A Competency-Based Approach to Pass/Fail Decisions: An Observational Study

Nazdar E. Alkhateeb[1], Ali Al-Dabbagh[1], Yaseen Mohammed[2], Mohammed Ibrahim[3]

[1] Department of Medical Education, College of Medicine, Hawler Medical University, Erbil, Kurdistan Region, Iraq

[2] Department of Community Health, Cihan University, Erbil, Kurdistan Region, Iraq

[3] Department of Child Health, School of Medicine, University of Dundee, Dundee, Scotland, United Kingdom

Received: 21 Oct. 2020; Accepted: 22 May 2021

Abstract- Any high-stakes assessment that leads to an important decision requires careful consideration in determining whether a student passes or fails. Despite the implementation of many standard-setting methods in clinical examinations, concerns remain about the reliability of pass/fail decisions in high stakes assessment, especially clinical assessment. This observational study proposes a defensible pass/fail decision based on the number of failed competencies. In the study conducted in Erbil, Iraq, in June 2018, results were obtained for 150 medical students on their final objective structured clinical examination. Cutoff scores and pass/fail decisions were calculated using the modified Angoff, borderline, borderline-regression, and holistic methods. The results were compared with each other and with a new competency method using Cohen's kappa. Rasch analysis was used to compare the consistency of competency data with Rasch model estimates. The competency method resulted in 40 (26.7%) students failing, compared with 76 (50.6%), 37 (24.6%), 35 (23.3%), and 13 (8%) for the modified Angoff, borderline, borderline regression, and holistic methods, respectively. The competency method demonstrated a sufficient degree of fit to the Rasch model (mean outfit and infit statistics of 0.961 and 0.960, respectively). In conclusion, the competency method was more stringent in determining pass/fail, compared with other standard-setting methods, except for the modified Angoff method. The fit of competency data to the Rasch model provides evidence for the validity and reliability of pass/fail decisions.

Acta Med Iran 2021;59(7):421-429.

Keywords: Pass/fail decision; Competence-based; Standard-setting; Rasch model

Introduction

Adopting a competency-based approach to teaching necessitates changing the assessment methods (1). One of the greatest challenges to institutions responsible for training and certifying physicians in assessing clinical competence (2-4), which is significant because it helps to protect patients by determining whether students can progress to higher levels of study and/or medical qualification.

Objective structured clinical examinations (OSCEs) are the most commonly used clinical competency assessment tools. If OSCEs are correctly designed and analyzed, they can benefit medical students' learning and future performance (5). A key but difficult task in OSCEs is making pass/fail decisions in cases of borderline performance. To handle this issue, many standard-setting methods have been introduced and implemented in a range of clinical examinations (6).

However, concerns regarding the reliability, validity, and acceptability of these methods remain an issue (7). The differences in cutoff scores among different standard-setting methods may reduce the legal defensibility of these cutoffs, especially when it led to differences in the pass/fail decision (8-9).

To recognize the relationship between the observed (actual) score on an examination and the underlying competence in the domain, which is commonly unobserved, a test theory model is necessary (10). Item response theory (IRT), which has received little attention in the medical education literature (11), provides deeper analysis and gives a range of information on the behavior of individual test items (difficulty), individual students (ability), and the underlying construct being examined

Corresponding Author: N.E. Alkhateeb
Department of Medical Education, College of Medicine, Hawler Medical University, Erbil, Kurdistan Region, Iraq
Tel: +964 7504534461, Fax: +964 7504534461, E-mail address: nazdar.alkhateeb@hmu.edu.krd

(10). By using IRT models (e.g., the Rasch model), standard setters can identify items that are not mapped to student ability (i.e., items that are either too difficult or too easy for a particular cohort (12). IRT models represent a powerful method for interrogating clinical assessment data, resulting in more valid measures of students' clinical competencies to inform defensible and fair decisions on students' progression and certification (13). Using Rasch measurement theory to establish competency level is a technique that can help decision-makers by simplifying data in a meaningful way and correcting for a mix of judge types (14).

This study aimed to examine pass/fail decisions using different standard-setting methods and to compare these with a new method based on the number of failed competencies. This will contribute to filling the knowledge gap regarding where to draw the line for the borderline performer and help to provide successful remediation. To achieve these aims, this study sought to answer the following questions: Can the number of failed competencies facilitate pass/fail decisions? How accurately do the -competency method data fit the Rasch model to provide a defensible pass/fail decision?

