Adv in Health Sci Educ (2013) 18:231–244
DOI 10.1007/s10459-012-9367-y

The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score in medical programme assessments

Boaz Shulruf · Rolf Turner · Phillippa Poole · Tim Wilkinson

Received: 20 December 2011 / Accepted: 9 March 2012 / Published online: 7 April 2012

Abstract The decision to pass or fail a medical student is a 'high stakes' one. The aim of this study is to introduce and demonstrate the feasibility and practicality of a new objective standard-setting method for determining the pass/fail cut-off score from borderline grades. Three methods for setting up pass/fail cut-off scores were compared: the Regression Method, the Borderline Group Method, and the new Objective Borderline Method (OBM). Using Year 5 students' OSCE results from one medical school we established the pass/fail cut-off scores by the abovementioned three methods. The comparison indicated that the pass/fail cut-off scores generated by the OBM were similar to those generated by the more established methods ($0.840 \leq r \leq 0.998$; $p < .0001$). Based on theoretical and empirical analysis, we suggest that the OBM has advantages over existing methods in that it combines objectivity, realism, robust empirical basis and, no less importantly, is simple to use.

Keywords Standard-setting · Borderline grades · Assessment · Cut-off score

B. Shulruf (✉)
Centre for Medical and Health Sciences Education, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand
e-mail: b.shulruf@gmail.com; b.shulruf@auckland.ac.nz

R. Turner
Department of Statistics, Faculty of Sciences, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand
e-mail: r.turner@auckland.ac.nz

P. Poole
Department of Medicine, Faculty of Medical and Health Sciences, University of Auckland, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand
e-mail: p.poole@auckland.ac.nz

T. Wilkinson
Health Care of the Elderly, University of Otago, The Princess Margaret Hospital, P.O. Box 800, Christchurch, New Zealand
e-mail: tim.wilkinson@otago.ac.nz

Introduction

Students who are eligible to gain entry to a medical programme are often considered to be the 'cream of the crop' in terms of being high-achieving and academically-focussed individuals. However, the reality is that within any medical programme, there is always a small group of students who receive borderline pass or fail grades in a range of assessments. This group causes a persistent challenge for teaching staff and administrators in determining whether or not these students should pass, and thence progress through the programme and ultimately into the medical workforce, or should receive a fail grade as they have not yet met the standard. Making a wrong decision in passing a borderline student could have negative consequences in perpetuating weaknesses in applied knowledge and performance (for example see: Hays et al. 2008). In medicine, this may literally be life-threatening. On the other hand, failing a student with the attributes to be a good doctor has adverse consequences for the student and is a loss for the society.

The literature on standard setting focuses mainly on making decisions about borderline students whose overall scores on a number of examinations, tests, or test components such as stations in Objective Structured Clinical Examination (OSCE) (Harden and Gleeson 1979) indicate that they have reached a level where it is not clear whether they should pass or fail the examination (Halpin and Halpin 1987). The AMEE Guide No. 18: Standard setting in student assessment, (Ben-David 2000) well describes the difficulty in choosing the proper method for standard setting or pass/fail decisions for borderline students. In this monograph, Ben-David (2000) classifies assessment into two types: (a) relative/norm-based assessment and (b) criterion/absolute-based assessment. Each type has some advantages. A criterion/absolute-based assessment compares student performance to a predefined, recognised standard which increases the confidence that students who met the standards are competent enough (Nedelsky 1954). Nedelsky's (1954) definition of a criterion/absolute-based standard has been a foundation pillar in this field:

> the passing score is to be based on the instructor's judgement of what constitutes an adequate achievement on the part of the student and not on the performance by the student relative to his/her class or to any other particular group of students. In that sense the standard to be used for determining the passing score is absolute.

