Original Article

Standard setting methods in objective structured clinical examination (OSCE): A comparative study of five methods

Reshma Mohamed Ansari[1*], Norhafizah Ab Manan[2], Nur Ain Mahat[3], Norfaizatul Shalida Omar[4], Atikah Abdul Latiff[5], Sara Idris[5], Azli Shahril Othman[6]

[1]Medical Education Department International Medical School Management & Science University, Selangor Malaysia
[2]Department of Public Health Faculty of Medicine University of Cyberjaya, Selangor Malaysia
[3]Kulliyah of Nursing International Islamic University, Kuantan, Pahang, Malaysia.
[4]Department of Biochemistry Manipal University College Malaysia, Malacca Malaysia
[5]Department of Anatomy Faculty of Medicine University of Cyberjaya, Selangor Malaysia
[6]Department of Pharmacology Faculty of Medicine and Defence Health, National Defence University Malaysia, Kuala Lumpur Malaysia

Article info

Article history:
Received 31 Jan. 2024
Accepted 20 Aug. 2024
Published 14 Dec. 2024

***Corresponding author:**
Reshma Mohamed Ansari, Medical Education Department International Medical School Management & Science University, Selangor, Malaysia.
Email: reshmaansari77@gmail.com

How to cite this article:
Mohamed Ansari R, Ab Manan N, Mahat NA, Shalida Omar N, Abdul Latiff A, Idris S, Shahril Othman A. Standard setting methods in objective structured clinical examination (OSCE): A comparative study of five methods. J Med Edu Dev. 2024; 17(55): 87-96.

Abstract

Background & Objective: Objective Structured Clinical Examination (OSCE) is a crucial component in medical school examinations to assess students' competency, particularly in clinical skills incorporating cognitive and affective domains. OSCE results are subjected to standard-setting methods, which yield different findings. Hence, in this study, five different standard-setting methods, namely norm reference, Angoff method, borderline group method (BGM), borderline regression method (BRM), and modified Cohen's method, were compared to determine the cut-off scores and failure rates determined by each method.

Material & Methods: Data of 170 second-year medical students who attended OSCE with eight stations for their First Professional Examination at the end of year 2 MBBS was taken for the study following ethical approval. Total scores for each station were standardized to 20 marks, and cut-off scores were determined using each of the five standard-setting methods.

Results: As a comparison of 5 methods, the Norm reference method yielded the highest number of stations with high cut-off scores, followed by BRM. This is reflected in the number of failures, too. On the contrary, using the Angoff method yielded the lowest cut-off scores in maximum stations, resulting in the least number of failed students. The Cochrane's Q test of the results yielded a $p < 0.001$, which signifies that the proportion of students who failed a particular OSCE station was significantly different when different methods were used to determine the cut score.

Conclusion: The study, which compared 5 common standard-setting methods employed in medical education assessments, found that norm-referenced and BRM had high cut-off scores and failures, with the opposite determined by the Modified Angoff method. The study concluded that the cut-off score and failure rate differed with different standard-setting methods, and the choice of the method is contextual depending on the available resources.

Keywords: standard setting, OSCE, Angoff, norm-reference, borderline group, borderline regression, Cohen

Introduction

Medical schools have continually emphasized the importance of Objective Structured Clinical Examinations (OSCEs) in assessments to ensure competency and patient safety (1). Medical educationists have named OSCE one of the most reliable, practical, and effective ways of assessing competency in all three domains: knowledge, clinical skills, and affective domain. This is because OSCE tests practical skills and the underlying cognitive and affective domains (2). Though deemed high stakes, the use of OSCE in medical schools is very diverse, necessitating standard-setting practices to ensure the validity and reliability of this examination method (3).

The standard setting determines the remarkable score that indicates the cut-off point differentiating competent students from the lesser competent ones (3). Hence,

selecting the most appropriate standard-setting method has a paramount influence on the examinee's performance results (4). The criteria that help decide the method of standard setting are the systematic nature of the method, reproducible results, being absolute and unbiased (5). Another factor to consider when choosing a standard-setting method is the available resources and expertise in the institutions (6). Also, validation of standard-setting methods is a crucial component of quality assurance in medical education as there are no gold standard methods of standard setting to be named (4-5).