Materials and Methods

Study context

This observational study was conducted at Hawler Medical University, College of Medicine.

Participants

The data used in this study were examination results obtained from 150 final-year medical students completing an OSCE as part of their exit examination in June 2018.

The OSCE

The OSCE included 23 work stations and two rest stations. A time of 5 minutes was allocated for each station, so each OSCE session took approximately 125 minutes. The examination was run in three parallel circuits on three different floors of the same hospital and was conducted in two rounds. Students were isolated between rounds to decrease the risk of sharing exam materials. Stations used real patients, trained simulated patients, manikins, or data and videos.

Each station's score sheet contained a detailed checklist of the items examined. The checklist was scored with a maximum of 100 points. A global rating was also included for each station (1=fail, 2=borderline, 3=pass, 4=above the expected level).

The content validity of the OSCE was established by using blueprinting to ensure an adequate sampling across subject areas and competencies. The following skills were included in the blueprint as competencies (15): history taking (five stations), physical examination (six stations), data interpretation (five stations), procedural skills (two stations), communication skills (two stations-one station testing counseling skills and one history-taking station also assessing communication skills) and patient management (four stations).

For the quality assurance of the stations, question selection was conducted at both the department and the faculty level. The stations were prepared and reviewed for accuracy by the OSCE Committee, which included members from all departments. Stations were written in advance of the examination date, including instructions for the students, notes for the examiners, scenarios for the standardized patients, a list of requirements for each station, and marking sheets. To ensure the consistency and fairness of the scores, training was conducted for both the examiners and the standardized patients.

Because each student taking the OSCE had to perform a number of different tasks at the stations, this wide sampling of cases and skills should result in a more reliable picture of a student's overall competence. Moreover, as the students moved through the stations, they were examined by a number of different examiners, serving to reduce individual examiner bias.

Standard-setting methods

Standards were set using four different methods: a holistic score of 50% (university regulation), the modified Angoff (MA) method, the borderline regression (BLR) method, and the borderline (BL) method.

Angoff standards (16,17) for all stations were set by a group of eight experts. All experts had participated in teaching clinical sessions and assessing OSCE examinations. Two meetings were arranged for the experts; in the first meeting, the experts were asked to outline the criteria for a borderline (minimally competent) student. The researchers used simple language and avoided educational terminology to define the borderline student as one who is good enough to pass the exam, but it is difficult to gain a score above the pass mark. Through discussion, the experts reached a consensus on characteristics of a borderline student who was performing at a level between 'pass' and 'fail' Using this definition, for each item on the checklist, the experts were asked to estimate the probability that a borderline student would perform that item correctly (on a scale of 0-100

Each item with expert's variation of more than 20% were highlighted by the researcher. A second meeting arranged to discuss the highlighted items among experts who had a chance to reconsider their judgement. The MA passing score was calculated for each station by averaging the estimates across experts and items. The MA passing score for the total exam was calculated by averaging the 23 station passing scores.

BLR (18) was the second method investigated. We performed a linear regression analysis, using student performance as total percentage scores and examiner global ratings (fail=1, borderline=2, pass=3; and above the expected level=4) to determine the cutoff score. The cutoff score was derived by substituting the borderline value (2) into the regression equation.

For the BL method, students with borderline performances were identified, and their checklist scores were collected. The mean score for this group was set as the passing score (19).

To assess the passing score for each competency, we first calculated the means of both checklist scores and global ratings of the stations assessing each individual competency, so each student had a mean checklist score and a mean global rating for each competency. The passing score of each competency was calculated using the BLR method, where students who failed more than half of the competencies (*i.e*. three or more of the six competencies) were determined to have failed the examination as shown in Figure 1. Rasch item fit statistics were used to show how well the data for the competency method fit the Rasch measurement model. We then compared the pass/fail decision according to each standard-setting method with the pass/fail decision considering those who failed three or more competencies to have failed the exam.

Data collection and data analysis

A retrospective analysis carried out using data acquired from final OSCE exams (raw scores for each of the 23 stations for the 150 students were obtained from the College of Medicine) with the goal of calculating the cutoff score for passing using different standard-setting methods and comparing these methods with the new competency method.

