


Comparing Standard Setting Methods for Objective Structured Clinical Examinations in a Caribbean Medical School

Journal of Medical Education and Curricular Development
Volume 7: 1–10
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2382120520981992



Neelam Rekha Dwivedi¹ , Narasimha Prasad Vijayashankar¹, Manisha Hansda¹, Arun Kumar Dubey¹, Fidelis Nwachukwu¹, Vernon Curran² and Joseph Jillwin¹ 

¹Xavier University School of Medicine, Oranjestad, Aruba. ²Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada.

ABSTRACT

BACKGROUND: OSCE are widely used for assessing clinical skills training in medical schools. Use of traditional pass fail cut off yields wide variations in the results of different cohorts of students. This has led to a growing emphasis on the application of standard setting procedures in OSCEs.

PURPOSE/AIM: The purpose of the study was comparing the utility, feasibility and appropriateness of 4 different standard setting methods with OSCEs at XUSOM.

METHODS: A 15-station OSCE was administered to 173 students over 6 months. Five stations were conducted for each organ system (Respiratory, Gastrointestinal and Cardiovascular). Students were assessed for their clinical skills in 15 stations. Four different standard setting methods were applied and compared with a control (Traditional method) to establish cut off scores for pass/fail decisions.

RESULTS: OSCE checklist scores revealed a Cronbach's alpha of 0.711, demonstrating acceptable level of internal consistency. About 13 of 15 OSCE stations performed well with "Alpha if deleted values" lower than 0.711 emphasizing the reliability of OSCE stations. The traditional standard setting method (cut off score of 70) resulted in highest failure rate. The Modified Angoff Method and Relative methods yielded the lowest failure rates, which were typically less than 10% for each system. Failure rates for the Borderline methods ranged from 28% to 57% across systems.

CONCLUSIONS: In our study, Modified Angoff method and Borderline regression method have shown to be consistently reliable and practically suitable to provide acceptable cut-off score across different organ system. Therefore, an average of Modified Angoff Method and Borderline Regression Method appeared to provide an acceptable cutoff score in OSCE. Further studies, in high-stake clinical examinations, utilizing larger number of judges and OSCE stations are recommended to reinforce the validity of combining multiple methods for standard setting.

KEYWORDS: OSCE, standard setting, relative method, mean borderline group method, borderline regression method, and modified Angoff's method, traditional method

RECEIVED: November 19, 2020. **ACCEPTED:** November 25, 2020.

TYPE: Original Research

FUNDING: The authors have received the publication costs from Xavier University School of Medicine, Aruba.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Neelam Rekha Dwivedi, #23, Santa Helenastraat, Xavier University School of Medicine, Oranjestad, Aruba. Email: neelammd@xusom.com

Introduction

Xavier University School of Medicine (XUSOM) is an off-shore medical school located on the island of Aruba. It offers a 4-year MD program similar to North American medical schools. Year 1 and 2 curriculum consist of basic sciences portion which is organized as an integrated organ-based system with both horizontal and vertical integration. Basic science is taught in a hybrid curriculum, using a combination of didactic lectures, Problem Based Learning (PBL), Team Based Learning (TBL), Clinical case presentations and other self-directed learning (SDL) teaching and learning methodologies.

The clinical skills training begins in the first semester of medical school as part of an "Early Clinical Exposure" course, where students visit local family physicians' and specialists' clinics, and the Government hospital in Aruba. In addition, the

ICMPD (Introduction to Clinical Medicine and Physical diagnosis) course is designed to teach the history taking and physical examination skills, diagnostic reasoning and train the students for OSCE's using standardized patients.

In 2013, XUSOM launched its own "Standardized Patient Program" in Aruba. Since then, the Objective Structured Clinical Examination (OSCE) using standardized patients (SPs), is incorporated in each organ system course, to reinforce teaching and assessment of clinical skills training, and also to prepare students for the United States Medical Licensing Examination Step 2 Clinical Skills examination (USMLE Step 2CS).

The use of the OSCE has grown in importance in medical education and assessment analysts are now labeling it as one of the most rational, effective, and dependable methods for



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

Table 1. Distribution of students across OSCE stations for 2 semesters with the number of students.

SPRING 2019	SUMMER 2019
Respiratory system (33 students)*	Respiratory system (22 students)**
Gastrointestinal system (33 students)*	Gastrointestinal system (20 students)**
Cardiovascular system (35 students)#	Cardiovascular system (30 students)*

*Same group (3 unique). **Same group (2 unique). #Unique.

assessing clinical performances.¹⁻⁵ It is one of the formats of competency-based assessment methods to evaluate not only medical knowledge, but also other core proficiencies, like practice-based learning and communication skills, which all are related to effective patient care.^{2,6-8}

The OSCE, with reference to Miller's pyramid of assessment (Miller 1990) assess that a candidate is able to "show how" one would perform in simulated settings.⁸ SP-based multiple station OSCEs are now a part of several high stakes examinations, including the Canadian Medical Council of Canada (MCC) qualifying examination and an examination for international medical graduates wishing to practice in Canada.^{9,10} The National Board of Medical Examinations (NBME) also uses OSCE in the U.S. licensing examinations. Use of OSCEs in such high-stake examinations emphasizes the need of standard setting procedures to accurately assess the examinees during their training in medical schools.