Thirty or more standard-setting methods have been grouped into relative, test-centered, and student-centered (7). Assessments in medical education use criterion-referenced methods (test/examinee) to determine mastery of the candidates and norm-referenced or relative methods to rank examinees (8). The most common methods used for standard setting are norm-referencing, Angoff method, Borderline Group Method (BGM), Borderline Regression Method (BRM), and Cohen/Modified Cohen's method (9). Though the Hofstee method is used in medical education assessments, researchers consider it less accurate due to its reliance on examiners (10). In norm-referenced methods, the pass/fail scores are determined by the relative scores of students in a particular exam (2). In the predetermined Angoff standard setting, the pass/fail scores are determined based on the items of the respective OSCE stations. In this method, after defining a borderline student, standard setters are asked to review a question as a whole and agree on whether the borderline group of students will pass the station. Then, the Angoff ratings are averaged to calculate the question pass mark (7, 11). In BGM, content experts are selected as judges to determine a 'borderline' student. Then, the mean scores of the examinees classified as 'borderline' are calculated for each station, which would then be the station's score (12). In the modified BGM, the station's mean score is averaged with the 'borderline' mean scores to achieve a final pass score (13).In the BRM method, known as the best method to use in OSCE (14), each examiner is asked to give a global rating score for each student for an OSCE station. The global rating score includes a good pass, pass, borderline, or fail. The global rating scores are then statically regressed against the checklist for the respective OSCE station. The pass mark is calculated using a linear equation by assigning the midpoint of the global rating scale against the borderline group's marks (14-15). Cohen's method, first mentioned in 2010 by Janke Cohen Schotanus, is based on the best cohort of students' performance and assumes that fluctuations in students' marks reflect the difficulty level of the exam or the quality of teaching. This method uses the 60% of 95th percentile students as the reference point for the pass score (4). Cohen's method was questioned for its fairness and the subjectivity of the 'multiplier' (0.6) used to calculate the pass scores (16). Cohen's method also relies on the assumption that the student's score in the 95th percentile is an accurate indicator of exam difficulty, which is consistent over time (16). Hence, Taylor, in 2011, devised a modified Cohen's method which used the formula 0.65 x P 90 where 0.65 is obtained by rearranging the mean score and the P 90, indicating students' score at the 90th percentile, which is consistent over time (10, 16-17).

Khalid et al. (2021), in a comparison between the relative method and other standard-setting methods in a written exam, found that the Angoff method produced credible and reliable pass scores closer to the relative method. However, Cohen and Modified Cohen gave divergent results. They recommended further studies with different assessment formats and sample sizes (18). Though the Angoff method is well-lauded and widely used, McLachlan (2021), who deduced a borderline candidate from the continuous assessment of students rather than judges' inference, found Modified Angoff scores to be variable. Tavakol et al. (2023), who analyzed the knowledge-based tests of 358 final-year medical students using the Angoff method, recommended the use of the three-parameter item response theory to reduce inter- and intra-judgmental inconsistencies, which questions the validity and reliability of the Angoff method. Smee et al. (2023) found that BRM was more standardized than BGM. Goldenberg et al. (2021) have explained that it is generally accepted that absolute methods over relative methods are more appropriate when making high-stakes or summative decisions. So, medical educators must consider the test's context and aim (18, 19), keeping in mind that the determination of the passing score must demonstrate transparency, reproducibility, credibility, and feasibility (18). It is worth mentioning that these methods yield disparaging outcomes and can be employed depending on specific contexts, availability of resources, test type, student's level, and judges. While employing a specific standard-setting method, the faculty should be aware of all these subtle yet determinant factors to justify the chosen method (3). Given the premise that different standard-setting methods yield different cut scores and the absence of a

'gold standard' in standard setting, this study was embarked to compare five different standard-setting methods (norm-referenced, Angoff, BGM, BRM, and Modified Cohen's methods) in the first professional OSCE exam conducted in the Faculty of Medicine of a medical school in Malaysia to look into the cut-off scores and failure rates determined by each method. The current study would help educators compare the outcomes of five different methods in the same context, the understanding of which would help them choose a suitable standard-setting method.

Material & Methods

Design and setting(s)

A non-experimental, cross-sectional study was conducted at the faculty of medicine in a medical university in Malaysia. Five standard-setting methods were compared on the OSCE scores obtained in the first professional exam by 170 students at the end of year 2. The medical program at the above school is a five-year course with one intake per year in September every year. An integrated curriculum is followed in the medical school with the first two years of pre-clinical study and the next three years of clinical study. The curriculum encompasses two professional exams, both with an OSCE component. The first professional exam takes place at the end of year two before the students proceed to the clinical years, while the final professional exam takes place at the end of year five. The faculty has always employed modified Angoff as their standard-setting method in all their professional exams.

Participants and sampling

Since this study was a non-experimental, cross-sectional study that followed previous researchers (7, 15) and was not intended to test the effectiveness of a specific intervention in a specific population, total population sampling was employed where we chose to examine the entire population (20). This decision was based on the nature of our research question, which aimed to examine the performance of different standard-setting methods across the entire cohort rather than making inferences about a sample from a larger population. Marks from all 170 second-year medical students who attended the OSCE with eight stations in the First Professional Examination conducted in August 2022 were analyzed. The data was complete, and no data was excluded from the analysis.

