A collaborative comparison of Objective Structured Clinical Examination (OSCE) standard setting methods at Australian medical schools

Bunmi S. Malau-Aduli[a], Peta-Ann Teague[a], Karen D'Souza[b], Clare Heal[a], Richard Turner[c], David Garne[d], Cees van der Vleuten[e]

[a]College of Medicine and Dentistry, James Cook University, Queensland, Australia
[b]School of Medicine, Deakin University, Victoria, Australia
[c]School of Medicine, University of Tasmania, Tasmania, Australia
[d]School of Medicine, University of Wollongong, New South Wales, Australia
[e]School of Health Professions Education, Maastricht University, Maastricht, Netherlands

Corresponding Author: Bunmi Malau-Aduli, College of Medicine and Dentistry, Division of Tropical Health and Medicine, James Cook University, Townsville, Australia

Tel: +61747814418; Fax: +6174781 5870

E-mail: bunmi.malauaduli@jcu.edu.au

Abstract

Background: A key issue underpinning the usefulness of the OSCE assessment to medical education is standard-setting, but the majority of standard-setting methods remain challenging for performance assessment because they produce varying passing marks. Several studies have compared standard setting methods; however, most of these studies are limited by their experimental scope, or use data on examinee performance at a single OSCE station or from a single medical school. This collaborative study between ten Australian medical schools investigated the effect of standard-setting methods on OSCE cut scores and failure rates.

Methods: This research used 5,256 examinee scores from seven shared OSCE stations to calculate cut scores and failure rates using two different compromise standard-setting methods, namely the Borderline Regression and Cohen’s methods.

Results: The results of this study indicate that Cohen’s method yields similar outcomes to the Borderline Regression method, particularly for large examinee cohort sizes. However, with lower examinee numbers on a station, the Borderline Regression method resulted in higher cut scores and larger difference margins in the failure rates.

Conclusion: Cohen’s method yields similar outcomes as the Borderline Regression method and its application for benchmarking purposes and in resource-limited settings is justifiable, particularly with large examinee numbers.

Introduction

Objective Structured Clinical Examinations (OSCEs) are used by many health professional courses, especially medical schools, to assess examinee clinical competence. To achieve this, OSCEs generally expose examinees to predetermined role-played medical scenarios featuring simulated patients (SPs), while examiners observe and assess examinees based on their interactions with the SPs (Keely et al. 2002; Hodges and McIlroy 2003). Examinee performance in OSCE stations provide a systematic means to assess their acquired clinical skill sets vital to their successful completion of their medical course and throughout their future careers (Hodges and McIlroy 2003; Payne et al. 2008).

A key issue underpinning the usefulness of OSCE assessment to medical education is standard setting, which is used to determine the minimum standard or passing mark required for successful examinee performance and subsequent progression within the medical course (Wass et al. 2001). Hence, examinee assessment and clinical competency outcomes are highly reliant on the method selected to calculate this minimum standard (Cusimano 1996). Presently, several standard-setting methods have been developed (Ben-David 2000; Norcini 2003; Barman 2008; Downing and Yudkowsky 2009; Cizek 2012).

Standard-setting methods must be transparent, reproducible, credible, feasible, and justifiable (Kaufman et al. 2000; Wass et al. 2001; Humphrey-Murto and MacFadyen 2002). Other major considerations in choosing an appropriate standard setting method are time and available resources and expertise. It is important to align the time needed to implement a method with the needs and resources of the testing program (Hambleton et al. 2012). However, while the majority of standard-setting methods meet most of these criteria, they remain challenging for performance assessment because they still produce varying passing marks (Humphrey-

Murto and MacFadyen 2002; Boursicot et al. 2006; George et al. 2006; Wood et al. 2006), therefore indicating that there is no single best method or gold standard.