On the other hand, in some circumstances there is a need to set up pass/fail standards which are relative to the examinee population (e.g. when places are limited, when the criterion-based standard has not yet been established, or when it is difficult to establish such a standard). In such cases, a relative/norm-based assessment is used. Nonetheless, assessments and standard setting methods of both types are vulnerable to critique (Ben-David 2000; Cizek and Bunch 2006; Norcini 2003).

The main shortcomings of a relative/norm-based method such as Contrasting Groups (Norcini 2003) or an extrapolation type, such as the Borderline Regression Method (Kramer et al. 2003; Woehr et al. 1991; Wood et al. 2006) are as follows:

(a) standards are not based on an agreed pre-defined level of competency;
(b) standards are not known in advance;
(c) the examination difficulty is not considered in advance (Ben-David 2000; Cizek and Bunch 2006).

Furthermore, critique of norm-based methods for cut-off scores may be method-specific. For example, when using the Regression Method (e.g. Kramer et al. 2003) students are given raw scores and grades (e.g. Fail, Borderline, Pass, Distinction) for their performance and the scores are linearly regressed to the grades to identify the score that falls on the borderline grade. This score is set up as the cut-off score. A criticism is that this method assumes a linear association between the scores and the grades, and that the grades are on an interval scale, which is usually not the case. In addition, the cut-off score is affected by extreme scores, which serves to undermine the reliability of the regression method (Cizek and Bunch 2006).

Another norm-based method is the Borderline Group Method (BGM) (Cizek and Bunch 2006; Livingston and Zieky 1982; Zieky and Livingston 1977) which uses the same scoring and grading method as the Regression Method. However, the BGM sets up the cut-off score as the median of the scores classified as Borderline. The main shortcoming of the BGM is that it does not consider the performance of most of the examinees, particularly those who met the most certain outcomes of Pass and Fail (Cizek and Bunch 2006).

The absolute/criterion-based methods for standard setting aim to determine examinees' performance based on a predefined competency level which is not related to the group performance. The most widely used standard setting method in medicine and health sciences is the the Angoff method, which has spawned several derivative methods (Hurtz and Auerbach 2003; Jalili et al. 2011). Nonetheless, absolute/criterion-based methods such as Angoff (Behuniak et al. 1982), Nelensky (1954), Ebel (1972) have significant shortcomings, namely:

(a) the standards rely on expert judgement which requires employing a large number of judges to achieve an acceptable reliability;
(b) the experts are normally not the examiners, hence the actual examiners and the expert panellists may have different criteria for what constitutes a Pass level; and
(c) the process may be time-.and resource—consuming as well as being inconvenient for clinical academics.

Hofstee's method for standard setting brings together expert judgement and norm methods, aiming to minimise the shortcomings of both norm and criterion based methods (Wayne et al. 2005). Although it appears to have advantages as the cut-off point is set up by both student performance (relative method) and expert judgement (criterion based method), it is a compromise, as the shortcomings of both types of assessment are manifested in this method. In particular, Hofstee's method relies upon experts to estimate the following:

(a) the highest percent correct cut score that would be acceptable, even if every examinee attains that score;
(b) the lowest percent correct cut score that would be acceptable, even if no examinee attains that score;
(c) the maximum acceptable failure rate; and
(d) the minimum acceptable failure rate.

Hence these criteria, set up by experts to determine the cut-off score, are also vulnerable to judgements about a hypothetical group of students. As a result this is more of a ompromise than an objective method for setting up pass/fail cut-off scores (Bowers and Shindoll 1989; Cizek and Bunch 2006).

The Contrasting Groups method (Cizek and Bunch 2006; Norcini 2003) is a method that aims to identify the pass/fail cut-off score that best separates the two groups (either Pass or Fail). It may use graphical methods (Norcini 2003) or logistic regression to determine the score at which the probability of categorisation is, say, 50 %, if that is decided to be the proportion of passing students. Again, the classification of passes and fails is made by judges that might allowsthe same score to be classified as both a Pass and a Fail (the overlapping range); which conceptually undermines the confidence in such classification (Cizek and Bunch 2006).