Tools/Instruments

The checklists for eight stations, four for history taking and four for examination, were prepared by the Clinical Skills Training (CST) team and vetted by the central vetting committee before the examination. **Table 1** gives the blueprint of the OSCE.The skills evaluated were history taking and examination of the systems taught in the pre-clinical years. Each station had marks ranging from 10 up to 20 marks, depending on the items on the checklist. The checklist consisted of the steps the student should perform in every station (15). Each point in the checklist is awarded appropriate marks, and the total indicates the whole mark of the particular station (15). The global rating scale consisted of 1: Clear fail, 2: Borderline, 3: Clear pass, and 4: Good pass, irrespective of the student's scores in the checklist (15). The Deputy Dean Academic and Dean approved all questions before submitting to the Exam Unit. On the exam day, the CST coordinator briefed the examiners and simulated patients (SP) at their respective stations. The exam was conducted in two streams, each comprising eight active and two rest stations. Each active station was timed for six minutes, one minute was given to read the questions, and five minutes were spent on performance. During the exam, examiners at each station evaluated students' performance by completing the checklist and global rating scales.

Table 1. Blueprint of OSCE

Station no	System	Type of station
1	Cardiovascular system	Examination
2	Respiratory system	History taking
3	Nervous system	Examination
4	Reproductive system	History taking
5	Musculoskeletal system	Examination
6	Gastrointestinal system	History taking
7	Gastrointestinal system	Examination
8	Urinary system	History taking

Data collection methods

The total score from the checklist in every station was converted to 20 marks for standardization across the stations. The five standard-setting methods belonging to both relative and absolute methods were used to determine the cut score, as mentioned below:

Norm reference

The mean of each station is calculated, and the cut scores are obtained by subtracting 1 SD from the means of each station, as followed by previous researchers (21).

Angoff Method

In this method, five examiners (content experts) are asked to determine the cut score for each station. The examiners' cut scores are then averaged to determine the estimated passing scores for all students (7).

Borderline Group Method (BGM)

Examiners evaluate students' performance based on global rating scales. The cut scores are obtained from the average score of students with borderline ratings (13).

Borderline Regression Method (BRM)

OSCE checklist scores are used to develop a cut score using linear regression. Regression of global rating scores to OSCE total scores generated a linear equation. The predicted cut scores of the borderline group are established by substituting the borderline rating values, which is two multiplied by the regression equation (15).

Modified Cohen's Method

The students' scores are arranged from the lowest to the highest, 90% confidence intervals are highlighted, and the mean is determined (10, 16, 17). Ultimately, 65% of the total mean score is calculated and considered a passing score. This is presented in the formula 0.65 x P 90 (16).

After getting the cut of scores of each method in each station, the scores of each station were converted to the amended score by using the following formula:

Amended score = actual score x (pass old score/new pass score)

Example: 13 (actual score) x 10 (old) /12 (new) = 10.83 (amended score)

Subsequently, the total score was calculated by adding each station's scores.

Though the pass rate varied in each station, only the students who scored less than 50% of the total mark were categorized as failing in the overall OSCE.

Data analysis

Data was analyzed using STATA version 16. Descriptive statistics such as frequency and percentage were used to describe the number of borderline students and the failure rate of each station. Simple linear regression analysis was conducted to produce a regression equation in BRM. Cronbach's alpha analysis was used to measure internal consistency. Meanwhile, Spearman's correlation was used to measure the correlation between global rating and checklist score. The Cochran Q test was employed to compare the proportions between methods. The significant value for all statistical tests was set at 5%.

Results

The overall alpha of the score was 0.67. **Table 2** displays the results of the score across eight OSCE stations as correlated by Spearman's correlation, alpha if deleted, and the number of borderline students in each station. If the alpha if deleted for all stations is less than the overall Cronbach's alpha, it suggests that no item should be removed to increase the reliability of the exam. The Spearman's correlation coefficient of each station, ranging from 0.57 to 0.7, indicates that the global rating was moderately correlated with the checklist score. For station four, only three students were rated as borderline, which is in contrast to 42 students being borderline in station 2. Figure 1 illustrates the line graph for the pass score of each station according to five standard setting methods.

