Journal of Contemporary Medical Education

available at www.scopemed.org, www.jcmededu.org

Original Research

Comparing the borderline group and borderline regression approaches to setting Objective Structured Clinical Examination cut scores

Katharine Jane Reid, Agnes Dodds

Medical Education Unit, Melbourne Medical School, at the University of Melbourne, AUSTRALIA

Received: February 04, 2013

Accepted: September 21, 2013

Published Online: November 30, 2013

DOI: 10.5455/jcme.20130921013523

Corresponding Author:
Katharine Jane Reid,
Medical Education Unit, Melbourne
Medical School, at the University of
Melbourne, Victoria, 3010, Australia
kjreid@unimelb.edu.au

Keywords: *Objective Structured Clinical Examinations, Standard setting, Assessment, Borderline Regression, Borderline Groups*

ABSTRACT

The aim of the study was to compare the Borderline Group and Borderline Regression approaches to setting standards on nine Objective Structured Clinical Examinations (OSCEs) undertaken in the final year of medical school. For each of nine OSCEs completed by 301 students we obtained a total score (maximum 40 marks) and a global score on a 5-point rating scale which examiners completed after scoring the OSCE using a checklist. We calculated cut scores for each OSCE using the Borderline Group method by computing the mean and the median score of the group of students whose performance was rated by examiners rated as borderline. We calculated cut scores for each OSCE using the Borderline Regression method by predicting total OSCE scores from global ratings using linear regression, and calculating the cut score by substituting the score of borderline candidates (2) into the regression equation for each OSCE. Both methods established higher standards than an arbitrary 50 per cent criterion. There was also a high degree of concordance between methods. Establishing conceptual standards on OSCEs requires ongoing consideration of appropriate methods and research modelling the effects on cut scores of applying different techniques.

INTRODUCTION

Objective Structured Clinical Examinations (OSCEs) are central to assessment in medical curricula and often establish competence in high stakes contexts [1]. Awareness has developed of the need to set defensible performance standards for OSCEs. Standards are examination scores selected as the boundary between performance that is acceptable or unacceptable. The standard divides candidates deemed competent or not competent on the abilities measured by the assessment [2]. Cut scores (commonly 50 per cent) unrelated to a particular standard of performance are frequently used, with the standard applied to individual OSCEs, a total OSCE score, or to some combination. These standards, though easy to understand and straightforward to apply, are increasingly difficult to justify because they do not consider examination difficulty [3]. With this recognition, medical schools need to: (a) choose defensible standard setting methods, (b) implement chosen methods in the context of available resources and the constraints of assessment schedules, and (c) educate students about the standard setting methodologies used to evaluate their performance.

Different methods of setting OSCE standards exist; each with strengths and weaknesses, both of a pragmatic and a conceptual nature [4,5]. In practice, developing cut scores involves selecting the type of standard and the method for deriving it, determining cut scores, and deciding how they will be used [2,4]. All standard setting procedures incorporate professional judgement. It is erroneous to believe 'true' cut scores exist that can be determined with specific methods or a large enough sample of judges [6, 7]. Given these considerations, it is necessary to select well-described and researched methods of establishing cut scores, which are defensible under scrutiny and which can be justified to students.

Main features of the Borderline Regression and Borderline Group methods

Advocates of the *Borderline Group method* assert that valid OSCE cut scores can be derived from the scores of 'borderline' candidates [8]. For this method, examiners score the OSCE and then rate student performance on a global rating scale. Individual OSCE cut scores are the average scores of candidates with borderline ratings [1,8-10]. Individual OSCE cut scores can be summed to create a total cut score [1,8].

The *Borderline Group method* is time efficient as examiners complete global ratings at the end of the OSCE [11,12]. It thus fulfils the requirement that standard setting methods must be attainable within available departmental resources [5]. Cut scores are easily computed without the need for advanced statistical procedures [12]. Unlike other methods (e.g., Angoff), the *Borderline Group method* is based on examinee performance, which maintains a sense of face validity [12].