Statistical analysis

SPSS, Version 23 and Excel 2010 were used for the data analysis. Cohen's kappa was used to measure agreement between standard-setting methods on a pass/fail decisions; this statistic can be interpreted as follows: Values ≤0 indicate no agreement, values of 0.01-0.20 indicate no agreement to slight agreement, values of 0.21-0.40 indicate fair agreement, values of 0.41-0.60 indicate moderate agreement, values of 0.61-0.80 indicate substantial agreement and values of 0.81-1.00 indicate almost perfect agreement (20).

Analytical plan

A one-parameter (Rasch) IRT model was fitted to the data on 150 students and six competencies. We estimated competency difficulty based on how the student answered. Item difficulty is the value along the latent variables continuum at which a student has a 50% probability of passing each competency. In the Rasch model, the disparity between student ability and item difficulty predicts the likelihood of a correct answer. For example, if the difference between student ability and item difficulty is zero, there will be a 50% likelihood of a student answering a question correctly. Higher item (competency) difficulty estimates indicate that students require a higher level of the ability to have a 50% probability of passing the competency (11).

An expected score will be calculated from each observed score in the Rasch process, using a *t*-test for each item. Item fitness to the Rasch model can be identified by means of infit and outfit statistics, which are expressed as 'infit mean square' or 'infit *t*' and 'outfit mean square' or 'outfit *t*.' A value of 1 for an outfit indicates a perfect fit, whereas values less than 0.70 indicate misfit, and values greater than 1.30 indicate overfit. Infit *t* values also show the degree to which a question fits the Rasch model. Observed data follow the Rasch model if the results of infit *t* are non-significant (*t* from -2 to 2) (11).

The item information function was calculated mathematically using the Rasch method by combining information on student ability and item difficulty. The sum over all items was plotted against student ability, giving the 'test information function' curve, which allowed the estimation of reliability at different levels of student ability. A tall narrow curve indicates a test containing highly discriminating items; less discriminating items provide less information but over a wider range (11).

Results

The MA method yielded the highest mean passing score, 61.19, which resulted in a failure rate of 50.6%. The lowest failure rate produced by the holistic method, only 8%. The other standard-setting methods' cutoff scores and failure rates are shown in Table 1.

To check for internal consistency reliability of the OSCE, Cronbach's alpha was computed across the 23 stations for all students (n=150) and was found to be 0.8.

Table 1. Standard-setting procedures applied to the 23 OSCE stations

Standard-setting method	Passing cutoff score	Number of students failing	Failure rate (%)
MA	61.19	76	50.6
BL	55.73	37	24.6
BLR	54.93	35	23.3
Holistic method	50	13	8
Competency method	≥3 competencies	40	26.7

OSCE: objective structured clinical examination; MA: modified Angoff; BL: borderline; BLR: borderline regression

Table 2 shows that failure in three or more competencies coincided with the student failure, as assessed by the holistic, BL, and BLR methods, with 100%, 82.8%, and 81% agreement, respectively.

Cohen's kappa values for the different standard-setting methods on 3450 decisions across the 23 OSCE stations are shown in Table 3.

Table 2. Percent agreement of the competency method with other standard-setting methods on which students failed the OSCE

Number of failed competencies	Standard-setting method N (%)			
	MA	**BL**	**BLR**	**Holistic**
<3	39 (51.3)	7 (19)	6 (17.2)	0 (0)
≥3	37 (48.7)	30 (81)	29 (82.8)	13 (100)
Total number of students failing	76	37	35	13

OSCE: objective structured clinical examination; MA: modified Angoff; BL: borderline; BLR: borderline regression

Table 3. Cohen's kappa for the different standard-setting methods

Standard-setting methods		Kappa	Sensitivity (identified a passing student)	Specificity (identified a failing student)	Mean kappa
MA	**Holistic**	0.169	54	100	0.392
	BL	0.483	65.5	100	
	BLR	0.457	64.3	100	
	Competency	0.457	64.9	94.9	
			62.175	98.727	
BL	**Holistic**	0.449	82.5	100	0.653
	MA	0.483	100	48.7	
	BLR	0.963	98.3	100	
	Competency	0.718	93.7	76.9	
			93.625	81.4	
BLR	**Holistic**	0.475	83.9	100	0.652
	BL	0.963	100	94.6	
	MA	0.457	100	46.1	
	Competency	0.713	94.6	74.4	
			94.625	78.775	
Holistic	**MA**	0.169	100	17.1	0.379
	BL	0.449	100	35.1	
	BLR	0.475	100	37.1	
	Competency	0.425	100	33.3	
			100	30.65	
Competency	**Holistic**	0.425	81	100	0.578
	BL	0.718	92	81.1	
	BLR	0.713	91.3	82.9	
	MA	0.457	97.3	48.7	
			90.4	78.175	