Table 2. Scores across 8 OSCE stations (n=170)

Station	R^2	Spearman' r	Cronbach's alpha if deleted	No of borderline student
1	0.38	0.60	0.65	30
2	0.60	0.75	0.64	42
3	0.43	0.57	0.63	33
4	0.29	0.58	0.66	3
5	0.52	0.70	0.64	36
6	0.55	0.70	0.60	20
7	0.42	0.63	0.63	34
8	0.51	0.65	0.63	23

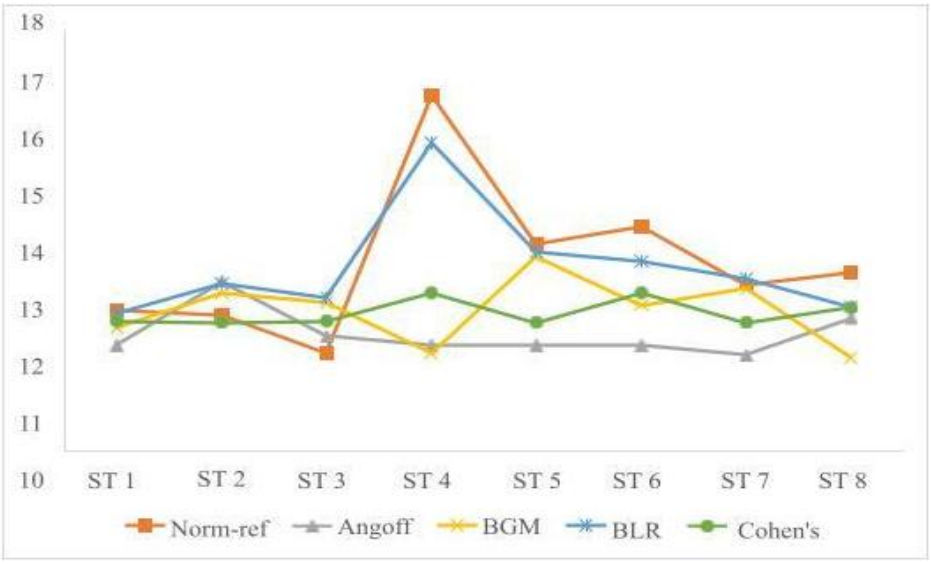

Figure 1. Line graph for pass score of each station according to five standard setting methods

Table 3 presents the cutoff scores for eight OSCE stations as determined by five standard setting methods. 62.5% of the highest cutoff scores were provided by the norm-reference method, as indicated in stations 1, 4, 5, 6, and 8. The highest cutoff scores in the remaining stations (Stations 2, 3, and 7) were given by the BRM method. The Angoff method yielded the lowest cutoff score in four stations (Stations 1, 5, 6, 7), followed by the BGM method for 2 stations (Stations 4 and 8) and with the Modified Cohen method for 1 station (Station 2) respectively.

Table 3. Cut off scores for 8 OSCE stations as determined by 5 standard setting methods

Station	Norm-ref	Angoff	BGM	BRM	M.Cohen's
1	12.7*	12.0†	12.3	12.6	12.5
2	12.6	13.2*	13.0	13.2*	12.4†
3	11.9†	12.2	12.8	12.9*	12.5
4	16.7*	12.0	11.9†	15.9	13.0
5	13.9*	12.0†	13.7	13.8	12.4
6	14.3*	12.0†	12.8	13.6	13.0
7	13.2	11.8†	13.1	13.3*	12.4
8	13.4*	12.5	11.8†	12.7	12.7

*Highest score/station
†Lowest score/station

Table 4 shows the failure rate of each station as per the cutoff score determined by the five standard setting methods and their analysis using the Cochrane Q test. At all stations, there were more failed students by the norm-referenced standard setting (stations 1, 4, 5, 6, and 8), followed by the BRM method (Stations 2, 3, and 7). On the contrary, the Angoff method determined the lowest cutoff scores in 4 stations (Stations 1, 5, 6, and 7). The p-value <0.001 of Cochrane's Q test means that the proportion of students who failed for the station was significantly different when different methods were used to determine the pass score.

The norm-reference method yielded the highest number of failures (6 stations) followed by the BRM method (5 stations), with stations 1, 5, and 7 yielding the same number of failures. It is interesting to note that these stations were examination stations of the cardiovascular system, musculoskeletal system, and gastrointestinal system. The Angoff method produced the lowest number of failures in 5 stations (Stations 1, 4, 5, 6, and 7), followed by BGM (Stations 2, 4, and 8) with station 4 yielding the same number of failures. The Norm reference method reported 2 stations with low failures (2 and 3) with station 2 similar to BGM and Modified Cohen's method results. BRM did not report any lowest number of failures.