Standard setting methods are broadly categorized into norm-referenced or relative method and criterion-referenced method or absolute method. Norm-referenced or relative methods identify the cut-off score relative to performance of the group or top scoring examinees taking the examination.¹¹⁻¹⁴ Criterion-based or absolute methods identify cutoff scores based on the level of competence expected of students on the content being examined and are, thus, preferred for competence-based assessments like OSCEs.^{15,16} This can further be categorized into test/examination-centered (eg, Angoff) methods and examinee-centered methods (eg, borderline group [BLG] and borderline regression).¹⁷ With the requirement for standards to be defensible, evidenced and acceptable, absolute standards are generally preferred.¹⁸ Medical schools commonly use examinee-centered standard setting methods such as Mean borderline group (BG) method and Borderline regression method (BLR).^{13,19,20} An alternative method is test item centered, Modified Angoff standard setting method. In this method, the pass mark is based on item or station characteristics and varies according to the difficulty level of the station determined by the characteristics of the items on checklist rather than the examinees performance.²¹

Until September 2018, XUSOM Aruba was applying an arbitrary cut off score of 70% as a passing score for OSCE. This decision was based on tradition, rather than on test content or examinee's performance. It was difficult to provide a defensible explanation of how this 70% passing standard was

set. Also, large variations in the performance of different cohorts, at XUSOM, were noticed while using the arbitrary 70% score as pass-fail criteria. The present study is proposed to address this problem by comparing 4 different standard setting methods with our traditional method, analyze the data and determine the method or combination of methods that would be most appropriate for the assessment of OSCE at XUSOM, Aruba.

The purpose of this study was to compare the utility, feasibility and appropriateness of different standard-setting methods for setting pass-fail cut-off scores for internal OSCE examinations.

Methods

The current study is a descriptive study design conducted at XUSOM, Aruba. Basic Sciences students (year 1 and year 2) undertaking the final OSCEs in the Respiratory system (RS), Gastrointestinal system (GIS) and Cardiovascular system (CVS), during Spring and Summer 2019 semesters, participated in this study. In Spring 2019 semester, 33 students undertook the final Respiratory system OSCE, 33 students the GIS OSCE and 35 students the CVS OSCE. In Summer 2019, 22 students completed the Respiratory system OSCE, 20 students the GIS OSCE and 30 students the CVS OSCE (Table 1). Thus, a total of 173 students participated in this study. RS and GIS are taught to 3rd semester (Year 1) students separately at different times and CVS is taught to 4th semester (Year 2) students, therefore same set of students were administered with RS and GIS OSCE in Spring 2019 at different times, and these students were administered with CVS OSCE in Summer 2019. The students of 3rd semester (year 1) were administered with RS OSCE and GIS OSCE at different times during Summer 2019. The students for CVS OSCE in their 4th semester (Year 2) during Spring 2019 were unique (Table 1). Five full time faculty members with a minimum MD qualification were involved in the study. The project was approved by the IRB (Institutional Review Board) of XUSOM, Aruba. Informed consent and confidentiality of information about participants was obtained for the study.

A 15-station OSCE, using standardized patients, was administered to 173 students over 6 months. 5 stations each were held per system in a single circuit (total of 15 stations for 3 organ system) as shown below (Figure 1). Students rotated