This study introduces a new pass/fail decision-making method for borderline grades namely the *Objective Borderline Model* (OBM). This method addresses some of the shortcomings of the currently-available methods, thus providing more objectivity in decision-making for borderline grades. Specifically, the OBM is a method that determines the pass/fail cut-off score by specifying a readily-calculable probability for a Borderline grade to be reclassified as Pass, thence choosing the cut-off so that the specified probability is achieved. This estimated probability uses only students' examination scores and three grade classifications: Fail, Borderline, and Pass.

The first underlying assumption of the OBM is that the test examiners may clearly decide three ranges of examination scores (Nedelsky 1954). A clear Pass grade (P) includes the scores at or above the point that *all* examiners agree upon; a student who reached that score has definitely passed the exam. A clear Fail grade (F) includes the scores at or below the point that *all* examiners agree is a failing score. A Borderline grade (B) includes all the scores which fall between the P and the F range; that is, those where there is uncertainty whether or not such scores should be determined as P or F.

The P, B, or F classification method is widely accepted (Nedelsky 1954; Prideaux et al. 2011; Wilkinson et al. 2001) and has strong face validity, since it uses raw student examination scores and the most conservative grade classification—all scores within either the F or the P grade ranges are those which were classified as such *without any doubts* by the examiners. Obviously, these examiners are deemed to have the best knowledge for determining the expected level of a Pass and Fail in that particular exam (Nedelsky 1954).

The second assumption is that within the borderline score range, the higher the score the more likely it is to be a Pass.

The aim of this study is, therefore, to demonstrate the feasibility and practicality of the OBM and collect evidence testing its validity (Schuwirth et al. 2011).This is done by comparing the pass/fail cut-off scores generated by the OBM to those derived from the BGM and the Regression method.

We decided to test the validity of the OBM using OSCE scores for a number of reasons. First, one set of full OSCE examination provides us with an opportunity to undertake 16 independent tests of validity (scores of each station were compared independently) which added robustness to the study. Second, we acknowledge that setting pass/fail standards for OSCEs has been an ongoing challenge for many scholars and practitioners, so demonstrating that the OBM works well with OSCE data might improve practice in an area of great need (Schuwirth et al. 2011).

We measured the validity of the OBM against two standard setting methods in common use. The Regression Method has been used at the University of Otago School of Medicine to make pass/fail decisions over similar OSCE examination data and it is well established as a valid method for standard setting (Schoonheim-Klein et al. 2009). To strengthen our findings we followed the relevant consensus statements and recommendations (particularly recommendation 2) from the Ottawa 2010 Conference (Schuwirth et al. 2011) by conducting additional validity check of the OBM against the Borderline Group Method (Cizek and Bunch 2006; Livingston and Zieky 1982; Zieky and Livingston 1977), which uses only the borderline grades for pass/fail decision making. To further strengthen our confidence in the OBM, we checked the OBM's validity against the Borderline Group Method and Regression method for each of the 16 OSCE station separately, which was summed up to 16 independent checks.

The Objective Borderline model (OBM)

When a collection scores is achieved by n students on a single examination, their resultant grade could be classified as 'Fail' (F), 'Borderline' (B), or 'Pass' (P). Note that F is a fail without any doubt, P is a pass without any doubt, and B is where there is indecision as to whether the final grade should be Pass or Fail, based on the examination score. It is assumed that there are thresholds such that a student's grade may be determined by noting which thresholds his or her score lies below, above, or between. Let the thresholds determining the B's be "l" (lower) and "u" (upper). In these terms a score "S" is assigned a grade of B if and only if $l < S < u$.

Our aim is to make a decision about the students with 'Borderline' grades and reassign them as being either 'Fail' or Pass, and to do so on a rational basis, using a probability-based model. This is done in terms of the numeric scores by determining a new threshold $\textsc{t}$, between 'l' and 'u,' and then assigning scores $S \geq \textsc{t}$ a P (Pass) and all other scores an F (Fail).

