A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools

Citation for published version (APA):

Malau-Aduli, B. S., Teague, P-A., D'Souza, K., Heal, C., Turner, R., Garne, D. L., & van der Vleuten, C. (2017). A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools. *Medical Teacher*, 39(12), 1261-1267.
<https://doi.org/10.1080/0142159X.2017.1372565>

Document status and date:

Published: 01/01/2017

DOI:

[10.1080/0142159X.2017.1372565](https://doi.org/10.1080/0142159X.2017.1372565)

Document Version:

Publisher's PDF, also known as Version of record

Document license:

Taverne

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Download date: 26 Apr. 2023

A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools

Bunmi Sherifat Malau-Aduli, Peta-Ann Teague, Karen D'Souza, Clare Heal, Richard Turner, David L. Garne & Cees van der Vleuten

To cite this article: Bunmi Sherifat Malau-Aduli, Peta-Ann Teague, Karen D'Souza, Clare Heal, Richard Turner, David L. Garne & Cees van der Vleuten (2017) A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools, Medical Teacher, 39:12, 1261-1267, DOI: [10.1080/0142159X.2017.1372565](https://doi.org/10.1080/0142159X.2017.1372565)

To link to this article: <https://doi.org/10.1080/0142159X.2017.1372565>

Published online: 22 Sep 2017.

Submit your article to this journal [↗](#)

Article views: 1093

View related articles [↗](#)

View Crossmark data [↗](#)

Citing articles: 4 View citing articles [↗](#)

A collaborative comparison of objective structured clinical examination (OSCE) standard setting methods at Australian medical schools

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

^aCollege of Medicine and Dentistry, James Cook University, Townsville, Australia; ^bSchool of Medicine, Deakin University, Geelong, Australia; ^cSchool of Medicine, University of Tasmania, Hobart, Australia; ^dSchool of Medicine, University of Wollongong, Wollongong, Australia; ^eSchool of Health Professions Education, Maastricht University, Maastricht, The Netherlands

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 10 Australian medical schools investigated the effect of standard-setting methods on OSCE cut scores and failure rates.

Methods: This research used 5256 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

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.

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 (Livingstone and Zieky 1982; Cizek 1996, 2012; Norcini 2003). Relative standards identify a group of passing and failing examinees relative to predetermined 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 predetermined 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-centered or examinee-centered (Livingstone and Zieky 1982). Test-centered standards are based on exam content; examples include the Nedelsky (1954), Angoff (1971), Ebel (1972) and Jaeger (1982) methods. Examinee-centered methods, on the other hand, focus judgment on the examinee performance and not the test content; examples include the contrasting groups, borderline group and borderline regression methods (Livingstone and Zieky 1982; Wood et al. 2006). The compromise method combines both test- and examinee-centered methods; examples include the Hofstee (1983) and Cohen-Schotanus and van der Vleuten (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-centered 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 that cannot be assumed in the case of the OSCE. Additionally, they are cumbersome and time-consuming (Hambleton et al. 2012). Conversely, examinee-centered 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 utilizing 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; Humphrey-Murto and MacFadyen 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 codeveloped and used by eight participating schools. The second phase of the study focused on the exit clinical examination in which four OSCE stations were codeveloped and used by all 10 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 10 and 12, with two to three rest stations, and a time of 10 min 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 examination. 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 minimize response error across sites (Cox 1980; Weijters et al. 2010). Additionally, to improve agreement between our raters, the global scales are behaviorally 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, 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 5256 examinee scores, distributed between the seven shared OSCE stations were analyzed 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 examination (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 (Figure 1). A similar pattern was observed in the exit examination (Figure 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

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 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 examination							
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 examination							
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.

the BRM, especially with large cohort sizes. According to Taylor (2011), the score of the 95% 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 examination. 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 cut score 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 heterogeneity 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 emphasizes 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, generalizations 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.

Conclusions

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 organization and implementation of the examinations.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Ethical approval

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

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, D.W., and Norcini, J.J. "How to set standards on performance-based examinations: AMEE Guide No. 85." *Medical Teacher* 2014. 36,2: 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, MPH and TM, 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

- Angoff WH. 1971. Scales, norms, and equivalent scores. In: Educational measurement. Washington DC: American Council on Education; p. 514–515.
- Barman A. 2008. Standard setting in student assessment: is a defensible method yet to come? *Ann Acad Med Singapore*. 37:957–963.
- Ben-David MF. 2000. AMEE Guide No. 18: standard setting in examinee assessment. *Med Teach*. 22:120–130.
- Boulet JR, De Champlain AF, McKinley DW. 2003. Setting defensible performance standards on OSCEs and standardized patient examinations. *Med Teach*. 25:245–249.
- 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 Theory Pract*. 11:173–183.
- Boursicot KAM, Roberts TE, Pell G. 2007. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Med Educ*. 41:1024–1031.
- Cizek GJ. 2012. An introduction to contemporary standard setting. In: Setting performance standards: foundations, methods, and innovations. 2nd ed. New York: Routledge; p. 3–14.
- Cizek GJ. 1996. Setting passing scores. *Educ Measure Issues Pract*. 15:20–31.
- 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*. 32:154–160.
- 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:S112–S120.
- 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. *Stud Educ Eval*. 35:174–182.
- Downing SM, Yudkowsky R. 2009. Assessment in health professions education. New York: Routledge.
- Ebel RL. 1972. Essentials of educational measurement. Englewood Cliffs (NJ): Prentice Hall; p. 492–494.
- George S, Haque MS, Oyeboode F. 2006. Standard setting: comparison of two methods. *BMC Med Educ*. 6:46.
- Hambleton RK, Pitoniak MJ, Copella JM. 2012. Essential steps in setting performance standards on educational tests and strategies for assessing the reliability of results. In: Setting performance standards: Foundations, methods, and innovations. New York: Routledge; p. 47–76.
- Hodges B, McIlroy JH. 2003. Analytic global OSCE ratings are sensitive to level of training. *Med Educ*. 37:1012–1016.
- Hofstee WKB. 1983. The case for compromise in educational selection and grading. In: On educational testing. Washington DC: Jossey Bass; p. 109–127.
- 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*. 77:729–732.
- Jaeger RM. 1982. An iterative structured judgment process for establishing standards on competency tests: theory and application. *Educ Eval Policy Anal*. 4:461–476.
- 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*. 75:267–271.

- Keely E, Myers K, Dojeiji S. 2002. Can written communication skills be tested in an objective structured clinical examination format? *Acad Med.* 77:82–86.
- 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.* 37:132–139.
- 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. *Creat Educ J.* 3(Special Issue):937–942.
- 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.* 38:263–271.
- McKinley DW, Boulet JR, Hambleton RK. 2005. A work-centered approach for setting passing scores on performance-based assessments. *Eval Health Prof.* 28:349–369.
- McKinley DW, Norcini JJ. 2013. How to set standards on performance-based examinations: AMEE Guide No. 85. *Med Teach.* 36:97–110.
- Nedelsky L. 1954. Absolute grading standards for objective tests. *Educ Psych Measure.* 14:3–19.
- Norcini JJ. 2003. Setting standards on educational tests. *Med Educ.* 37:464–469.
- 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.* 83:900–905.
- Taylor CA. 2011. Development of a modified Cohen method of standard setting. *Med Teach.* 33:e678–e682.
- Wass V, van der Vleuten CPM, Shatzer J, Jones R. 2001. Assessment of clinical competence. *Lancet.* 357:945–949.
- 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.* 36:62–67.
- 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 Theory Pract.* 11:115–122.