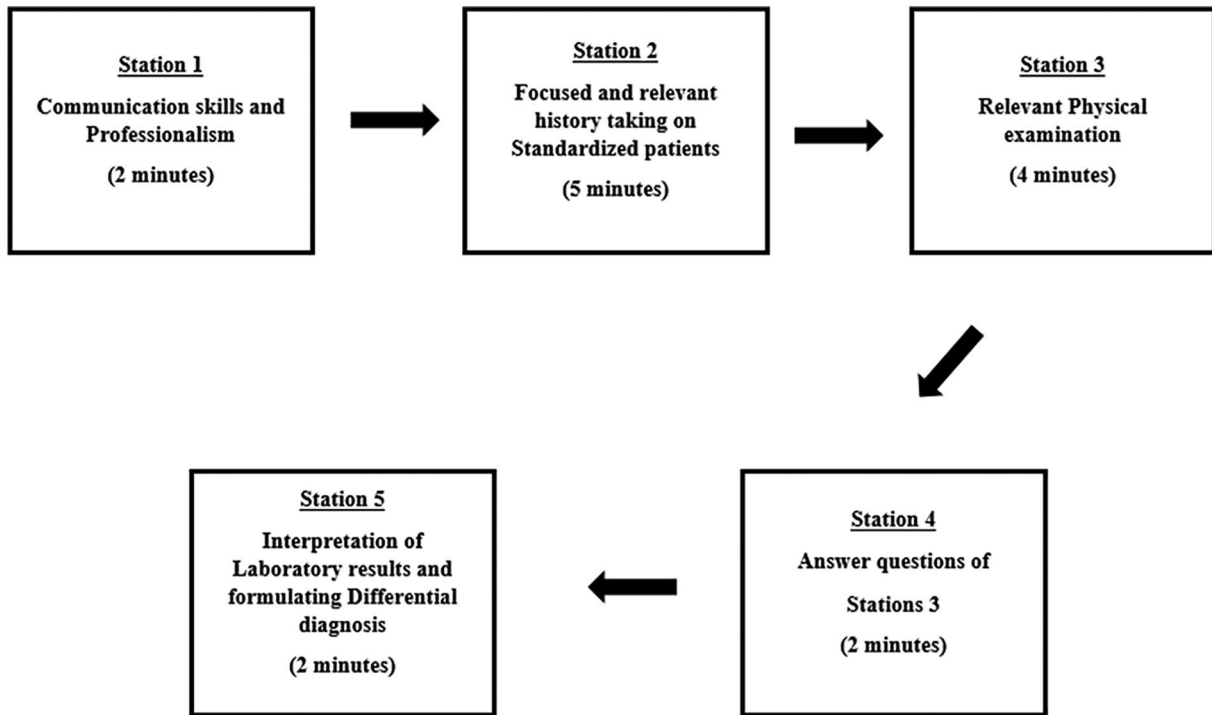


Figure 1. OSCE circuit.

through the stations completing single circuit. Every student was examined on one-on-one basis by a single examiner.

An established blueprinting process was followed to ensure stations assessed a variety of appropriate domains of clinical skills such as knowledge, psychomotor and affective. Each station was of a 2-minute duration to assess communication skills and professionalism, 5 minutes for history taking, 6 minutes for physical examination skills and 2 minutes for interpretation of laboratory results and reaching most likely diagnosis. Each station was reviewed and validated by the faculty members prior to OSCE administration. The students were properly instructed, prior to the OSCE session, regarding the information related to the presenting problem, the task and time frame for completing the encounter and strict policy to be followed.²² One full-time faculty member with several years' experience of teaching ICMPD and administering OSCEs recruited other members of the study gradually over a period of time. To ensure consistency and fairness of scores, all the faculty involved were trained gradually via workshops for conducting OSCEs during which they were clearly informed about the objectives, outcomes, roles and responsibilities and were allowed to shadow and observe. All the faculty members were involved in training the students for OSCE, designing the stations, creating rubrics for the OSCE stations based on the blueprint, in training of SPs and grading the student's performance during OSCE. The Station checklists were reviewed and validated by all members of faculty involved in the study. All SPs were well trained for their consistent role to ensure that each student is presented with the same challenge.²³ The data analysis was done using ANOVA and t-test by using SPSS 20 version. The

effect size was calculated using Cohen's "d" values for the pooled standard deviation. The effect size was deemed Large if $d > 0.8$, Medium if $d > 0.5$, and Small if $d > 0.2$.

Standard setting methods

Traditional method. Each examiner graded individual student performance using a station checklist comprising varying number of items per checklist based on the task assessed at each station with a minimum number of 5 items per checklist and a maximum number of 25 items per checklist. Based on the performance of the student, the score was converted to percentage. For each system, 5 OSCE checklists were used and checklist item criteria for communication skills and professionalism were common across all systems. The items for other clinical skills varied according to the clinical condition and organ system. Each student's "System Total Score" was calculated by adding the 5 station checklist scores and dividing by 5 to produce an average system score for the student. Traditionally the arbitrary cut-off passing score of 70% had been applied to determine pass-fail.

Modified borderline group method (MBGM). After completing the station checklists, the examiners also provided a global Likert rating (from 1 to 5) for every individual student, based on overall impression of the student's performance, independent of the checklist score. Likert ratings included: 1=Clear Fail, 2=Borderline Fail, 3=Borderline Pass, 4=Clear Pass, and 5=Excellent Performance. A linear regression analysis was conducted to ascertain whether station global Likert ratings

Table 2. Summary of traditional method standard setting approach.

SYSTEM TOTAL	NUMBER OF STUDENTS (N)	MEAN	STANDARD DEVIATION	95% CONFIDENCE INTERVAL	NUMBER OF FAILURES (BELOW 70)	% OF FAILURES (BELOW 70)
CVS	65	64.40	7.13	63 to 67	54 out of 65	83.1%
GIS	53	69.63	8.26	67 to 72	27 out of 53	50.9%
RS	55	67.26	6.10	66 to 69	35 out of 55	63.6%

correlated with the station checklist score. After establishing the relationship between Likert ratings and checklist scores, students obtaining Likert ratings between 2 (Borderline Unsatisfactory) and 3 (Borderline Satisfactory) at each station, were selected, to calculate a mean checklist score for the “Borderline Group” at each station.