We propose to determine $\textsc{t}$ in terms of probabilities calculated on the basis of the observed fractions or proportions of F's, B's, and P's from the initial grade assignments. In determining $\textsc{t}$ it is crucial to take into consideration the ordinal nature of the grades. Explicitly, F is not as good as (or "is less than") B, and similarly B is not as good as ("is less than") P.

To take the ordinal nature of the grades into consideration, it is expedient to consider the grades in pairs, namely (F, B) and (B, P). We think in terms of applying to each B grade two independent tests, one in terms of the (F, B) pair and one in terms of the (B, P) pair, and converting the B to a P only if it is successful in both tests. We set the probability that it succeeds in the first test in terms of the numbers of F's and B's in the observed sample. Let these numbers be n_F and n_B respectively. We set the probability (p_{r1}) that a grade passes the first test to be

$$p_{r1} = n_B/(n_F + n_B)$$

which is the probability that a grade randomly selected from amongst the observed F's and B's will be a B. Likewise, we set the probability that a B grade is successful in the second test to be

$$p_{r2} = n_P/(n_B + n_P)$$

where n_P is the number of P's in the observed sample; note that $n_F + n_B + n_P = n$. The value of p_{r2} is the probability that a grade randomly selected from amongst the observed B's and P's will be a P.

Since the two tests are, by assumption, independent, the probability that both are passed is simply the product $p_r = p_{r1} \times p_{r2}$. We now determine $\textsc{t}$ as that point between 'l' and 'u' so that the probability that a score S corresponding to a B grade will be greater than $\textsc{t}$ is p.

That is, we take $\textsc{t}$ to be the $(1 - p_r)$th quantile of the empirical distribution of the values of S corresponding to B grades. In other words, we select $\textsc{t}$ so that the proportion of values of S, corresponding to B grades, which are greater than or equal to $\textsc{t}$ is p.

To simplify the calculations we could choose $\textsc{t}$ so that the fraction of the interval [l, u) lying above $\textsc{t}$ is equal to p_r. This amounts to assuming that scores S corresponding to B grades are uniformly distributed on [l, u). This assumption may well be a reasonable approximation to the truth. Such a score S chosen at random between l and u, will have a probability of falling above $\textsc{t}$ equal to p. Hence we would set

$$\mathrm{T} = p_{\mathrm{r}} \times \mathrm{l} + (1 - p_{\mathrm{r}}) \times \mathrm{u}$$

Note that if we choose one of the B grades at random, its corresponding score S will have probability p_r of being larger than T. Hence the probability that this randomly chosen grade B will be converted to a P is p_r, the probability that the B met both of our conceptual tests, as desired.

In practical terms, to establish the cut-off score given the grade distribution $n_F + n_B + n_P = n$, we establish the probability $p_r = p_{r1} \times p_{r2} = (n_B/(n_F + n_B)) \times (n_P/(n_B + n_P))$ The cut-off score is then the (1 – p_r)th quantile of the borderline score (See "Appendix 1" for practical instructions and an example).

Methods: model testing

To test the utility of the OBM we used results from the final summative OSCE for Year 5 medical students from a medical school in New Zealand, which comprised 16 stations, each treated as an independent test. The data include scores (possible range 0–20) from 16 OSCE stations and examiners' estimates at the time of the overall grade of each student in each station. The grades were: Below the expected level; Borderline; At the expected level; and Above the expected level. We calculated the pass/fail cut-off score for the students in each OSCE station using three different methods. The first method was a modified Borderline Group Method (henceforth: mBGM) (Cizek and Bunch 2006; Zieky and Livingston 1977) which was also used by Wilkinson et al. (2001) on similar data. This method sets up the pass/fail cut-off score as the mean of the scores, which were classified as Borderline. The mBGM differs from the Borderline Group Method only in using the mean rather the median (which were very similar) of the borderline grades to set up the pass/fail cut-off score. Note that a student's score was the mean of the two scores given by the two examiners in each OSCE station; and a score was defined as a borderline when at least one examiner identified the student performance as Borderline (Wilkinson et al. 2001).