Fundamentally, a standard-setting method should deliver a true representation of examinee performance; hence, only clinically competent examinees should pass an OSCE assessment. There are three major types of standard setting categories, namely criterion-referenced (absolute), norm-referenced (relative) and compromise methods (Livingston and Zieky 1982; Cizek 1996; Norcini 2003; Cizek 2012). Relative standards identify a group of passing and failing examinees relative to pre-determined passing scores without considering the difficulty of the test or ability of the examinees (Cohen-Schotanus and van der Vleuten 2010). Relative standard setting methods are easy to set but less defensible because the two important factors (test difficulty and examinee ability) that could affect the passing scores are not considered (McKinley and Norcini, 2013). Hence, the absolute method has been preferred for testing clinical competencies (Norcini, 2003). Absolute standards are based on a pre-determined level of competency that does not depend on the performance of a well-defined group (Downing and Yudkowsky, 2009). These methods require a desired level of mastery and the passing criteria are determined from the judgments of a group of subject matter experts. Absolute standard setting methods are either test-centred or examinee-centred (Livingston and Zieky 1982). Test-centred standards are based on exam content; examples include the Nedelsky (1954), Angoff (1971), Ebel (1972) and Jaeger (1983) methods. Examinee-centred methods, on the other hand, focus judgement on the examinee performance and not the test content; examples include the contrasting groups, borderline group, and borderline regression methods (Livingston and Zieky 1982; Wood et al. 2006). The compromise method combines both test- and examinee-centred methods; examples include the Hofstee (1983) and Cohen’s (2010) methods.

Providing a detailed description of these standard setting methods is beyond the scope of this manuscript, however this information is widely available in the medical education literature (Livingstone and Zieky 1982; Cizek 1996; 2012; McKinley and Norcini 2013). The test-centred methods are used widely in large-scale assessment and have been shown to provide reliable and valid cut-scores (McKinley et al. 2005). However, they assume an underlying unidimensional structure which cannot be assumed in the case of the OSCE. Additionally, they are cumbersome and time-consuming (Hambleton et al. 2012). Conversely, examinee-centred methods are more commonly seen in the medical education literature in setting cut-scores for OSCEs (Boulet et al. 2003; Kramer et al. 2003; McKinley et al. 2005; Boursicot et al. 2007). The test format of performance assessments such as the OSCE necessitates the use of methods that consider examinees' complete score profile.

The borderline regression method (BRM) has been identified as superior to the modified borderline group method. This is due to the BRM utilising all examinee scores to calculate the pass mark rather than just those examinee scores ranked as borderline (Ben-David 2000; Wood et al. 2006). This standard-setting method has been deemed preferential to other methods due to its ability to be derived immediately after the conclusion of the OSCE and its high validity in representing actual examinee performance (Humphrey-Murto and MacFadyen 2002; Wood et al. 2006). The BRM has been successfully validated (Kaufman et al. 2000, Kramer et al. 2003); its superiority rests not just in its ability to set standards quickly, but in its use of all examinee/assessor interaction at station and scoring form level to both determine the standard and provide detailed station level quality metrics for diagnostic processes. However, little is known in the OSCE literature about another simple and cost-effective method, Cohen's method.

Cohen's method was developed by Janke Cohen-Schotanus in 2010 and it is based on the best cohort of examinees' performance. It assumes that fluctuations in examinee performance reflect test difficulty or teaching quality and it uses the 65% of 95th percentile examinee as the reference point for the passing mark (Cohen-Schotanus and van der Vleuten 2010; Taylor 2011). Cohen's method has principally been applied to standards in knowledge tests; however, it has also been previously used in the OSCE setting (Kaufman et al. 2000) but the findings were inconclusive. According to Taylor (2011), the score of the 95% percentile examinee is an accurate indicator of exam difficulty and is consistent over time. Paradoxically, Cohen's method is considered limiting because of its intrinsic reliance on ranking examinee performance and using this rank to determine pass or fail rather than actual examinees' clinical competency (Barman 2008).

Several studies have compared standard setting methods in OSCEs (Kaufman et al. 2000; Humphry-Murto and MacFayden 2002; Boursicot et al. 2007). However, to the best of our knowledge, most comparative standard-setting studies are limited by their experimental scope, using examinee performance at a single OSCE station or from a single medical school.

This research was undertaken as a collaborative study between ten Australian medical schools. The study compared the outcome of two compromise standard-setting methods (Borderline Regression and Cohen's methods) on examinee performance in seven shared OSCE stations used to assess clinical competence in the early and exit phases of clinical exams. The study aimed to answer the research question – to what extent do the cut scores and failure rates from both standard setting methods differ?