Table 4. Failure rate of each station determined by 5 methods and analysed by Cochran's Q test

Station	Norm-reference n (%)	Angoff n (%)	BGM n (%)	BRM n (%)	M.Cohen n (%)	Cochran's Q (df)	P value
1	30 (18) *	16 (9) †	22 (13)	30(18) *	22 (13)	40.0 (4)	<0.001
2	29 (17) †	36(21)*	29 (17) †	36(21) *	29 (17) †	28.0 (4)	
3	25 (15) †	28 (17)	41(24) *	41(24) *	28 (17)	53.6 (4)	
4	21 (12) *	3 (2) †	3 (2) †	17 (10)	5 (3)	58.6 (4)	
5	20 (12) *	5 (3) †	20 (12) *	20 (12)*	11 (7)	48.9 (4)	
6	23 (14) *	8 (5) †	13 (8)	18 (11)	13 (8)	37.2 (4)	
7	21 (12) *	8 (5) †	21 (12)	21(12) *	21 (12)	52.0 (4)	
8	24 (14) *	18 (11)	14 (8) †	18 (11)	18 (11)	25.6 (4)	
Overall	10 (6)	2 (1)	3 (2)	8 (5)	3 (1.8)	24.2 (4)	

*- Highest number of failures/station
†- Lowest number of failures/station
Note: Cochran's method used to analyze scores determined by five standard setting methods.

Discussion

The standard setting method used in high-stakes examinations is a policy decision that should be defensible, as it is expected to be consistent and reflective of best practice (13). This study aimed to compare the cut-off scores and the resulting effects on student pass rates in an 8-station OSCE. The study found that 62.5% of the highest cut-off scores to pass were provided by the norm-reference method (5 stations), followed by the BRM method (3 stations). The pass cut-off score of the Angoff method was the lowest among 4 stations, followed by the BGM method (2 stations). The Angoff method yielded more stations with the lowest number of failures (n=5), followed by the BGM method (n=3). Higher numbers of failures were yielded by the norm-reference and BRM methods, notably in physical examination stations.

There are various reasons for the differing failure rates of different stations. This could be due to the varying difficulty levels of each station, especially those that assess physical examination skills rather than history taking (15). According to a review by Chong et al (2017), the type of station could be a possible cause for differences in scores. Research indicates a weak, yet existing relationship between examiner scoring and exam content. History taking or communication skills might involve less assessor interaction, resulting in higher scores, compared to examination stations. Examiner fatigue at examination stations may contribute to a higher failure rate (22).

In our current study, comparing all stations, more students failed by the norm-referenced standard setting, followed by the BRM method, probably due to the high cut-off score set by the respective methods. On the contrary, the Angoff method yielded the least failed students due to the lowest predicted cut-off score. Our Cochrane's Q test yielded a $p < 0.001$, which signifies that the proportion of students who failed a particular OSCE station significantly differed when different methods were used to determine the pass score. The Angoff method yields the least failures because it is a criterion-referenced (absolute) method that intends a desirable competency in the candidates so all candidates can either pass or fail the test (3). On the contrary, the norm-referenced (relative) method requires a fixed number of candidates to pass, causing the standard to rely on the cohort of students, thereby increasing the failure rate (3). Also, relative standard-setting methods are less defensible because factors such as test difficulty and examinee ability that could influence passing scores are not considered (4).

A recent study (23) had shown that the modified Angoff method, compared to four other methods of standard-setting (traditional, modified borderline group method, Kaufman's relative method, and BRM), to be consistently reliable and practically suitable to determine cut-off scores in multiple OSCE stations. Similar to our study, their study showed that the modified Angoff method yielded lower failure rates (typically less than 10% per station) compared to modified BGM and BRM, which had failure rates ranging from 28% to 57%. Comparison between BRM and BLM indicates that BRM is more reliable as it utilizes all scores rather than only the scores of borderline students. This ensures high validity, and we can derive the scores immediately after the exam (4). This result was also echoed in a study conducted by Elabd et al. (2023). However, BRM resulted in more failures due to the higher cut-off score than other methods (15). Since the BRM is based on a subjective assessment of student performance, it could

also be not very objective in specific settings (15). A way to reduce bias in BRM is to prevent the examiner from calculating the marks of the checklist and marking the students' performance (15). Calculating the marks might distract the evaluation process, leading to biased marks (15). Similar results were found in a study that analyzed the OSCE scores among 112 nursing program students in Canada, which reported lower cut-off scores with modified BGM than BRM (13). The cut-off score in Cohen's method is 60% of the 95% percentile examinee, which is considered an accurate indicator of exam difficulty and is consistent over time. On the contrary, its limiting factors include considering the highest examinee performance rather than the actual performance of all the examinees, which is combated by Modified Cohen's method (4, 8, 16).

Another recent study compared the norm-referenced method to BGM and BRM in an OSCE examination administered to 107 4thyear medical students in Korea (24). They found that the cut scores determined by BGM and BRM were higher than those determined by the norm-referenced method. However, no comparisons were made using the Angoff method or other methods of standard setting that employed expert judgment. BGM is feasible for ease of cut score calculation but does not hold well when there are insufficient borderline scorers, ensuring that BRM is a better reliable method (24). The efficacy and feasibility of BRM have been substantiated in a study conducted by Wood et al. (2006) as well (25). A study on two cohorts sitting for the same OSCE was done on a group of junior and senior sports medicine residents (24). Comparing the modified Angoff method to BGM and BRM methods, they found that the cut-off scores between the two methods did not differ when tested in senior residents, who had one to four years of extra training compared to the junior residents. The modified Angoff method yielded lower cut-off scores than the BGM and BRM methods when tested in a group of junior residents, similar to our study (24).