The second method was the Regression method (e.g. Kramer et al. 2003). In the Regression method, student scores are regressed to the respective grades (Below the expected level = 1; Borderline = 2; At the expected level = 3; and Above the expected level = 4). In the regression model, we regressed the mean scores on the mean grade. Since there were two examiners, the cut-off score was defined as the score where the linear regression line crossed the grade which both examiners identified as Borderline (i.e. 2). Choosing the regression model and the mBGM to test the validity of the OBM deemed to be most appropriate as these methods incorporate characteristics of both norm- and criterion- based methods (Cizek and Bunch 2006). Moreover, these two methods do not use judges to set up the cut-off score which inevitably if used, would question the validity of the 'gold standard' set for this comparison (see section on validity: Schuwirth et al. 2011, p. 228).

The final method was the OBM that has been described above. However, since the OSCE data did not include a score range for Borderline grades, we had to establish a set of ranges by varying the distance from the cut-off score as defined by the mBGM. Table 1 shows the four ranges used to test the impact of the range on the cut-off score generated by the OBM, as well as how the range is determined for the mBGM and the Regression method. Note that the Borderline Group Method's criterion identified a relatively large proportion of borderline grades in each OSCE station ranging from 11.2 to 52 % (see Table 2).

Table 1 Definition of a borderline score

Range name	Definition of a borderline score
1SD	All scores fall within 1 standard deviation from mBGM's cut-off score
0.5SD	All scores fall within 0.5 standard deviation from mBGM's cut-off score
2SE	All scores fall within 2 standard error of the mean from mBGM's cut-off score
1SE	All scores fall within 1 standard error of the mean from mBGM's cut-off score
mBGM	All scores identified by at least one examiner as a borderline grade
Regression	All scores identified by at least one examiner as a borderline grade

Table 2 Comparison of proportion of students who were Borderline and the pass/fail cut-off score per OSCE station, by method

Station	Objective Borderline model (OBM)								%	mBGM	%	Regression
	%	1SD	%	0.5SD	%	2SE	%	1SE				
1	33.5	7.25	11.7	7.50	6.1	7.25	5.1	7.25	22.9	7.09	6.1	6.87
2	35.0	6.25	19.3	6.00	9.1	6.00			52.0	5.87	22.3	5.71
3	44.2	8.25	20.8	7.75	15.2	7.75	7.1	7.59	15.1	7.60	5.0	7.65
4	26.9	6.06	8.1	6.10	8.1	6.10	3.0	6.00	24.6	5.99	6.1	5.68
5	34.0	6.25	21.3	6.25	10.7	6.25	5.6	6.25	35.8	6.20	11.2	5.90
6	27.4	8.00	16.2	7.89	11.2	7.75	6.1	7.75	16.2	7.91	7.3	7.69
7	25.4	6.75	13.2	7.00	5.6	7.25	3.6	7.25	24.6	7.13	6.1	6.92
8	41.6	7.25	25.4	7.00	13.2	7.00	10.7	7.00	26.3	6.89	8.9	6.61
9	44.2	8.00	19.8	7.01	12.7	7.50	8.1	7.25	40.2	7.17	15.1	6.89
10	39.6	8.50	20.3	8.00	13.2	7.82	6.6	8.00	26.3	7.88	7.3	7.86
11	28.4	7.50	11.2	7.20	4.1	7.00	0.5	7.25	20.1	7.17	5.6	6.81
12	31.5	7.45	17.3	7.00	7.1	7.00	6.1	7.00	19.6	6.99	3.9	6.86
13	19.8	11.50	8.1	12.00	6.6	12.00	2.0	12.00	23.5	12.15	12.3	12.22
14	39.6	9.00	12.2	8.50	12.2	8.50	5.6	8.50	15.1	8.38	7.8	8.20
15	19.8	7.00	9.6	7.00	9.6	7.00	5.1	7.00	11.2	7.06	3.9	6.99
16	40.1	7.75	19.8	7.50	17.8	7.50	8.1	7.50	21.2	7.44	6.1	7.35