Methods

Sample

Ten geographically dispersed Australian medical schools participated in this collaborative study by sharing OSCE stations which were co-developed by an expert committee.

The collaborative project is known as the Australian Collaboration for Clinical Assessment in Medicine (ACCLAiM). All schools have similar horizontally and vertically integrated outcomes-based curricula, accredited by the same body (the Australian Medical Council). The selected year groups (early clinical and exit level) were chosen because of their comparable levels of intended learning outcomes.

OSCE stations

There were two phases of the collaboration in which a total of seven OSCE stations were collaboratively developed, and after achieving consensus on content and marking criteria, were incorporated into the 2015-2016 summative clinical examination cycle of the participating schools. The first phase of this study focused on the early clinical exam in which three OSCE stations were co-developed and used by eight participating schools. The second phase of the study focused on the exit clinical exam in which four OSCE stations were co-developed and used by all ten participating schools. The examination procedure was similar for both phases. The core clinical competencies assessed by all seven stations were selected from prospectively reviewed clinical blueprints of the specific clinical skills and medical problems, representing a fair and reasonable assessment, and mapped to the Medical Deans of Australia and New Zealand (MDANZ) medical competencies project (MDANZ 2014). The OSCE stations focused on clinical reasoning, communication skills, risk assessment, investigation and management plan.

Examination procedure

Each collaborative set of OSCE stations was embedded into each participating medical school’s OSCEs, where the total number of live stations varied between ten and twelve, with two to three rest stations, and a time of 10 minutes allocated for each station. The stations were used by the collaborating schools as deemed suitable to their curriculum. Participating schools were required to use at least two (2) stations per exam. The participating schools arranged the shared station ‘paperwork’ to fit with their local practice, to ensure that the shared OSCE stations appeared identical to the local medical school stations. Due to large numbers of examinees, concurrent multiple circuits of each station were used at each school. All schools had one internal local examiner per station who were experienced clinicians involved in examinee teaching and examination. Examiners rated examinee performance on each OSCE station using anchored checklists with descriptors for five performance categories (fail, borderline fail, borderline pass, clear pass, exceptional). They also gave global ratings of overall station performance using a 7-point Likert scale (where 1=very poor performance, 2=well short of expected standard, 3=short of expected standard, 4=expected standard, 5=better than expected standard, 6=much better than expected standard, 7=exceptional performance). Given that the study involved multi-institutional collaboration, a 7-point global rating scale was used to provide the optimum number of categories that would allow for increased reliability of ratings and minimise response error across sites (Cox 1980; Weijters et al. 2010). Additionally, to improve agreement between our raters, the global scales are behaviourally anchored with explicit performance category descriptors. The examiners were also provided with a calibration exercise specific to each station, during which they were able to become familiar with using the global rating. Details of the ACCLAiM examination procedure and protocols have been described in our previously published work (Malau-Aduli et al. 2012; Malau-Aduli et al. 2016).

Standard-setting methods

This research used two standard-setting methods to compute passing scores for each shared OSCE station: the Borderline Regression (BRM) and Cohen's methods. For comparison, standard setting procedures for the two methods were applied to the examinee scores for each OSCE station and their differential effects on cut scores and failure rates were determined.

For the BRM (Wood et al. 2006), we used linear regression analysis of examinee performance, as total percentage scores, and examiner global rankings to determine the cut score. The cut score was derived by substituting the value for the borderline examinee (3.5) into the regression equation.

For Cohen's method (Cohen-Schotanus and van der Vleuten 2010), we used the performance of the top 95th percentile of the test scores as the benchmark and the cut score was set as 65% of the 95th percentile with the formula: $CS = K \times P_{95}$ - where CS is the cut score, K is the multiplier (0.65) and P_{95} the score of the examinee at the 95th percentile (Cohen-Schotanus and van der Vleuten 2010; Taylor 2011).

Results

A total of 5,256 examinee scores, distributed between the seven shared OSCE stations were analysed in this study. The demographic profile of the participating examinees showed that their mean age was 25.2 ± 4.7 years; 52% were females and 86% were domestic students.