Borderline regression method (BLR). This method was conducted to predict checklist scores for a Borderline student. We used 2 borderline categories, 2 = Borderline Unsatisfactory and 3 = Borderline Satisfactory, with a mean Likert rating of 2.5 used as Borderline. For each of the 15 stations, we used a linear regression model in which the student’s checklist scores and global Likert rating scores were considered as dependent and independent variables, respectively. Then we calculated the checklist score cut-off on the regression equation for the global Likert rating cut-off set at 2.5.^{21,24,25}

Modified Angoff method (MAM). In this method, individual OSCE stations were scored by 6 selected Chairs of the clinical departments at XUSOM who also acted as judges. Each of the clinical chairs had more than 15 years of experience in teaching and assessing the students of year 3 and year 4 of the MD program and training the residents at their clinical departments. All the Clinical chairs had extensive expertise in conducting OSCEs, they were briefed about the purpose and steps of standard setting processes, followed by a brief discussion on qualities of a borderline (minimally competent) student. Each judge was asked to determine the probability of a borderline (minimally competent) student to perform the test item in each station correctly in the percentage from 0 to 100. Following the individual ratings, the judges displayed their ratings, then discussed the reasoning behind any discrepancies. Following the discussion, each judge again rated each station answering the same question. The judges’ estimates were averaged for each station and the mean of the averages was used as cut off scores.²⁶

Relative method. Kaufman et al¹³ used the best student’s performance as reference point, with such students generally well prepared for the examination and fluctuations in these students’ scores reflecting variations in exam difficulty. This method used the score that ranks at the 95th percentile and defined “passing” as a score that is equal to 60% of the 95th percentile. Similar methodologies of standard settings are shown to be

more practical, and overcome certain disadvantages of criterion and norm-based methods.²⁷ But this definition resulted in zero failures in our study. We modified Kaufman’s method to define “passing” as the score that is equal to 70% of the 95th percentile score of the best students, which is consistent with our traditional cutoff standard of 70%.

Results

Standard setting methods

Traditional method. Table 2 summarizes the system total mean scores and percentage of students failing each system using the traditional method standard setting approach. The CVS and RS mean scores were significantly below the traditional standard of 70 ($P < .05$), and the CVS mean score was significantly lower than the GIS mean ($P < .05$).

(CI = Mean score \pm 1.96 * Standard error of the mean for 95% CI)

Standard error = S.D./ \sqrt{N}

Modified borderline group method (MBGM). Pell et al²⁸ determined an acceptable R^2 value to be above 0.50, and all values in our data, were above 0.50, indicating a strong linear relationship between the Likert rating and the checklist score for each station (Table 3). In our study, the R^2 values are higher indicating that the checklist scores and the global scores show a strong positive correlation.

Figures 2 to 4 depicting the average scores of the students for that particular system, they show consistent scatter plots as the individual station scores.

Table 4 shows the number of students identified as Borderline for each station, along with the Borderline Groups’ station checklist mean score. For each System OSCE, the sum of the 5 stations check list mean score was divided by 5 to determine the cut-off score. Failure rate was defined as the number/percent of students with a system total score lower than the cut-off score.

Borderline regression method (BLR). Table 5 summarizes the predicted checklist score for a 2.5 global Likert rating at each station. For each system, the 5 predicted checklist scores were summed and divided by 5 to calculate a system total cut-off score. Failure rate was defined as the number/percent of students with a system total score lower than the

Table 3. Borderline methods: R^2 coefficient of determination.

STATION	R^2
1	0.97
2	0.58
3	0.78
4	0.92
5	0.94
6	0.97
7	0.88
8	0.93
9	0.91
10	0.97
11	0.98
12	0.77
13	0.87
14	0.86
15	0.95

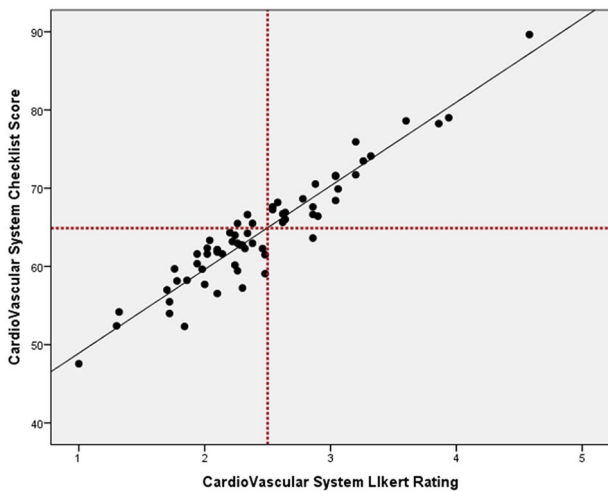


Figure 2. Inter-grade discrimination for cardiovascular system.

cut-off score. Figures 2 to 4 represent the results of the Borderline Regression Method for each station by determining intergrade discrimination and indicating the relationship between checklist score and global Likert rating score on the slope of the regression line. Scores that fall along the slope of the regression line indicates checklist scores correlate well with the scores of Likert rating scale, which in turn reflects the validity of final score. Slopes of regression for Cardiovascular, Gastrointestinal and Respiratory systems indicate an average increase in checklist scores (y axis) corresponding to an increase of 1 grade on the global Likert rating scales (x axis).²⁵

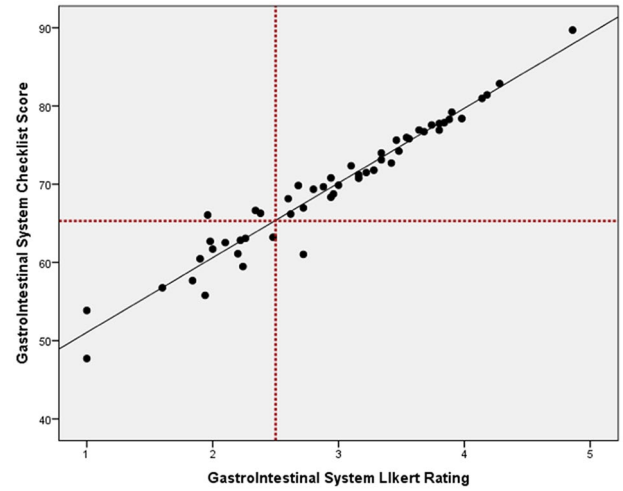


Figure 3. Inter-grade discrimination for gastrointestinal system.