%, The percentage of scores identified as borderline among all scores

To compare the models, we calculated the mean cut-off scores of all stations by each model as well as the correlations between the models.

Results

The proportion (%) of borderline scores identified by each method and the cut-off scores identified for each station classified by method and borderline range are presented in Table 2. Each of the three models (OBM [with borderline ranges of 1SD, 0.5SD, 2SE and 1SE]; mBGM; and Regression) identified very similar cut-off scores. These are shown in Fig. 1). There were very high correlations among the mean cut-off scores of the models (Table 3).

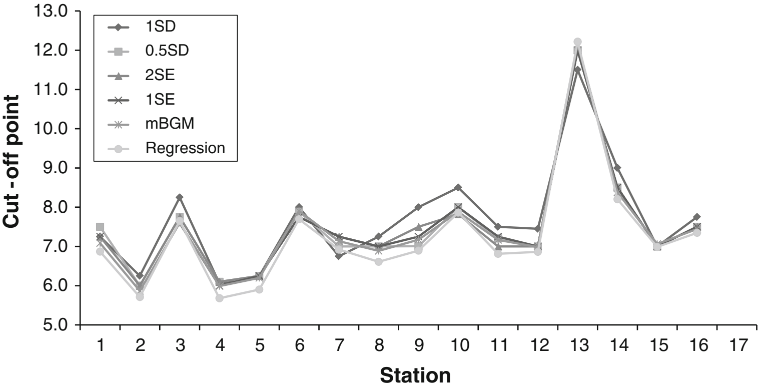


Fig. 1 Mean cut-off scores of 16 OSCE stations by method

Table 3 Correlation among pass/fail cut-off scores across methods

	Model	OBM				mBGM
		SD1	SD0.5	SE2	SE1	
Regression		0.964	0.994	0.994	0.995	0.998
mBGM		0.963	0.995	0.996	0.998	
OBM	SE1*	0.964	0.995	0.996		
	SE2	0.968	0.992			
	SD0.5	0.964				

n = 16, $p < .0001$

* n = 15

These comparisons of 16 independent OSCE stations demonstrate that the OBM provides pass/fail cut-off scores which are very similar to the Regression and the mBGM methods, despite the fact that the OBM has established those cut-off scores based on a different paradigm and statistical method.

Discussion

This study aimed to describe and show aspects of validity of a new standard setting model named the Objective Borderline Method (OBM), by comparing pass/fail cut-off scores defined by the OBM with other two well-established methods, namely the modified BGM method (Cizek and Bunch 2006; Wilkinson et al. 2001; Zieky and Livingston 1977) and the Regression method (Kramer et al. 2003).

The results indicate that the OBM is as effective as the other methods in that it generated very similar cut-off scores to the mBGM and the Regression methods (Tables 2, 3; Fig. 1). We argue, however, that the OBM is preferable for a number of reasons.