The Angoff method has been compared to the Hofstee method in the past to establish the minimum passing scores for advanced cardiac life support procedures (26). Similar to our study, researchers discovered that the Angoff method resulted in lower minimum passing scores compared to the Hofstee method. However, their study did not compare their findings to statistical-based methods like BGM and BRM. It's important to note that both the Angoff and Hofstee methods rely solely on expert judgment, as evidenced by their study, which involved a panel of twelve diverse health professionals (26). Another study involving 54 2nd year physician assistant students also yielded similar results. This study examined the reliability of BGM versus the Angoff method in a multi-station standardized patient clinical skills examination (27). The study found that the BGM method set an overall cut score of 76% (95% CI +/- 5), while the Angoff method set a lower cut score at 62% (95% CI +/- 9), which aligns with our current study. In a study involving 78 4th-year medical students taking a multiple-choice examination, standard setting by the Angoff method resulted in a pass rate of 100%, compared to the norm-reference method, where only 85% of students passed (21). This study also reported higher inter-rater reliability and moderate test-retest reliability using the Angoff method (21).

Contrary to our current study, research conducted in Egypt (10) and Iran (28) comparing four standard setting methods each showed that both the BRM and the Cohen's method yielded lower pass marks compared to the Modified Angoff method. Previous studies on other types of assessments have also shown limited agreement between the Modified Angoff method and other standard setting methods. Studies have also revealed higher cut-off scores with the Angoff method, such as a study on standard setting for multiple choice questions in which the Angoff method yielded a cut score of 54.98%, which was higher than the cut score suggested by the Hofstee method at 44% (29). A study by Elfaki & Salih (2015) on One Best Answer (OBA) scores found that the passing score by the Angoff method was higher compared to the norm reference method (48 vs 35) with 36% agreement. This led to a higher failure rate with the Angoff method compared to the norm reference method (61% vs 12%) (5).

In a study to explore the possibility of using the Angoff method to determine the cut score for a nursing licensing examination, it was found that standard setting using the Angoff method yielded a higher cut score of 74.4% and 76.8% in two mock exams. The cut-off score was much higher than the traditional method of 60% of the total score, which was the standard for the licensing examination (30). Another study to determine cut scores for a national licensing medical examination found that the difference between the cut scores produced by the Angoff and Hofstee methods did not exceed 2% points (31).

The Angoff method relies on applying judgment in defining a borderline student and observing and evaluating students' performance in the exam, hence deemed subjective (3). George et al. (2006) defined a

borderline candidate as "the one who has a precisely 50:50 probability of passing or failing the test (21). Angoff method is considered appropriate as the set standard is defensible as it gathers judgment from experts in an unbiased way considering the level of the candidates and content of the examination compared to a predefined score (3). Acceptable results in the Angoff method rely on selecting appropriate numbers and mixing the judges for various viewpoints (3). A range of 5-20 judges with a group of 10 is deemed suitable for this process (3). Some have argued that the maximum number of judges can be 30 as more significant numbers yield more valid findings and reproducible results (5, 21). It is also paramount that the judges are content experts and be meticulously selected to yield a good mixture of age, gender, ethnicity experience, familiarity with the student, seniority, and sub-specialization (5, 21). Another factor to be considered is the adequate training of judges in practice, characterizing borderline students by delineating skills, discussion, achieving consensus, and practice (5). Apart from being credible experts in their field, judges must be familiar with the performance level of the students taking the test (32). The process validity of the Angoff method is enhanced by aptly defining minimum competency and assumption of response probability (30). The advantages of the Angoff method are that it is highly intuitive, has knowledge of pass-fail scores before the exam, and has sufficient intra-panel and inter-panel reliability. The disadvantages are the laborious process, time-consuming, varied quality of judges, and judgmental inconsistencies given the judges' conceptualization and competency (32). The number of examinees also influences the cut-off score, as evidenced by Malau-Aduli et al. (2017). This is because the higher the examinee number, the more the error margin shrinks, reducing heterogeneity in variance and allowing better correlations (4).