Figure 1 shows a comparison of cut scores and failure rates as determined by the two standard setting methods for the early clinical OSCE stations. Cohen's method resulted in higher cut scores and failure rates ($p<0.01$) than the BRM for two of the three stations used for the early clinical exam (Table 1). The BRM yielded the highest cut score (63%) and failure rate (32%) for station 2, which had the lowest number of examinees ($n=511$) in the early clinical exam (Fig 1). A similar pattern was observed in the exit exam (Fig 2), where station 3 had the lowest number of examinees ($n=324$) and the BRM resulted in the highest cut score (66%) and failure rate (37%). For the exit exam, both BRM and Cohen's method yielded the same cut scores and failure rates for two of the four stations that were used (stations 1 and 2), and these stations had the highest number of examinees, 805 and 995 respectively.

Table 1 also portrays the difference in fail decisions between the two standard setting methods for each OSCE station. Cohen's method generally resulted in a higher failure rate if the examinee numbers were high. However, with lower examinee numbers on a station, the BRM resulted in higher cut scores and larger difference margins in the fail decisions. In the early clinical exam, there were 10.4% more fails on station 2 ($n=511$; $p<0.01$)) with the BRM, while station 3 had only 5.4% more fails with Cohen's method and the observed difference was not significant. For the exit exam, there were 16% ($p<0.001$) more fails with the BRM than with Cohen's method on station 3, which had the lowest numbers of examinees. Cohen's method yielded 10.6% ($p<0.0001$) more fails on station 4 than the BRM. Additionally, there were significant differences ($p<0.001$) in passing rates between schools for both methods. Table 2 portrays the impact of the Borderline Regression and Cohen's methods for borderline examinees on each OSCE station. Overall, there were higher correlations between both standard setting methods, in the pass-fail decisions for the borderline group on stations with higher examinee numbers. There were no observable associations between the assessed competencies and pass-fail decisions for both methods. Nevertheless, future research could explore this in more detail.

Discussion

Validation of standard-setting methods has been a major part of quality assuring assessment processes in medical education. This is to ensure fair representation of actual examinee competence levels while providing an objective and defensible outcome (Kaufman et al. 2000; Humphrey-Murto and MacFadyen 2002). Holistic standard-setting methods with a set arbitrary passing mark (usually 50 or 60%) are considered inappropriate for OSCE assessment as high failure rates can arise with their rigid application and arbitrary adjustments of passing marks (Kaufman et al. 2000; George et al. 2006).

This research shows that selection of a standard-setting method has potentially severe implications on perceived examinee performance. It has demonstrated that examinees could pass or fail OSCE assessments with the same performance dependent on the standard-setting method, supporting the findings of Boursicot et al (2006). This has implications affecting comparability and benchmarking of examinees' clinical competence between medical schools and OSCE stations using different standard-setting methods. The observed differences in passing rates across schools triangulates with published work in other settings (Boursicot et al. 2006; 2007) and this may imply differences in competence levels of examinees at different schools. It also highlights the importance of case specificity. Consequently, selection of standard-setting method should be a research-informed decision.

Nonetheless, the results of this study indicate that Cohen’s method results in similar performance outcomes as the BRM, especially with large cohort sizes. According to Taylor (2011), the score of the 95% percentile examinee is an accurate indicator of exam difficulty and is consistent over time. These could be used to explain the similarities observed between the cut scores and failure rates obtained for both BRM and Cohen’s methods at most of the OSCE stations in this research. Both methods use all examinee data in setting the passing standard and this provides a fair representation of examinees’ performance on the whole exam. The similar pass-fail outcomes obtained for borderline examinees, particularly in cases with large total examinee numbers, further confirms the utility of Cohen’s method for OSCEs. Our findings suggest that the stability of the cutscore across the two standard setting methods is dependent on the number of examinees. Cohen’s method relies on the performance of the examinees in the higher cohort quartile and our results imply that with more examinee numbers in the higher performance pool, the error margin shrinks, resulting in reduced heterogenity of variance and therefore allowing for better correlations. Based on our findings, we would recommend the use of Cohen’s standard setting method for multi-institutional OSCEs, where total examinee numbers are over 800. The concordance between the two methods is encouraging, providing some level of reassurance that the less resource-intensive Cohen’s method may be implemented with high confidence, particularly for multi-site benchmarking of clinical performance, and also in resource limited settings. The BRM has an added advantage of providing conceptual assessment of the examinees’ clinical competence levels. However, the examinee ranking based on Cohen’s method can also be credible and with similar logical outcomes as the BRM.