Although the *Borderline Group method* is straightforward to implement, it has some limitations. Boursicot and Roberts [13] cautioned that achieving consensus on the competencies of 'borderline' students is difficult. They suggested that clinicians may be unaccustomed to thinking about 'borderline' examinees. Boulet, Champlain & McKinley [14] noted that examiners might assign borderline ratings based on factors other than those assessed by the exam. For these reasons, examiner training is essential to clarify how global ratings are assigned.

Only scores of borderline candidates are used in the *Borderline Group method*, rather than the scores of all students completing the OSCE. When the number of borderline candidates is low, the cut score may have questionable validity. Using the average score of borderline candidates tends to bias cut scores upward; producing an average that is higher than the middle of the category [12]. Ben-David [2] suggested using the median score of borderline candidates, rather than the mean. Moreover, she noted that borderline scores should cluster together for a reasonable standard; widely spread scores may indicate that this method is not applicable.

Recognition of the drawbacks associated with the *Borderline Group method* led to the development of the *Borderline Regression method*, which uses all OSCE scores to develop a cut score using linear regression. Regressing global rating scores onto OSCE total scores produces a linear equation; a predicted score of the borderline candidate (or cut score) is determined by substituting the value of the borderline rating into the regression equation. Though similar to the *Borderline Group method* and sharing its advantages, the *Borderline Regression method* has an additional advantage in that it uses data from all candidates to set the cut score. It is therefore less affected when there are few borderline candidates, or if no candidates are rated borderline. Though it is more computationally complex, the *Borderline Regression method* could still be undertaken within the time constraints of examination periods. Wood et al. [12] undertook a comparison of the *Borderline Group* and *Borderline Regression method*s. They identified that the *Borderline Regression method* generally produced slightly lower but more accurate estimates of cut scores (as evidenced by smaller confidence intervals).

Aim

The purpose of this paper is to model the implementation of the *Borderline Group* and the *Borderline Regression method* for setting the standard on OSCEs undertaken in the final year of medical school.

METHODS

We chose final OSCEs for one cohort of students (*N* = 301) completing a medical course to model the effects of implementing a *Borderline Grou*p or a *Borderline Regression* approach to setting OSCE standards. These results were chosen because of the high stakes of final examinations in establishing competence to proceed to internship; ten OSCEs in final semester comprised 85% of the final assessment. A radiology OSCE station completed without an examiner was excluded from these analyses. These analyses were conducted after this student cohort had graduated, and there was no relationship between the OSCEs students undertook, student outcomes and the current study.

Students completed each nine minute OSCE station in the examination period at the end of their final year. Four OSCEs were undertaken at the university, with the remainder undertaken at the students' assigned clinical school. OSCEs are developed at the university and are specifically designed to measure curriculum outcomes for this medical course. Students undertake OSCEs throughout their training and they are familiar with their structure and the assessment process. These OSCEs focused broadly on interview skills (OSCEs 1–6) and physical examination skills (OSCEs 7–9).

One hundred and thirty-seven examiners assessed on average 17 students (range 5–40) during the OSCE assessment period, with most examiners assessing only one OSCE station. Examiners used a structured checklist for each OSCE which yielded a maximum possible score of 40 marks. Examiners completed a global rating (comprising a five-point Likert scale where 1 = *Fail*, 2 = *Borderline*, 3 = *Pass*, 4 = *Good Pass*, 5 = *Excellent*) after scoring the OSCE using the checklist. Specific examiner training was conducted prior to the scheduled OSCEs in assessing OSCEs and in identifying borderline candidates through completing global ratings. During the examiner briefing, examiners were told that the global rating was for providing an overall impression of the candidates, that the global rating may be used for establishing passing grades, but that the global rating was not the student's grade. No OSCE standard setting method had been implemented in the curriculum for this student cohort and neither students nor examiners would have been familiar with the specific methods compared in this paper.

For the *Borderline Group method*, we calculated predicted cut scores for each OSCE by computing the mean and median score of students who received a borderline global rating. For the *Borderline Regression method*, we developed a linear regression equation for each OSCE which predicted total OSCE scores (the dependent measure) from global ratings (the independent measure). The predicted cut score was obtained by substituting the value for the borderline candidate (in this case 2) into the regression equation. In line with Pell and Roberts [3], we raised this mark by one standard error to limit the number of incompetent candidates passed. Cut scores were rounded to the nearest whole number.