Some researchers have questioned the credibility of the Angoff method as it is based on cognitive judgments (7). On the other hand, opponents of this view claim that Angoff is a user-friendly, well-researched, technically sound method continually used in UK medical schools (7). A recent review conducted by Saaiq (2024) concluded that there needs to be a universal consensus regarding the best method for standard setting in assessments testing the acquisition of knowledge and skills in medical education. However, test-centered or item-centered methods are preferred for written assessments, and BRM is more suited for skills assessment. The choice of the method follows the available facilities and the nature of the assessment (9). The practical applicability of the study is that medical educationists can use the results to determine the type of standard-setting method for high-stakes examinations. This study has compared five methods and provided reproducible results. Medical educationists should work towards adopting valid and reliable methods of standard setting to produce fair results.

Conclusion

The study that compared five standard-setting methods in the year 2 OSCE examination for medical students concluded that the cut-off score and failure rate differed with different standard-setting methods. The choice of method is contextual depending on the available resources such as faculty members such as judges, statisticians, students' level of competence, and organizational support. The norm-referencing method yielded the maximum number of highest-cut-off scores/station with a higher failure rate, and the modified Angoff method yielded the maximum number of lowest cut-off scores/station with a lower failure rate.

It is important to note that this study was conducted on the OSCE scores obtained by preclinical students in a single institution, which may limit the generalizability of the results. The context-specific nature of each standard-setting method further complicates direct comparisons. Given these potential sources of variability, we recommend conducting large-scale studies in both written exams and clinical examinations in different curricular setups to further our understanding.

Ethical considerations

This study received an ethical exemption from the University of Cyberjaya (UOC/CRERC/ EXEMPTION/01/2023) for data collection. Only student scores were taken for analysis. The students were not identified. The data was accessible only to the authors who analyzed it.

Artificial intelligence utilization for article writing

No artificial intelligence was used to write this article.

Acknowledgments

Nil.

Conflict of interest statement

The authors declare no conflict of interest.

Author contributions

NAbM and NAM conceived the research concept. The data was collated by NSO, AAL and SI. NAbM conducted the statistical analysis. NAM and ASO confirmed the findings. The outline of the manuscript was written by RMA. It's important to note that all authors, without exception, equally contributed to the writing of the article. The final manuscript was approved by all the authors, reflecting our collective effort and commitment.

Supporting resources

No grants or other resources were used to fund this study.

Data availability statement

The data will be available upon special request.

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. Annals of Medicine and Surgery. 2020;49:19-23. [https://doi.org/10.1016/j.amsu.2019.11.013]
2. 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. Medical Teacher. 2013;35(9):e1437-46. [https://doi.org/10.3109/0142159X.2013.818634]
3. Hejri SM, Jalili M. Standard setting in medical education: fundamental concepts and emerging challenges. Medical Journal of the Islamic Republic of Iran. 2014;28:34.
4. 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. Medical Teacher. 2017;39(12):1261-7. [https://doi.org/10.1080/0142159X.2017.1372565]
5. Elfaki OA, Salih KM. Comparison of two standard setting methods in a medical students MCQs exam in internal medicine. American Journal of Medicine and Medical Sciences. 2015;5(4):164-7. [https://doi.org/10.5923/j.ajmms.20150504.04]
6. Abd-Rahman AN, Baharuddin IH, Abu-Hassan MI, Davies SJ. A comparison of different standard-setting methods for professional qualifying dental examination. Journal of Dental Education. 2021;85(7):1210-6. [https://doi.org/10.1002/jdd.12600]
7. Tavakol M, O'Brien D, Stewart C. Determining intra-standard-setter inconsistency in the Angoff method using the three-parameter item response theory. International Journal of Medical Education. 2023;14:123. [https://doi.org/10.5116/ijme.64ed.e296]
8. Cohen-Schotanus J, van der Vleuten CP. A standard setting method with the best performing students as point of reference: practical and affordable. Medical Teacher. 2010;32(2):154-60. [https://doi.org/10.3109/01421590903196979]
9. Saaiq M. Standard setting methods for the assessment of knowledge and skills in medical education. Journal of Health Professions Education and Innovation. 2024;1(2):14-20. [https://doi.org/10.21608/JHPEI.2024.266664.1019]
10. Kamal D, Sallam M, Gouda E, Fouad S. "Is there a "best" method for standard setting in OSCE exams? Comparison between four methods (a cross-sectional descriptive study). Journal of Medical Education. 2020;19(1). [https://doi.org/10.5812/jme.106600]
11. 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 Medical Education. 2016;16:1-7. [https://doi.org/10.1186/s12909-015-0506-z]
12. Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. Medical Education. 2004;38(2):199-203. [https://doi.org/10.1111/j.1365-2923.2004.01755.x]
13. Smee S, Coetzee K, Bartman I, Roy M, Monteiro S. OSCE standard setting: three borderline group methods. Medical Science Educator. 2022;32(6):1439-45. [https://doi.org/10.1007/s40670-022-01667-x]
14. 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. Medical Education. 2003;37(2):132-9. [https://doi.org/10.1046/j.1365-2923.2003.01429.x]
15. Elabd K, Abdul-Kadir H, Alkhenizan A, Alkhalifa MK. A Comparison of the checklist scoring systems, global rating systems, and borderline regression method for an objective structured clinical examination for a small cohort in a Saudi medical school. Cureus. 2023;15(6). [https://doi.org/10.7759/cureus.39968]
16. Taylor CA. Development of a modified Cohen method of standard setting. Medical Teacher. 2011;33(12):e678-82. [https://doi.org/10.3109/0142159X.2011.611192]
17. McLachlan JC, Robertson KA, Weller B, Sawdon M. An inexpensive retrospective standard setting method based on item facilities. BMC Medical Education. 2021;21:1-7. [https://doi.org/10.1186/s12909-020-02418-5]
18. Khalid MN, Shafiq F, Ahmed S. A Comparison of standard setting methods for setting cut-scores for assessments with constructed response questions. Pakistan Journal of Educational Research and Evaluation. 2021;9(2):74-85.
19. Goldenberg M, Ordon M, Honey JR, Andonian S, Lee JY. Objective assessment and standard setting for basic flexible ureterorenoscopy skills among urology