The recent global call by licensing bodies for the development of national frameworks for standard setting of assessment in medical schools emphasises the need for benchmarking of examinee performance across multiple sites and institutions (Wilkinson et al. 2014). Cohen's method has been implemented with great success in other educational settings (Dochy et al. 2009; Cohen-Schotanus and van der Vleuten 2010; Taylor 2011) and the fact that this method uses the top performing examinees as the reference point ensures valid and accurate cut scores because this top cohort of examinees is usually stable and performs equally well between different year groups. The resulting similar cut scores and identification of failing examinees for Cohen's and BRM, particularly with larger examinee numbers, demonstrate that Cohen's method provides equally feasible and reproducible outcomes as the BRM. However, Cohen's method has the additional benefit that it is less cumbersome, requires shorter time and less resources for its implementation and validation of examinee performance across multiple sites and institutions. This standard setting approach is therefore recommended as a justifiable standard setting method.

Research Strengths and Limitations

This research is the first study that compares the BRM and Cohen's standard setting approaches and also examines the outcome of standard setting methods across multiple institutions, with large numbers of examinees. However, generalisations of the findings to other settings are limited primarily from the use of data collected over only a single year of study. In addition, whilst all attempts have been made to control for minor local differences in the delivery of shared OSCE stations, these may have affected the obtained examinee performance scores. Furthermore, generalizability of these results may be limited by the sample, OSCEs, medical curricula, and assessments. Variability of these factors at other institutions may produce different cuts. Practitioners should consider these factors in the decisions they make based on those cuts.

Conclusion

Standard-setting methods have a profound effect in determining examinees' clinical competency and subsequent pass-fail decisions in OSCE assessments. However, this research demonstrates that with higher examinee numbers, resultant pass-fail decisions are very similar for BRM and Cohen's methods. Future research using broad-scale comparisons in other settings could be used to complement these research findings.

Acknowledgements

The authors would like to thank all participating ACCLAiM medical schools and their administrative and academic staff, who supported the organisation and implementation of the examinations.

Declaration of interest: The authors report no conflicts of interest.

Ethical approval: All participating schools obtained ethics approval from their local Ethics Committee. All information was de-identified before data analysis.

Practice points

- Standard setting method affects pass-fail decisions.
- With higher examinee numbers, resultant pass-fail decisions are very similar for Cohen’s and BRM.
- With similar outcomes, Cohen’s method has the attraction of using less resources

Glossary

Standard setting: Standard setting is the process of defining or judging the level of knowledge and skill required to meet a typical level of performance and then identifying a score on the examination score scale that corresponds to that performance standard

Relative Standards: Standards that are established based on a comparison of those who take the assessment to each other are relative standards.

Absolute Standards: Standards set by determining the amount of test material that must be answered (or performed) correctly in order to pass are absolute standards

Reference: McKinley, Danette W., and John J. Norcini. "How to set standards on performance-based examinations: AMEE Guide No. 85." Medical teacher 36.2 (2014): 97-110.

Notes on contributors

BUNMI MALAU-ADULI, BSc, MSc, PhD, is a Senior Lecturer in Medical Education and the Academic Lead for Assessment and Evaluation at the College of Medicine and Dentistry, James Cook University

PETA-ANN TEAGUE, MBChB, DRCOG, MRCGP, Dip Med Ed, FRACGP, is an Associate Professor and the Director of the Generalist Medical Training (GMT) Program at James Cook University

KAREN D'SOUZA, MBBS(Hons), is a Senior Lecturer in Medical Education (Clinical Skills) and the Coordinator, Doctor and Patient Theme for the School of Medicine at Deakin University

CLARE HEAL, MBChB DRACOG, FRACGP, MPHandTM, Dip GU Med, PhD, is a Professor of General Practice and Rural Medicine for James Cook University in Mackay

RICHARD TURNER, MBBS, BMedSc, FRACS, is a Professor of Surgery and the Director of the Hobart Clinical School at the School of Medicine, University of Tasmania

DAVID GARNE, MBChB, MSC, MPhil, is an Associate Professor and the Associate Dean of Community, Primary, Remote and Rural Health at the School of Medicine at the University of Wollongong

CEES VAN DER VLEUTEN, MA, PhD, is the Scientific Director of the Graduate School of Health Professions Education at Maastricht University

References

Barman A. 2008. Standard setting in examinee assessment: is a defensible method yet to come? Ann Acad Med Singapore 37 (11):957-63.