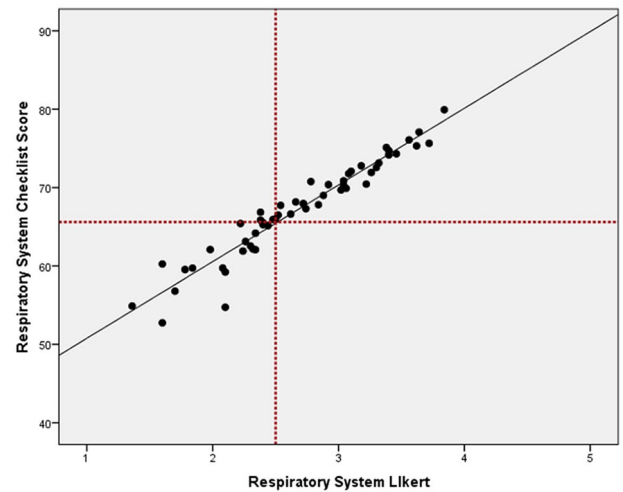


Figure 4. Inter-grade discrimination for respiratory system.

Modified Angoff method (MAM). Table 6 presents the MAM mean and standard deviation for each system, along with the failure rates, defined as number/percent of students with a system total score lower than the MAM cut-off score for the system.

Relative method. Table 7 represents the 95th percentile for each system total score, along with the score that equals 70% of the 95th percentile and failure rates, defined as the number/percent of students who had a System Total score less than 70% of the 95th Percentile.

Before analyzing the data to compare the different standard setting methods, the internal consistency of the OSCE across all the stations was assessed for all 173 students. The acceptable value of Cronbach’s alpha is 0.70 or above.²⁸ In our study, Cronbach’s alpha was 0.711, which is above the 0.70 acceptable value. This suggests that the checklist scores are internally consistent. As shown in Table 8, the “Alpha If Item Deleted” column estimates what Cronbach’s Alpha would be if we removed a specific station score. If a station’s “Alpha if deleted” value is lower than the total 0.711, it suggests reliability would decrease

Table 4. Modified borderline group method (MBGM): station means, standard deviation, and failure rate.

STATION	NUMBER OF STUDENTS WITH LIKERT 2 TO 3	CHECKLIST MEAN SCORE	STANDARD DEVIATION	NUMBER OF FAILURES	% OF FAILURES
Cardiovascular system (CVS)					
1	30	65.83			
2	19	63.65			
3	31	62.31			
4	32	66.28			
5	41	64.65			
System total*		64.54	1.62	37 out of 65	56.9%
Gastrointestinal system (GI)					
6	17	65.55			
7	21	67.55			
8	14	64.13			
9	19	70**			
10	17	64.42			
System total*		66.3	2.45	18 out of 53	34%
Respiratory system (RS)					
11	23	64.44			
12	32	65.39			
13	30	65.47			
14	25	69.16			
15	30	65.86			
System total*		66.06	1.81	23 out of 55	41.8%

*Total mean score of respective organ system OSCE. **The Station 9 distribution of scores for the Borderline sample was severely negatively skewed. Therefore, the median (not the mean) is the most appropriate measure of central tendency and the median is reported instead of the Mean.

if we removed that station. In other words, any station with an “Alpha if Deleted” value lower than our total 0.711 is a station that strengthens our reliability because removing that station would weaken internal consistency.

As shown in Table 8, 13 of the 15 stations performed well, with “Alpha if Deleted” values lower than 0.711. On the other hand, if we deleted Stations 4 and 11, Cronbach’s Alpha would increase above 0.711, which means that including Stations 4 and 11 weakens our internal consistency. However, the “Alpha if Deleted” values for stations 4 and 11 are not drastically above 0.711, so they do not severely detract from the reliability of checklist score.

The effect size was calculated using Cohen’s “d” values for the pooled standard deviation of 5.57. A One-sample t-test shows that the CVS mean ($M=64.40$, $SD=7.13$) was significantly below the traditional standard of 70, $t(64)=-6.33$, $P<.001$. The effect size was “large” $d=0.84$. Likewise, a One-sample t-test shows that the RS mean ($M=67.2$, $SD=6.10$)

was significantly below the traditional standard of 70, $t(54)=-3.32$, $P<.002$. The effect size was “medium” $d=0.59$. The GIS mean ($M=69.6$, $SD=8.26$) was not different than the traditional standard of 70, $t(52)=-.31$, $P>.02$. The effect size was “small” $d=0.05$. For these multiple comparisons, alpha was adjusted from .05 to .02 to reduce the risk of a Type 1 error.