OSCE: objective structured clinical examination; MA: modified Angoff; BL: borderline; BLR: borderline regression

Cohen's kappa values ranged from 0.169 between the holistic and MA methods to 0.963 between the BL and BLR methods. The BL method had the highest mean kappa value (0.653), whereas the holistic method had the lowest mean kappa value (0.379). MA had the highest mean specificity to detect a failing student (98.727), whereas the holistic method had the lowest specificity (30.65).

The outfit and infit statistics showed that all competencies were within the acceptable range (both for mean square and *t* values) and accurately fit the Rasch measurement model (Table 4).

Discrimination describes how well the OSCE items (competencies) separate students with abilities below the competency location from those with abilities above the competency location. One-parameter IRT often assumes fixed discrimination among all competency items. In practice, a high discrimination parameter (>1) means that the probability of a correct response increases more rapidly as ability increases. Here, the discrimination value was 1.31, which indicates that the competencies better discriminate between high- and low-ability students than expected for items of this difficulty.

As shown in Figure 2, the competency of 'examination' was the least difficult, and 'management' was the most difficult. The change in difficulty shifts the item characteristic curves (ICCs), along with ability. The probability of success was higher for the competency of 'examination' than for the other competency items at any ability level. A student would only need an ability level greater than -1.78 on the competency of 'examination' to be expected to succeed on the competency.

Similarly to the ICCs shown in Figure 2, the item information curves (IICs) shown in Figure 3 demonstrate that the 'management' item provides the most information about students' high ability levels (the peak of its IIC is farthest to the right) and the 'examination' item provides the most information about students' lower ability levels (the peak of its IIC is farthest to the left). All ICCs and IICs for the competencies have the same shape in the Rasch model (*i.e.*, all competencies are equally good at providing information about ability).

The test information relied on the competencies used and the students' ability. The test information can be calculated by summing all the competency information together. Figure 4 illustrates that the amount of information had a maximum at an ability level of approximately -1 when it is about 2.5. In other words, the competencies model is most informative when the ability of the student is equal to the difficulty of the competencies and becomes less informative as the student's ability moves away from the competency difficulty (*i.e.*, when the competency is either too easy or too difficult for the students).

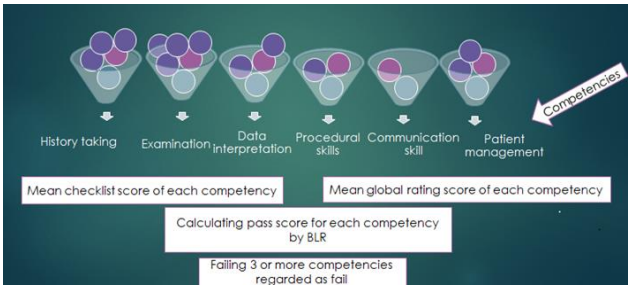

Figure 1. Competency method for pass/fail decision, number of stations for each competency represented in circle shapes

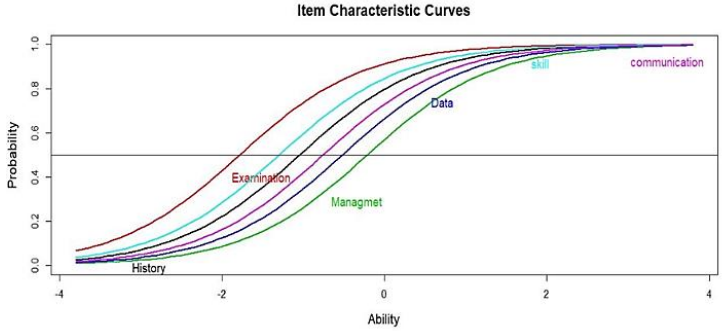

Figure 2. Plots of item characteristic curves for competencies in an item response theory model with competency difficulty levels of -4, 0, and 4. The competencies spread apart, representing varying levels of difficulty