Ben-David MF. 2000. AMEE Guide No. 18: Standard setting in examinee assessment. Med Teach. 22 (2):120-130.

Boulet JR, De Champlain AF, McKinley DW. 2003. Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach. 25 (3):245-249.

Boursicot KAM, Roberts TE, Pell G. 2006. Standard setting for clinical competence at graduation from medical school: A comparison of passing scores across five medical schools. Adv Health Sci Educ. 11 (2):173-183. doi: 10.1007/s10459-005-5291-8.

Boursicot KAM, Roberts TE, Pell G. 2007. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ. 41 (11):1024-1031. doi: 10.1111/j.1365-2923.2007.02857.x

Cizek GJ. 2012. An introduction to contemporary standard setting. In: Setting performance standards: Foundations, methods, and innovations. 2nd Ed; pp 3-14. New York: Routledge.

Cizek, GJ. 1996. Setting passing scores. Educational Measurement: Issues and Practice 15. doi: 10.1111/j.1745-3992.1996.tb00809.x.

Cohen-Schotanus J, van der Vleuten CPM. 2010. A standard setting method with the best performing examinees as point of reference: practical and affordable. Med Teach. 32 (2):154-160.

Cox EP. 1980. The optimal number of response alternatives for a scale: a review. J Marketing Res, 17, 407–422

Cusimano, MD. 1996. Standard setting in medical education. Acad Med. 71 (10):S112-S120. doi: 10.1097/00001888-199610000-00062.

Dochy F, Kyndt E, Baeten M, Pottier S, Veestraeten M. 2009. The effects of different standard setting methods and the composition of borderline groups: A study within a law curriculum. Studies in Educational Evaluation 35 (4):174-182.

Downing SM, Yudkowsky R. 2009. Assessment in health professions education: Routledge.

George, Sanju, M Sayeed Haque, and Femi Oyebode. 2006. Standard setting: comparison of two methods. BMC Med Educ. 6 (1):46.

Hambleton, RONALD K, MARY J Pitoniak, and JENNA M Copella. 2012. "Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results." Setting performance standards: Foundations, methods, and innovations:47-76.

Hodges B, McIlroy JH. 2003. Analytic global OSCE ratings are sensitive to level of training. Med Educ. 37 (11):1012-1016. doi: 10.1046/j.1365-2923.2003.01674.x.

Humphrey-Murto S, MacFadyen JC. 2002. Standard setting: A comparison of case-author and modified borderline-group methods in a small-scale OSCE. Acad Med. 77 (7):729-732. doi: 10.1097/00001888-200207000-00019.

Kaufman DM, Mann KV, Muijtjens AMM, van der Vleuten CPM. 2000. A comparison of standard setting procedures for an OSCE in undergraduate medical education. Acad Med. 75. doi: 10.1097/00001888-200003000-00018.

Keely E, Myers K, Dojeiji S. 2002. Can Written Communication Skills Be Tested in an Objective Structured Clinical Examination Format? Acad Med. 77 (1):82-86.

Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten CPM. 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. Med Educ. 37. doi: 10.1046/j.1365-2923.2003.01429.x.

Livingstone SA, Zieky MJ. 1982. Passing scores: A manual for setting standards of performance on educational and occupational tests. Princeton: Educational Testing Services.

Malau-Aduli BS, Mulcahy S, Warnecke E, Otahal P, Teague PA, Turner R, Van der Vleuten C. 2012. Inter-rater reliability: Comparison of checklist and global scoring for OSCEs. Creative Educ J, Special Issue, 3: 937-942. DOI:10.4236/ce.2012.326142

Malau-Aduli BS, Teague PA, Turner R, Holman B, D'souza K, Garne D, Heal C, Heggarty P, Hudson JN, Wilson IG. 2016. Improving assessment practice through cross-institutional collaboration: An exercise on the use of OSCEs. Med Teach. 38 (3):263-271.