Comparison of system checklist means. A One-Way Between-Subjects ANOVA was conducted to compare the system mean scores, and the results were significant, $F(2, 170)=7.80$, $P<.001$. Bonferroni post-hoc comparisons show that the CVS mean ($M=64.4$, $SD=7.13$) is significantly lower than the GIS mean ($M=69.6$, $SD=8.26$). We are 95% Confident that the CVS mean is between 2 to 8.5 points lower than the GIS mean. The RS mean was are not different than either CVS or GIS.

Table 9 summarizes the cut-off score and % of failed students for each system, using each of the five standard setting methods.

Table 5. Borderline regression method (BLR): regression equation, borderline score, and failure rate.

STATION	REGRESSION EQUATION ($Y=A + BX$)	BORDERLINE (2.5) SCORE	NUMBER OF FAILURES	% OF FAILURES
Cardiovascular system (CVS)				
1	$9.396X + 42.23$	65.72		
2	$9.777X + 40.49$	64.92		
3	$9.821X + 39.122$	63.67		
4	$10.115X + 0.309$	65.60		
5	$10.865X + 7.269$	64.43		
System total*		64.87	37 out of 65	56.9%
Gastrointestinal system (GIS)				
6	$9.651X + 41.299$	65.43		
7	$8.822X + 43.871$	65.93		
8	$9.308X + 41.541$	64.81		
9	$9.365X + 42.089$	65.50		
10	$9.914X + 40.118$	64.90		
System total*		65.31	15 out of 53	28.3%
Respiratory system (RS)				
11	$9.604X + 41.525$	65.54		
12	$8.886X + 43.321$	65.54		
13	$9.368X + 41.199$	64.62		
14	$8.212X + 46.938$	67.47		
15	$9.867X + 40.225$	64.89		
System total*		65.61	20 out of 55	36.4%

*Mean checklist score of borderline students in respective organ system OSCE.

Table 6. Summary of judges' estimates using MAM method.

SYSTEM	MEAN	STANDARD DEVIATION	NUMBER OF FAILURES	% OF FAILURES
Cardio vascular	55.88	3.85	6 out of 65	9.2%
Gastro intestinal	59.74	4.58	6 out of 53	11.3%
Respiratory	57.91	2.67	4 out of 55	7.3%

Discussion

The results indicate that the traditional arbitrary score of 70 had the highest failure rate, with the majority of students failing all 3 organ systems. The MAM and Relative methods yielded the lowest failure rates, which were typically less than 10% for each system. Failure rates for the Borderline methods ranged from 28 to 57% across the systems. Mean scores for the CVS was the lowest, and mean scores of GIS was relatively high. One of the reasons that CVS scores were low could be

attributed to the CVS station checklists being more challenging with more items assessing clinical reasoning skills. On the contrary, high GIS scores could be because the items in the checklist of the stations in GIS were uniformly process based requiring relatively less clinical reasoning. Therefore, to correct this disparity we proposed that the checklists of these 2 systems be reviewed and be comparable and standardized with regards to communications skills items and clinical reasoning items in future. We also proposed that items of CVS checklist be made

Table 7. Relative method: percentile scores and failure rate.

SYSTEM	SYSTEM TOTAL SCORE THAT RANKS AT 95TH PERCENTILE	70% OF 95TH PERCENTILE	NUMBER OF FAILURES	% OF FAILURES
Cardio vascular	75.92	53	3 out of 65	4.6%
Gastro intestinal	80.97	57	4 out of 53	7.5%
Respiratory	75.64	53	1 out of 55	1.8%

Table 8. Cronbach's alpha if deleted data.

STATION	ALPHA IF DELETED
1	0.705
2	0.698
3	0.687
4	0.734*
5	0.694
6	0.702
7	0.677
8	0.673
9	0.677
10	0.662
11	0.716*
12	0.709
13	0.697
14	0.705
15	0.695

* "Alpha if deleted" value above the value of Cronbach's alpha (0.711).

more liberal encompassing more items to test the basic aspects of history taking in CVS rather than items assessing advanced clinical judgement, and to incorporate more items that assess the clinical reasoning of examinee in the checklists of GIS.

Kaufman et al¹³ reported that Angoff and borderline methods were shown to provide a reasonable and defensible approach to standard setting and were of practical value when used by non-psychometricians in medical schools.²⁶ In contrast, Kramer et al examined standard setting in postgraduate general practice training, identifying that the Borderline Regression Method (BRM) was more credible and acceptable than the modified Angoff method (MAM). Kramer et al²⁹ used 84 examiners and a significant number of them also performed the modified Angoff method. These conflicting findings may be the results of some known difficulties with the Angoff method. Verheggen et al³⁰ demonstrated considerable variation between judges, especially when judges had less expertise in certain item areas.

In our study, though we used only 6 clinical chairs as judges compared to larger numbers used in other similar studies, the

Table 9. Summary: comparison of standard setting methods.