trainees using simulation-based methods. Journal of Endourology. 2020;34(4):495-501. [https://doi.org/10.1089/end.2019.0626]
20. Martínez-Mesa J, González-Chica DA, Duquia RP, Bonamigo RR, Bastos JL. Sampling: how to select participants in my research study? Anais Brasileiros de Dermatologia. 2016;91(3):326-30. [https://doi.org/10.1590%2Fabd1806-4841.20165254]
21. George S, Haque MS, Oyebode F. Standard setting: comparison of two methods. BMC Medical Education. 2006;6:1-6. [https://doi.org/10.1186/1472-6920-6-46]
22. Chong L, Taylor S, Haywood M, Adelstein BA, Shulruf B. The sights and insights of examiners in objective structured clinical examinations. Journal of Educational Evaluation for Health Professions. 2017;14. [https://doi.org/10.3352/jeehp.2017.14.34]
23. Dwivedi NR, Vijayashankar NP, Hansda M, et al. Comparing standard setting methods for objective structured clinical examinations in a Caribbean medical school. Journal of Medical Education and Curricular Development. 2020;7:2382120520981992. [https://doi.org/10.1177/2382120520981992]
24. Park SY, Lee SH, Kim MJ, Ji KH, Ryu JH. Comparing the cut score for the borderline group method and borderline regression method with norm-referenced standard setting in an objective structured clinical examination in medical school in Korea. Journal of Educational Evaluation for Health Professions. 2021;18;25. [https://doi.org/10.3352/jeehp.2021.18.25]
25. 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. Advances in Health Sciences Education. 2006;11:115-22. [https://doi.org/10.1007/s10459-005-7853-1]
26. Wayne DB, Fudala MJ, Butter J, et al. Comparison of two standard-setting methods for advanced cardiac life support training. Academic Medicine. 2005;80(10):S63-6. [https://doi.org/10.1097/00001888-200510001-00018]
27. Carlson J, Tomkowiak J, Knott P. Simulation-based examinations in physician assistant education: a comparison of two standard-setting methods. The Journal of Physician Assistant Education. 2010;21(2):7-14. [https://doi.org/10.1097/01367895-201021020-00002]
28. Jalili M, Mortazhejri S. Standard setting for objective structured clinical exam using four methods: Prefixed score Angoff borderline regression and Cohens. Strides in Development of Medical Education. 2012;9(1):77-84.
29. Kamal D, ElAraby S, Kamel MH, Hosny S. Evaluation of two applied methods for standard setting in undergraduate medical programme at the faculty of medicine, Suez Canal university. Education in Medicine Journal. 2018;10(2):15–25. [https://doi.org/10.21315/eimj2018.10.2.3]
30. Yim MK, Shin S. Using the Angoff method to set a standard on mock exams for the Korean nursing licensing examination. Journal of Educational Evaluation for Health Professions. 2020;17(14):1149169. [https://doi.org/10.3352/jeehp.2020.17.14]
31. Kim DH, Kang YJ, Park HK. Possibility of independent use of the yes/no Angoff and Hofstee methods for the standard setting of the Korean medical licensing examination written test: a descriptive study. Journal of Educational Evaluation for Health Professions. 2022;19:33. [https://doi.org/10.3352/jeehp.2022.19.33]
32. Verheggen MM, Muijtjens AM, Van Os J, Schuwirth LW. Is an Angoff standard an indication of minimal competence of examinees or of judges? Advances in Health Sciences Education. 2008;13:203-11. [https://doi.org/10.1007/s10459-006-9035-1oi]