RESULTS

The percentage of candidates rated by examiners as borderline ranged from 5.6% (*n* = 17) to 15.0% (*n* = 45), resulting in reasonable numbers of borderline candidates for each OSCE. The percentage of candidates below the 50% standard was relatively low (median 4.3%) with the exception of OSCE 1 (16.9% below standard). Table 1 shows for each OSCE: average station marks, percentages of borderline candidates, predicted cut scores for each standard setting approach, and numbers and percentages of candidates rated below standard. OSCEs varied in difficulty, with average station scores ranging from 25.0 to 34.2 marks out of 40.

Cut scores developed using the mean and the median of the borderline group were identical for six of the nine OSCEs and differed by a single mark for the remainder. In these cases, the mean cut score was higher. The *Borderline Regression* cut score equalled the *Borderline Group* mean and median cut scores on four OSCEs, was equal to either of these cut scores on two OSCEs and was at least one mark higher than both the *Borderline Group* mean and median cut score on three OSCEs.

Using the mean of the *Borderline Group* as the standard, the majority of students met the required standard on all nine OSCEs (57.8%), 24.3% on eight OSCEs and 12.0% met the required standard on seven OSCEs. The profile using the median of the borderline group as the standard was almost identical: 61.1% met the required standard on every OSCE, 24.6% on eight and 9.6% on seven OSCEs. Using the *Borderline Regression* method 57.5% of students met the required standard on all OSCEs, 23.3% on eight OSCEs and 12.3% on seven OSCEs.

Table 1. Predicted Standard for OSCEs using the Borderline Group and Borderline Regression methods for 301 Final Year Medical Students

			50 per cent	Borderline Group (mean)			Borderline Group (median)			Borderline Regression		
OSCE Station[a]	**Mean OSCE score**	**% rated border-line**	**% below standard**	**Predicted cut score**	**No. below standard**	**% below standard**	**Predicted cut score**	**No. below standard**	**% below standard**	**Predicted cut score**	**No. below standard**	**% below standard**
OSCE 1	28.2	9.3	2.7	22	24	8.0	22	24	8.0	22	24	8.0
OSCE 2	25.0	15.0	16.9	20	51	16.9	19	34	11.3	20	51	16.9
OSCE 3	26.7	6.3	4.3	21	21	7.0	21	21	7.0	21	30	10.0
OSCE 4	25.7	8.6	8.6	19	20	6.6	18	12	4.0	20	26	8.6
OSCE 5	29.0	7.7	2.7	23	24	8.0	23	24	8.0	23	32	10.6
OSCE 6	29.3	11.3	2.7	24	32	10.6	24	32	10.6	24	32	10.6
OSCE 7	32.8	5.6	0.0	27	18	6.0	26	16	5.3	27	18	6.0
OSCE 8	34.2	6.0	0.7	26	14	4.7	26	14	4.7	27	17	5.6
OSCE 9	32.5	7.0	0.3	24	14	4.7	24	14	4.7	25	17	5.6
Total OSCE	284.5	–	0.0	205	3	1.0	203	2	0.7	207	3	1.0

[a] OSCE stations 1–6 were interview based and OSCE stations 7–9 were physical examinations.

The same three students were identified as not reaching the standard on the total OSCE score using either the mean of the *Borderline group* or the *Borderline Regression method*. A stricter conjunctive criterion requires students to pass a minimum number of stations in addition to passing overall. However, only two students did not pass a minimum of five stations and these students also did not meet the total OSCE standard.