	CUT-OFF SCORE	% OF FAILED STUDENTS
Traditional method		
CVS	64.4	83.1
GIS	69.3	50.9
RS	67.6	63.6
Relative method		
CVS	53.0	4.6
GIS	57.0	7.5
RS	53.0	1.8
Modified Angoff method		
CVS	55.9	9.2
GIS	59.7	11.3
RS	57.9	7.3
Borderline group method		
CVS	64.5	56.9
GIS	66.3	34.0
RS	66.1	41.8
Borderline regression method		
CVS	64.9	56.9
GIS	65.3	28.3
RS	65.6	36.4

results of the modified Angoff method (MAM) appeared to be credible and acceptable. The judge's clinical experience, subject mastery and involvement in supervising OSCEs might have led them to correctly answer all the items in the checklists used to determine the scores of a borderline student. This may be the reason, which explains achievement of an acceptable cut off score. In a study by Dwyer et al²¹ similar observation was made about modified Angoff method (MAM) to set acceptable and credible cut-scores compared to the borderline method and Borderline Regression Method (BRM).

Boursicot et al²⁰ observed that the Borderline method is more consistent in determining the pass score than the Angoff

method. Wood et al²⁴ reported that a cut score derived from a Borderline Regression Method (BRM) was more accurate than 1 derived using the modified borderline group method, supporting the finding of our study as well. Hejri et al²⁵ also reported that Borderline Regression Method (BRM) is much more convenient and less resource consuming compared to other procedures like Angoff. Also, BRM has the advantage of generating a number of indices that are useful in measuring the validity of the OSCE. Considering the fact that BRM is widely used as a standard setting method, assessing its reliability is of paramount importance.

In our study, both Modified Angoff method (MAM) and Borderline Regression Method (BRM) were shown to be reliable by which there were consistently similar cut-off scores across different organ systems, thereby providing way to decide acceptable cut off scores. The pass/fail standard can be reliably set before the OSCE by using Modified Angoff method (MAM), which would be further useful in the setting of competency based medical education (CBME).³¹⁻³³ Borderline Regression Method (BRM) provides objective analysis, and statistical approach to decide about accurate cut off score and to ensure validity of the OSCE by assessing degree of correlation (R^2) between the checklist score and the overall global rating score.

Kaufman et al¹³ showed that the relative and traditional methods gave inconsistent results which was similar to the findings of our study. The findings of our study discourage the use of absolute / traditional method and relative method to determine pass-fail cut off score. Further, according to our study modified borderline group method (MBGM) seems not to be reliable when applied to a small scale OSCEs. It appears that none of the methods of standard setting are perfect when used alone. The standard setting results in our study have either a very high failure rate or very low cut-off score. This creates a disparity in the assessment of the borderline group of students, whose result depends on the cut-off scores. If the cut-off scores are too lenient then it gives advantage to the weak students who have just followed the process scale (communication skills), but did not have any clinical intuition in approaching a case. If the mean cut-off scores are too strict, then it becomes a disadvantage to the set of students who completed a difficult station.

Therefore, to minimize this disparity we propose, to use a combination of standard setting methods by combining 2 methods that could establish a reliable cut-off score and determine an acceptable percent of students failing an examination. It was determined that a combination of BRM and MAM could be practical. In MAM, subject experts determine the cut off score relying on their subjective and professional judgment based on the characteristics of test items/station. In BRM, the cutoff scores are determined by the assessment of examinee's actual performance by expert examiners thereby providing an objective and statistical approach to determine an acceptable cut off score of borderline students. Also, in our study analysis

of R^2 coefficient and intergrade discrimination values obtained by BRM ensures the quality of the overall OSCE and consistency of the examiner grading across all stations during OSCE.²⁷ The MAM average across all 3 systems is 58 and the BRM average is 62, so the combined average is 60. This gives us acceptable and reliable minimum score that a student has to achieve, to be able to pass the OSCE and also helps us differentiate a borderline pass vs. a borderline fail. Yousef et al³⁴ similarly observed that combining and averaging 2 standard settings, Angoff and Hofstee methods yielded a desirable higher cutoff passing scores than the fixed arbitrary passing score of 60% that was used in their school.

Confounding factors creating disparity in the scores such as knowledge, attitude and practices of the Year 1 and Year 2 students toward applying concepts of clinical reasoning in OSCE's, which may require revisions in other phases / steps of the examination cycle.

Conclusion

None of the standard setting methods, when used alone was of pragmatic value to determine an acceptable and reliable cut off score. However, using a combination of Modified Angoff method (MAM) with Borderline Regression Method (BRM) seems to produce a reliable and valid determination of cut off score. Further studies, in high-stake clinical examinations, utilizing larger number of judges and OSCE stations are recommended to reinforce the validity of combination of multiple methods for standard settings.