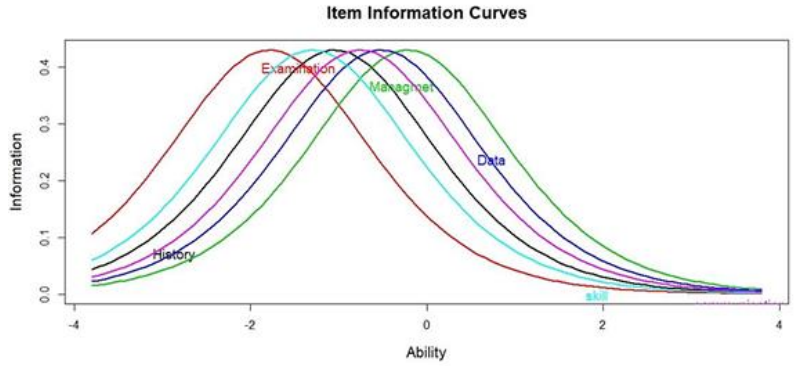

Figure 3. Plots of item information curves

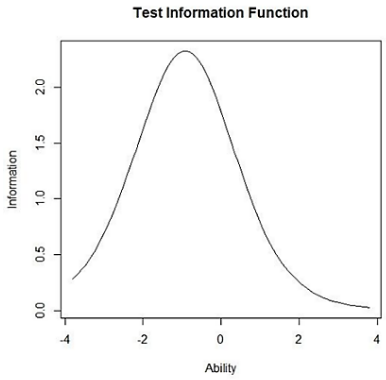

Figure 4. Item information function

Table 4. Item difficulty, standard error, and outfit and infit statistics

Competency	Item difficulty	Standard error	Outfit		Infit	
			MNSQ	ZSTD	MNSQ	ZSTD
History	-1.05	0.2	0.919	-0.51	0.937	-0.54
Examination	-1.78	0.27	0.82	-0.62	0.847	-0.95
Management	-0.21	0.17	1.07	0.6	1.097	1.19
Data	-0.52	0.18	0.706	-2.84	0.754	-2.99
Skill	-1.3	0.22	1.198	1.09	1.094	0.77
Communication	-0.76	0.19	1.048	0.43	1.041	0.45
Discrimination	1.31	0.16				

Note: Item difficulty measured in logits (negative values indicate easier questions)

Discussion

The assurance of sufficient quality and robust standard-setting is central to the delivery of any successful competency-based assessment (21). One of the most challenging aspects of clinical assessment is making pass/fail decisions for borderline grades allocated by examiners without adequate information to make these decisions (22). This study has proposed a new way of making pass/fail decisions in a high-stakes OSCE exam, incorporating the number of failed competencies, and examined the fitness of this new method to the Rasch model.

Different standard-setting methods may identify different cutoff scores for the same examination. Our study has shown diverging results, indicating there is no

one valid 'gold standard' method. The ranking of the existing standard-setting methods from most to least rigorous is as follows: MA, BL, BLR, and holistic methods. Despite the differences among standard-setting methods, the outcomes of the authentic methods (*i.e.*, cutoff scores and pass/fail decisions) should be similar (if not the same) for the same examination (23).

Prior studies have addressed this issue and investigated the validity of different standard-setting methods for OSCEs (23). Lee *et al*., (2018) used the MA method to calculate cutoff scores for three different domains (history taking, physical examination, and physician-patient interaction). One major drawback of his approach was that the cutoff score for passing in the MA method (with reality check) increases or decreases when performance data are provided to standard setters (9). In the present study, we combined the OSCE station scores measuring the same competency, calculated the cutoff score for each competency, and then we determined the pass/fail decision-based on the number of failed competencies.

Different standard-setting methods make different assumptions and determine cut scores differently. The competency method showed disagreement with the MA method, where only 48.7% of students who failed according to the MA method also failed when using the competency method. This difference might be caused by judges thinking about an average student instead of focusing on a borderline performer, leading to the substitution of a criterion-based concept with a norm-referenced one (24) and the setting of high cutoff scores. In contrast, the BL and BLR methods had the highest agreement with the other examined methods (mean kappa=0.63 and 0.62, respectively). A previous study indicated that the BL and BLR methods produce more realistic cutoff scores compared with the Angoff method (25). In the present study, the competency method has been shown to be more stringent than the other standard-setting methods except for the MA method in terms of pass/fail decisions; using the competency method would therefore increase the number of students failing the OSCE. However, this result may be desirable because the negative consequences of certifying an incompetent examinee (false positive) may far outweigh those of not certifying a competent one (false negative). It is important to minimise passing incompetent students (26). According to course and programme leaders in a previous study, examiners in clinical examinations were too lenient and tended to avoid failing students, especially by giving borderline students the benefit of doubt (22). Such a practice has the potential to have major adverse implications for medical practice (27).

