent with my research (Kamal et al., 2020).

Since our exam was dealing with the assessment of multiple competencies, so Hofstee method is more likely to produce a standard appropriate with the purpose of assessment. Secondly the cut scores were calculated by the academic staff who were very much familiar with the OSCE as an assessment tool, the curriculum and the students as well. They were all content experts, fair and open-minded. Some of them were teaching in Taibah university the same content so they were well aware of the acceptable students' performance. The examiners were meeting the criteria set by Downing et al., so their decision regarding the cut scores and estimation of number of failing students was accepted (Downing et al., 2006).

Schindler et al in his research paper has applied Hofstee cut off scores and found that it can even be used for a multi assessment surgical clerkship and for assigning grades as well and concluded that this method has all the characteristics of an appropriate standard setting method (Schindler et al., 2007).

V. CONCLUSION

Since different competencies reflect the different level of difficulties, the cut scores need to be set for each station dealing with that competency. The arbitrary 60 % method is not appropriate to the purpose of an OSCE exam, but a cut off score calculated by using data from experts' judgments provides a reasonable result with acceptable failing rates. Thorough and thoughtful preparation on the judges' part is deemed important. The data gathered from this exam can be reviewed and acted in accordance with to create a standard each academic year.

Notes on Contributors

The author herself contributed to the design of the research, carried out the data acquisition and analysis, interpreted the data and prepared the manuscript.

Ethical Approval

Approval was obtained from the Institutional Research Ethics Committee (IREC) for the collection and publication of student data with approval No. HA-03-M-122-046. Informed consent is taken from the students and special permission from the dean of the college is obtained for the use of students result for this research purpose.

Data Availability

The data generated and analysed in this study are included in this published article. The raw data set analysed can be provided by the author if special interest in it is shown.

Acknowledgement

I would like to thank Dr. Yasir Naseem Khan from the department of basic medical sciences, Al-Rayan college of medicine for useful suggestions and guidance at the final stage of the project.

I would also like to thank the Dean and the Exam and evaluation Unit (EEU) of the Al-Rayan college of medicine for their support.

Funding

The study does not require any funding.

Declaration of Interest

The author declares that she has no competing interests.

References

Boulet, J. R., De Champlain, A. F., & McKinley, D. (2003). Setting defensible performance standards on OSCEs and standardised patient examinations. *Medical Teacher*, *25*(3), 245-249. https://doi.org/10.1080/0142159031000100274

Cusimano, M. D. (1996). Standard setting in medical education. *Academic Medicine*, *71*(10), S112-20. https://doi.org/10.1097/00001888-199610000-00062

Dent, J., Harden, R. M., & Hunt, D. (2021). *A practical guide for Medical Teachers* (6th ed.). Elsevier.

Downing, S. M., Tekian, A., & Yudkowsky, R. (2006). Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine*, *18*(1), 50-57. https://doi.org/10.1207/s15328015tlm1801_11

Dudas, R. A., & Barone, M. (2014). Setting standards to determine core clerkship grades in pediatrics. *Academic Pediatrics*, *14*(3), 294-300. https://doi.org/10.1016/j.acap.2014.01.008

Dwivedi, N., Vijayashankar, N. P., Hansda, M., Ak, D., Nwachukwu, F., Curran, V., & Jillwin, J. (2020). Comparing standard setting methods for objective structured clinical examinations in a Caribbean medical school. *Journal of Medical Education and Curricular Development*, *7*. https://doi.org/10.1177/2382120520981992

Gormley, G. (2011). Summative OSCEs in undergraduate medical education. *The Ulster Medical Journal*, *80*(3), 127.

Harden, R. M., & Gleeson, F. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education*, *13*(1), 39-54. https://doi.org/10.1111/j.1365-2923.1979.tb00918.x

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

Hodges, B. (2003). OSCE! Variations on a theme by Harden. *Medical Education*, *37*(12), 1134-1140. https://doi.org/10.1111/j.1365-2923.2003.01717.x

Hofstee, W. K. (1983). The case for compromise in educational selection and grading. *On Educational Testing*, 109-127.

Humphrey-Murto, S., & MacFadyen, J. C. (2002). Standard setting. *Academic Medicine*, *77*(7), 729-732. https://doi.org/10.1097/00001888-200207000-00019

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, *34*(4), 353-366. https://doi.org/10.1111/j.1745-3984.1997.tb00523.x

Kamal, D., Sallam, M. A., Gouda, E., & Fouad, S. (2020). Is there a "best" method for standard setting in OSCE exams? Comparison between four methods (a cross-sectional descriptive study). *Journal of Medical Education*, *19*(1), Article e106600. https://doi.org/10.5812/jme.106600

Kaufman, D., Mann, K., Muijtjens, A., & Van Der Vleuten, C. P. (2000). A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Academic Medicine*, *75*(3), 267-271. https://doi.org/10.1097/00001888-200003000-00018

Khan, K., Ramachandran, S., Gaunt, K., & Pushkar, P. (2013). The objective structured clinical examination (OSCE): AMEE Guide No. 81. Part I: An historical and theoretical perspective. *Medical Teacher*, *35*(9), e1437-e1446. https://doi.org/10.3109/0142159x.2013.818634

Khan, K., 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

McKinley, D., & Norcini, J. J. (2014). How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher*, *36*(2), 97-110. https://doi.org/10.3109/0142159x.2013.853119

Norcini, J. J., Anderson, M. B., Bollela, V. R., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R. M., Hays, R., Kent, A., Perrott, V., & Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, *33*(3), 206-214. https://doi.org/10.3109/0142159x.2011.551559

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the OSCE: A review of metrics – AMEE Guide No. 49. *Medical Teacher*, *32*(10), 802-811. https://doi.org/10.3109/0142159x.2010.507716

Pugh, D., & Smee, S. (2013). Guidelines for the development of objective structured clinical examination (OSCE) cases. *Ottawa: Medical Council of Canada*. https://doi.org/10.13140/RG.2.1.4622.0003

Schindler, N., Corcoran, J., & DaRosa, D. A. (2007). Description and impact of using a standard-setting method for determining pass/fail scores in a surgery clerkship. *The American Journal of Surgery*, *193*(2), 252-257. https://doi.org/10.1016/j.amjsurg. 2006.07.017

Smee, S., Coetzee, K., Bartman, I., Roy, M., & Monteiro, S. (2022). OSCE standard setting: Three borderline group methods. *Medical Science Educator*, *32*(6), 1439-1445. https://doi.org/10.1007/s40670-022-01667-x

Zayyan, M. (2011). Objective structured clinical examination: The assessment of choice. *Oman Medical Journal*, 219-222. https://doi.org/10.5001/omj.2011.55

*Dr. Uzma Khan
Department of Clinical Sciences,
Al Rayan College of Medicine,
Al Rayan National Colleges
Madina Munawara, Saudi Arabia
Contact: +966542754680
Email: uziik2019@gmail.com, uk.yasser@amc.edu.sa