ORCID iDs

Neelam Rekha Dwivedi  <https://orcid.org/0000-0003-2903-8195>

Joseph Jillwin  <https://orcid.org/0000-0002-2592-6761>

REFERENCES

1. Ataro G. Methods, methodological challenges and lesson learned from phenomenological study about OSCE experience: overview of paradigm-driven qualitative approach in medical education. *Ann Med Surg (Lond)*. 2019;49:19-23.
2. Abdelaziz A, Hany M, Atwa H, Talaat W, Hosny S. Development, implementation, and evaluation of an integrated multidisciplinary Objective Structured Clinical Examination (OSCE) in primary health care settings within limited resources. *Med Teach*. 2016;38:272-279.
3. Malau-Aduli BS, Teague PA, D'Souza K, et al. A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools. *Med Teach*. 2017;39:1261-1267.
4. Majumder MAA, Kumar A, Krishnamurthy K, Ojeh N, Adams OP, Sa B. An evaluative study of objective structured clinical examination (OSCE): students and examiners perspectives. *Adv Med Educ Pract*. 2019;10:387-397.
5. Cömert M, Zill JM, Christalle E, Dirmaier J, Härter M, Scholl I. Assessing communication skills of medical students in objective structured clinical examinations (OSCE)—a systematic review of rating scales. *PLoS One*. 2016;11:e0152717.
6. Ananthkrishnan N. Objective structured clinical/practical examination (OSCE/OSPE). *J Postgrad Med*. 1993;39:82-84.
7. Bakhsh TM, Sibiany AM, Al-Mashat FM, Meccawy AA, Al-Thubaity FK. Comparison of students' performance in the traditional oral clinical examination and the objective structured clinical examination. *Saudi Med J*. 2009;30:555-557.
8. Kamran ZK, Sankaranarayanan R, Kathryn G, Piyush P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. *Med Teach*. 2013;35:e1437-e1446.

9. Reznick R, Smee S, Rothman A, et al. An objective structured clinical examination for the licentiate: report of the pilot project of the Medical Council of Canada. *Acad Med.* 1992;67:487-494.
10. Vu NV, Barrows HS. Use of standardized patients in assessments: recent developments and measurement findings. *Educ Res.* 1994;23:23-30.
11. Schoonheim-Klein M, Muijtjens A, Habets L, et al. On the reliability of a dental OSCE, using SEM: effect of different days [published correction appears in *Eur J Dent Educ.* 2008;12:252. Muijtjens A [corrected to Muijtjens A]]. *Eur J Dent Educ.* 2008;12:131-137.
12. Cizek GJ, Bunch MB. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* SAGE; 2007.
13. Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Acad Med.* 2000;75:267-271.
14. Zieky MJ, Perie M, Livingston SA. *Cutscores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Educational Testing Service; 2008.
15. Norcini JJ Jr. Standards and reliability in evaluation: when rules of thumb don't apply. *Acad Med.* 1999;74:1088-1090.
16. Turnbull JM. What is . . . normative versus criterion-referenced assessment. *Med Teach.* 1989;11:145-150.
17. Naveed Y, Claudio V, Rukhsana WZ. Standard setting methods for pass/fail decisions on high-stakes objective structured clinical examinations: a validity study. *Teach Learn Med.* 2015;27:280-291.
18. Norcini JJ. Setting standards on educational tests. *Med Educ.* 2003;37:464-469.
19. Kilminster S, Roberts T. Standard setting for OSCEs: trial of borderline approach. *Adv Health Sci Educ Theory Pract.* 2004;9:201-209.
20. Boursicot KA, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Med Educ.* 2007;41:1024-1031.
21. Dwyer T, Wright S, Kulasegaram KM, et al. How to set the bar in competency-based medical education: standard setting after an Objective Structured Clinical Examination (OSCE). *BMC Med Educ.* 2016;16:1.
22. Pugh D, Smee S. *Guidelines for the Development of Objective Structured Clinical Examination (OSCE) Cases.* Medical Council of Canada. <https://mcc.ca/media/OSCE-Booklet-2014>
23. Sim JH, Abdul Aziz YF, Mansor A, Vijayanathan A, Foong CC, Vadivelu J. Students' performance in the different clinical skills assessed in OSCE: what does it reveal? *Med Educ Online.* 2015;20:26185.
24. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. *Adv Health Sci Educ Theory Pract.* 2006;11:115-122.
25. Hejri SM, Jalili M, Muijtjens AM, Van Der Vleuten CP. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *J Res Med Sci.* 2013;18:887-891.
26. Ahmed MA, Ounsa GE, Mohamed EY, Taha E. Assessment of MBBS medical students using (OSCE), comparing the modified borderline group method (MBGM), the modified Angoff method (MAM) and the holistic method of 50%. *Int J Adv Res Innov Ideas Educ.* 2016;2:396-400.
27. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. *Med Teach.* 2010;32:154-160.
28. Pell G, Fuller R, Homer M, Roberts T International Association for Medical Education. How to measure the quality of the OSCE: a review of metrics—AMEE guide no. 49. *Med Teach.* 2010;32:802-811.
29. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations [published correction appears in *Med Educ.* 2003;37:574]. *Med Educ.* 2003;37:132-139.
30. Verheggen MM, Muijtjens AM, Van Os J, Schuwirth LW. Is an Angoff standard an indication of minimal competence of examinees or of judges? *Adv Health Sci Educ Theory Pract.* 2008;13:203-211.
31. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Velden U. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *Eur J Dent Educ.* 2009;13:162-171.
32. Norcini JJ, Shea J. The reproducibility of standards over groups and occasions. *Appl Meas Educ.* 1992;5:63-72.
33. Norcini JJ, Shea J. The credibility and comparability of standards. *Appl Meas Educ.* 1997;10:39-59.
34. Yousef MK, Alshawwa L, Tekian A, Park YS. Challenging the arbitrary cutoff score of 60%: standard setting evidence from preclinical Operative Dentistry course. *Med Teach.* 2017;39:S75-S79.