The second aim of the present study was to determine how accurately the data of the competency method fit the Rasch model. Classical test theory and IRT are widely used to address measurement-related issues that arise from commonly used assessments in medical education, including OSCEs. Traditionally, post-examination analysis results are often based on classical test theory. However, statistics in classical test theory are based on the aggregate, and their values are sample-size-dependent. Medical educators need to investigate the relationship between students' ability (independent of item sample size) and the ease or difficulty of questions (independent of student sample size). IRT and one of its main models (Rasch) offers a comprehensive and forensic analysis of exam data that can be used to enhance test quality (11).

Furthermore, Rasch analysis provides beneficial graphical displays that aid test constructors in appraising the effectiveness of their assessments and, in the context of pass/fail decisions, enables us to establish the cutoff score for each competency according to student ability level. To judge the compatibility of the observed data with the Rasch model, mean square values were used: A value of 1 indicates a precise fit, whereas values from 0.70 to 1.30 indicate a good fit. However, values <0.70 or >1.30 are termed misfitting and overfitting, respectively, and should lead to an analysis of the items (28). Our results showed that the examined competencies accurately fit the Rasch model, with all competencies' mean square values within the acceptable range. The Rasch analysis showed that the competency of 'management' was the most difficult, requiring greater student ability to pass.

In contrast to other standard-setting methods, which reduce all of the information obtained from an assessment to a binary pass/fail judgment and simplify high-stakes decision making such that a minimally competent student is treated the same as a maximally competent student-meaning that both can graduate as doctors (29), the competency method provides rich data on each student's strengths and weaknesses. This presents an opportunity for students to learn from the assessment, guiding those who fail the examination in the remediation process and helping them to concentrate on their deficient competencies. The mantra that 'assessment drives learning' is often repeated with the belief that the effect of assessment is always useful (29). However, according to Kalet *et al*., (2012), effective remediation requires good data (30). In the present study, implementing the competency method in an OSCE exam provided data that

could facilitate effective remediation.

The main strength of the present study is the use of the Rasch statistical IRT model to enhance the credibility of competency-based pass/fail decisions. Therefore, this new competency method is more dependable, as it is derived from mathematical principles, whereas other methods are based on an overall impression of the examination difficulty and provide a less defensible cut score. This provides evidence for the validity and reliability of pass/fail decisions made using this method. Furthermore, the competency method can be used to set a cutoff score reflecting the desired student ability for each competency.

The variety of the student ethnic backgrounds as well as a retrospective analysis of their scores without them being aware are two factors that play a role in enhancing the external validity of the results.

However, obtaining data from final-year medical students in a single institution from one geographical region may limit the generalisability of our findings. It would be useful to include more students from different medical colleges in the region. Thus, our findings need to be interpreted with caution when applied to other institutional settings.

Our findings indicate the importance of combining the results of OSCEs based on content similarities of stations, which is more meaningful for a competency-based assessment and would enable faculty to draw more meaningful conclusions and provide actionable feedback.

Future research needed to test to what extent the suggested method

Acknowledgements

We thank Jennifer Barrett, PhD, from Edanz Group (www.edanzediting.com/ac) for critically reviewing and editing a draft of this manuscript.