McKinley DW, Boulet JR, Hambleton RK. 2005. A work-centered approach for setting passing scores on performance-based assessments. Evaluation and the Health Pofessions 28 (3):349-369.

McKinley DW, Norcini JJ. 2013. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 36 (2):97-110.

Norcini JJ. 2003. Setting standards on educational tests. Med Educ. 37. doi: 10.1046/j.1365-2923.2003.01495.x.

Payne NJ, Bradley EB, Heald EB, Maughan KL, Michaelsen VE, Wang X, Corbett EC Jr. 2008. Sharpening the Eye of the OSCE with Critical Action Analysis. Acad Med. 83 (10):900-905. doi: 10.1097/ACM.0b013e3181850990.

Taylor CA. 2011. Development of a modified Cohen method of standard setting. Med Teach. 33 (12):e678-82. doi: 10.3109/0142159x.2011.611192.

Wass V, van der Vleuten CPM, Shatzer J, Jones R. 2001. Assessment of clinical competence. Lancet 357 (9260):945-9. doi: 10.1016/s0140-6736(00)04221-5.

Weijters B, Cabooter E, Schillewaert N. 2010. The effect of rating scale format on response styles: The number of response categories and response category labels. Inter J Res Marketing 27(3): 236–247.

Wilkinson D, Schafer J, Hewett D, Eley D, Swanson D. 2014. Global benchmarking of medical examinee learning outcomes? Implementation and pilot results of the International Foundations of Medicine Clinical Sciences Exam at The University of Queensland, Australia. Med Teach 36 (1):62-67.

Wood TJ, Humphrey-Murto SM, Norman GR. 2006. Standard setting in a small scale OSCE: A comparison of the modified borderline-group method and the borderline regression method. Adv Health Sci Educ. 11 (2):115-122. doi: 10.1007/s10459-005-7853-1.

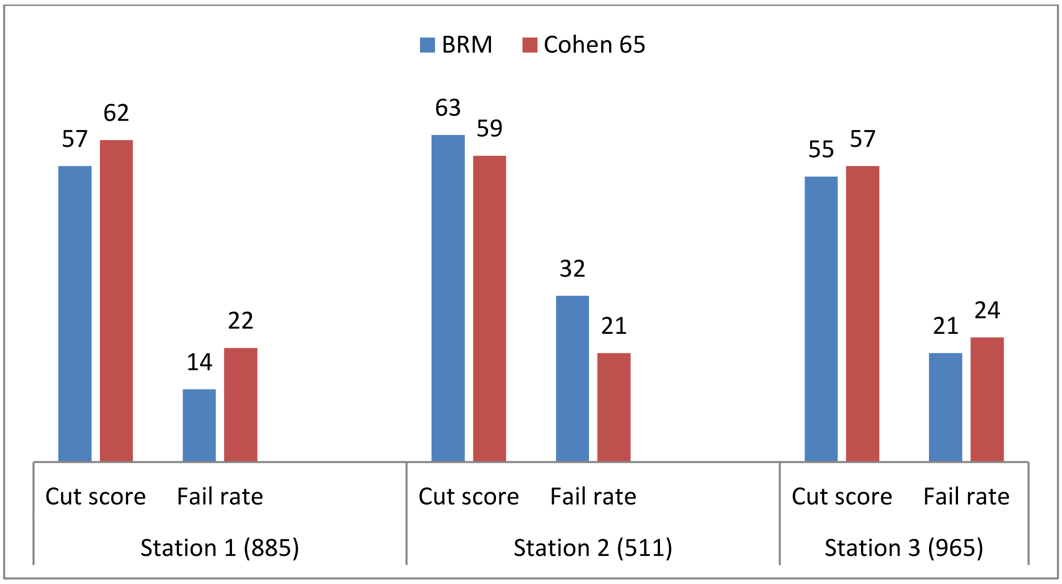

Figure 1: Comparison of cut score (%) and failure rate (%) for early clinical OSCE stations. The number of examinees included is shown in parentheses.