DISCUSSION

The importance of using standard setting procedures for OSCEs is frequently asserted but there are few accessible, practical guides which compare the outcomes of using different methods using student assessment data. Comparisons of this nature are important for medical schools in assessing the practical implications of applying the methods to real student data. Often, there is little available time between conducting assessments and deadlines for finalising results. Overly complex or time consuming methods may not be feasible to implement, and appropriately skilled personnel would need to be available on every assessment occasion. Methods that set unreasonably high cut scores would be undesirable because of the implications of organising supplementary assessments for large numbers of students. The research reported in this study, provides a real world example of the implementation of two standard setting methods for OSCEs for our medical school, and other institutions, to consider the implications of implementing one of these methods in their assessment schedule.

The current study demonstrated high concordance on OSCE standards developed using two empirical standard setting methods. The number of students identified as not meeting established standards using either a compensatory (total OSCE score) or conjunctive approach (passing a minimum number of stations in addition to overall) was extremely small. Each method established more rigorous standards and identified a small number of students who fell below predicted cut scores, yet met an arbitrary 50 per cent criterion.

The concordance between methods is encouraging, providing some level of reassurance that the less resource intensive *Borderline Group method* may be implemented with some confidence; however, the likelihood that there may be few (or no) candidates rated as borderline, favours the *Borderline Regression method*. Both methods, do, however, rely heavily on examiner judgements of borderline candidates. This is a conceptual evaluation, based on assessing whether the candidate exhibits the requisite knowledge and skills of a minimally competent candidate. Examiners complete global assessments after completing the examination checklist; thus, it would expected that examiners global assessment would be influenced by their checklist ratings. However, it is clear that examiner judgements reflect their expectation that a minimally competent candidate should be expected to have mastery of a greater proportion of the knowledge and skills required to complete an easier OSCE (e.g., the physical examinations assessed on OSCEs 7–9) with lower attainment expected on a more difficult OSCE (e.g., the interviews assessed on OSCEs 1–6). At the time of this research, examiners received minimal guidance in assigning global ratings for OSCEs. The validity of the standard setting methods described in this report is reliant on a common understanding of the characteristics of borderline candidates. Ongoing attention should be paid to providing examiners with a clear definition of the borderline or minimally competent to encourage a shared understanding of this concept.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

1. Dauphinee W, Blackmore D, Smee S, Rothman A, Reznick R. Using the judgments of physician examiners in setting the standards for a national multi-center high stakes OSCE. Advances in Health Sciences Education 1997;2(3):201-11.
2. Ben-David M. AMEE Guide No. 18: Standard setting in student assessment. Medical Teacher 2000;22(2):120-30.
3. Pell G, Roberts T. Setting standards for student assessment. International Journal of Research & Method in Education 2006, 29(1):91-103.
4. Norcini J. Setting standards on educational tests. Medical Education 2003;37(5):464-69.
5. Talente G, Haist S, Wilson J. A model for setting performance standards for standardized patient examinations. Evaluation & the Health Professions 2003;26(4):427.
6. Barman A. Standard setting in student assessment: Is a defensible method yet to come? AnnAcad Med Singapore. 2008;37(11):957-63.
7. Dwyer C. Cut scores and testing: Statistics, judgment, truth, and error. Psychological Assessment 1996;8(4):360-62.
8. Humphrey-Murto S, MacFadyen J. Standard setting: a comparison of case-author and modified borderline-group methods in a small-scale OSCE. Academic Medicine 2002;77(7):729.

9. Smee S, Blackmore D. Setting standards for an objective structured clinical examination: the borderline group method gains ground on Angoff. Medical education 2001;35(11):1009-10.

10. Wilkinson T, Newble D, Frampton C. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. Medical Education 2001;35(11):1043-49.

11. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Medical Education 2003;37(2):132-39.

12. Wood T, Humphrey-Murto S, Norman G. Standard setting in a small scale OSCE: a comparison of the modified borderline-group method and the borderline regression method. Advances in Health Sciences Education 2006;11(2):115-22.

13. Boursicot K. Setting standards in a professional higher education course: Defining the concept of the minimally competent student in performance-based assessment at the level of graduation from medical school. Higher Education Quarterly 2006;60(1):74-90.

14. Boulet J, De Champlain A, McKinley D. Setting defensible performance standards on OSCEs and standardized patient examinations. Medical Teacher 2003;25(3):245-49.