References

1. Torbeck L, Wrightson AS. A method for defining competency-based promotion criteria for family medicine residents. Acad Med 2005;80:832-9.
2. Yang YY, Lee FY, Hsu HC, Huang CC, Chen JW, Lee WS, et al. A core competence-based objective structured clinical examination (OSCE) in evaluation of clinical performance of postgraduate year-1 (PGY1) residents. J Chinese Med Assoc 2011;74:198-204.
3. General Medical Council. Outcomes for graduates (Tomorrow's Doctors), 2015. (Accessed at: https://www.gmc-uk.org/-/media/documents/outcomes-for-graduates-jul-15-1216_pdf-61408029.pdf.)
4. Cleaton N, Yeates P, McCray G. Exploring the relationship between examiners' memories for performances, domain separation and score variability. Med Teach 2018;40:1159-65.
5. Gormley G. Summative OSCEs in undergraduate medical education. Ulster Med J 2011;80:127-32.
6. Lee M, Hernandez E, Brook R, Ha E, Harris C, Plesa M, et al. Competency-based Standard Setting for a High-stakes Objective Structured Clinical Examination (OSCE): Validity Evidence. MedEdPublish 2018;7:1-15.
7. Shulruf B, Jones P, Turner R. Using Student Ability and Item Difficulty for Making Defensible Pass/Fail Decisions for Borderline Grades. High Educ Stud 2015;5:107-18.
8. Wyse AE. Five Methods for Estimating Angoff Cut Scores with IRT. Educ Meas Issues Pract 2017;36:16-27.
9. Tavakol M, Dennick R. The foundations of measurement and assessment in medical education. Med Teach 2017;39:1010-5.
10. De Champlain AF. A primer on classical test theory and item response theory for assessments in medical education. Med Educ 2010;44:109-17.
11. Tavakol M, Dennick REG. Psychometric evaluation of a knowledge based examination using Rasch analysis : An illustrative guide : AMEE Guide No . 72 Psychometric evaluation of a knowledge based examination using Rasch analysis : An illustrative guide : AMEE Guide No . 72. Med Teach 2013;35:838-48.
12. Tavakol M, Doody GA. Making students' marks fair: standard setting, assessment items and post hoc item analysis. Int J Med Educ 2015;6:e838-9.
13. Tor E, Steketee C. Rasch analysis on OSCE data : An illustrative example. Australas Med J 2011;4:339-45.
14. Boone W, Staver J, Yale M. Rasch Analysis in the Human Sciences. In: Boone W, Staver J, Yale M, eds. Chapter 20. Germany: Springer Science & Business Media, 2013:482.
15. Core Committee, Institute for International Medical Education. Global minimum essential requirements in medical education. Med Teach 2002;24:130-5.
16. Bandaranayake RC. Setting and maintaining standards in multiple choice examinations: AMEE Guide No. 37. Med Teach 2008;30:836-45.
17. Norcini JJ, Shea JA, Hancock EW, Webster GD BR. A criterion-referenced examination in cardiovascular disease? Med Educ 1988;22:32-9.
18. 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 Heal Sci Educ 2006;11:115-22.
19. Liu M, Liu KM. Setting Pass Scores for Clinical Skills Assessment. Kaohsiung J Med Sci 2008;24:656-63.

20. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22:276-82.
21. Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: A review of metrics-AMEE guide no. 49. Med Teach 2010;32:802-11.
22. Shulruf B, Damodaran A, Jones P, Kennedy S, Mangos G, O'Sullivan AJ, et al. Enhancing the defensibility of examiners' marks in high stake OSCEs. BMC Med Educ 2018;18:10.
23. Yousuf N, Violato C, Zuberi RW. Standard Setting Methods for Pass/Fail Decisions on High-Stakes Objective Structured Clinical Examinations: A Validity Study. Teach Learn Med 2015;27:280-91.
24. Zieky M, Perie M, Livingston S. A Primer on Setting Cut Scores on Tests of Educational Achievement Excerpts From Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Prince- ton, NJ: Educational Testing Service, Inc, 2006
25. Schoonheim-Klein M, Muijtens A, Habets L, Manogue M, Van Der Vleuten C, Hoogstraten J, et al. On the reliability of a dental OSCE, using SEM: effect of different days. Eur J Dent Educ 2008;12: 131-7.
26. Cusimano MD. Standard setting in medical education. Acad Med 1996;71: S112-20.
27. Albanese M. Rating educational quality: factors in the erosion of professional standards. Acad Med 1999;74:652-8.
28. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences Second Ed. In: Bond TG, Fox CM, eds. Chapter 12. New Jersey: Lawrence Erlbaum Associates, 2007:355.
29. Harrison C. Feedback in the context of high-stakes assessment: Can summative be formative? [dissertation]. Maastricht: Maastricht University.; 2017.
30. Kalet A, Tewksbury L, Ogilvie J, Buckvar-Keltz L, Porter B, Yingling S. Remediation of Learners Who Perform Poorly on an OSCE. In: Zabar S, Kachur E, Kalet A, Hanley K, eds. Objective Structured Clinical Examinations. New York, NY: Springer New York, 2012:35-8.