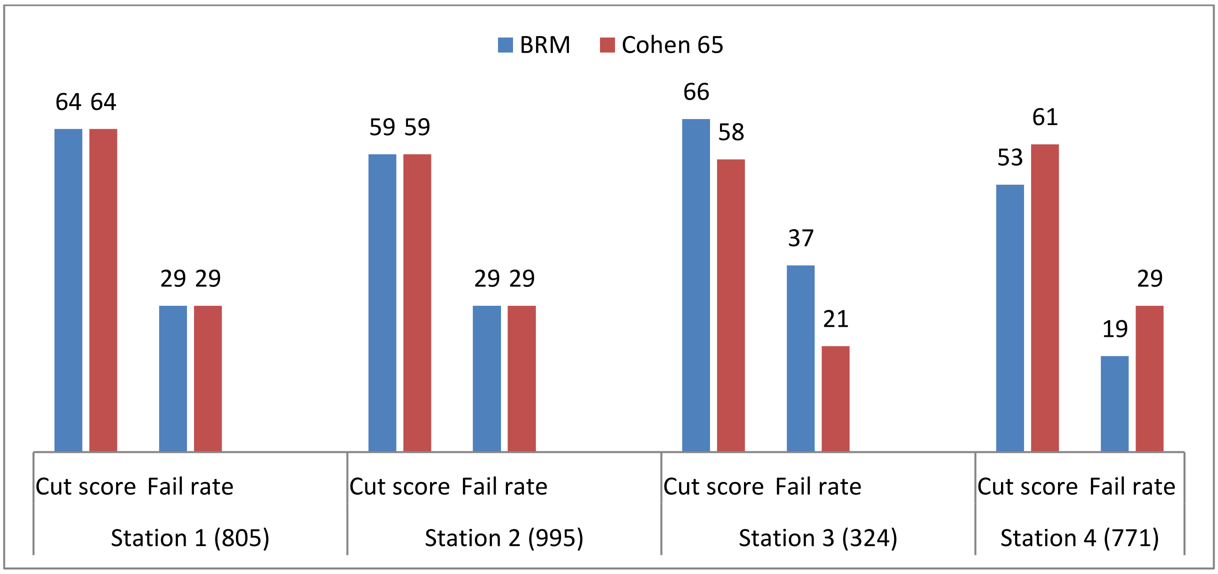

Figure 2: Comparison of cut score (%) and failure rate (%) for exit clinical OSCE stations. The number of examinees included is shown in parentheses.

Table 1: Differences in standards between the Borderline Regression and Cohen's methods for each OSCE station

OSCE Station	Major Competency Assessed	Number of Examinees	Mean OSCE Score ± STDev	BRM		Cohen's		Pass-Fail Decision	p-value
				Cut Score	No of Examinees who failed	Cut Score	No of Examinees who failed		
	Early Clinical Exam								
Station 1	Interpretation of Relevant Investigation	885	74.7±15.2	57	120	62	197	8.7% more fails with the Cohen method	0.0001
Station 2	Clinical Reasoning	511	68±13.2	63	161	59	108	10.4% more fails with the BR method	0.01
Station 3	Clinical Reasoning	965	66.9±13.6	55	205	57	232	5.4% more fails with the Cohen method	0.2
	Exit Exam								
Station 1	Communication Skills	805	71.1±16.2	64	233	64	233	Same number of fails	-
Station 2	Clinical Reasoning	995	66.6±14.8	59	285	59	285	Same number of fails	-
Station 3	Investigation and Management Plan	324	69.4±13.6	66	119	58	67	16% more fails with the BR method	0.001
Station 4	Suicide Risk Assessment	771	69.8±17.0	53	142	61	224	10.6% more fails with the Cohen method	0.0001

Table 2: Impact of the Borderline Regression and Cohen's methods for borderline examinees on each OSCE station

OSCE Station	Total No of Examinees	No of Borderline Examinees*	BRM		Cohen's		Difference between BRM and Cohen's Borderline Fails
			Cut Score	No of BRM Fails	Cut Score	No of Cohen Fails	
Early Clinical Exam							
Station 1	885	136	57	72	62	99	27
Station 2	511	241	63	147	59	88	59
Station 3	965	324	55	168	57	177	9
Exit Exam							
Station 1	805	301	64	189	64	189	0
Station 2	995	438	59	232	59	232	0
Station 3	324	154	66	102	58	59	43
Station 4	771	149	53	77	61	105	28

*number of examinees who were awarded global rating score 3 or 4