The OBM is based on standards set up by the examiners for the first pass or fail decision (that is, score must be a clear pass or a clear fail), which is in line with the fundamental

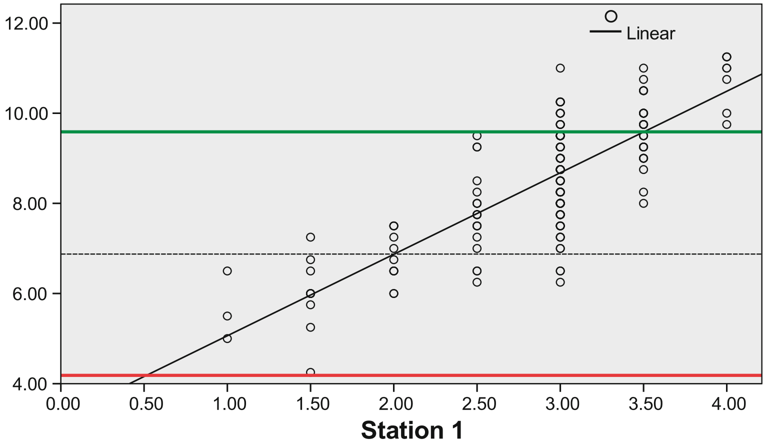


Fig. 2 Example of Regression method using data from OSCE station 1

criterion for pass/fail decision within the education assessment context (Nedelsky 1954). Since this study found that the range of borderline grades has negligible impact on the cut-off score, we suggest that the easiest way to achieve agreement among all examiners is to set up the broadest borderline range suggested by any of the examiners. This is reasonable as the OBM then identifies the clear passes and fails without any doubt. As presented in the introduction, many standard setting methods do not establish pass/fail decisions based on Nedelsky's foundations, but rather use experts' estimate of the expected competence of the examinees (in absolute/criterion based methods) or by identifying a cut-off score, which in most methods is a measure of central tendency of the Borderline scores (Ben-David 2000; Cizek and Bunch 2006; Cusimano 1996; Norcini 2003; Norcini and Shea 1997). Even the Contrasting Groups method is based on the discrepancy of judgement, where the same score might initially be identified as pass for one examinee and fail for another (in the same examination/OSCE station) (Cizek and Bunch 2006). Moreover, these authors contend that "the choice of 50 is one that sets false positive and false negative classification errors as equally serious, which again highlights that the choice of this value is a policy, rather than a technical decision" (Cizek and Bunch 2006, p. 109).

Figure 2 demonstrates just how problematic such methods (e.g. Regression and mBMG) are. Scores that were considered as Borderline ranged between 4.25 and 9.50, but at the same time scores as high as 6.5 were considered as Fail and as low as 6.25 were considered as Pass.

On the other hand, the OBM sets up in advance the Pass, Fail and Borderline ranges, so each score is classified into one of these categories whereby only the Borderline scores (which by definition could be either Pass or Fail) are then reclassified into Pass or Fail based on a probability model.

The OBM estimates the probability of a Borderline score to be considered as a Pass, based on the distribution of the Fail, Borderline and Pass grades. We propose a probability model is more realistic and makes fewer assumptions than other methods, yet is equally robust. The OBM does not claim to set up an absolute, 'set-in-stone', cut-off score for every population. The OBM rather sets up *absolute standards for clear passes and fails* (standard based method) that applies to *all populations*. Then, the OBM sets up the cut-off score based on the *probability* that Borderline scores in a particular examination (e.g. OSCE station) for a particular population meet the Pass standard.

The advantages of the OBM are numerous:

1. the OBM is based on absolute rather than relative standards;
2. the OBM does not rely upon a panel of experts to set up the standards but rather trusts the examiners to set up standards for Pass and Fail. The examiners are responsible for the quality of the standards and the decision whether students pass or fail the examination. The OBM method clearly specifies that the scores which do not fall within the ranges that all agreed would be either clear Pass or Fail are defined as borderline; hence, subject to the OBM;
3. the OBM does not consider hypothetical groups of examinees to set up cut-off scores, but rather uses data from the actual examinees;
4. the OBM considers the examinees' population but without compromising the acceptable level of the pass performance;
5. the OBM is simple, does not require any statistical skills, and can be easily calculated;
6. although the cut-off score cannot be identified before the examination, the examinees may know in advance what score ranges are associated with Pass, Fail and Borderline grades. They may also know how the cut-off score is calculated, which minimises uncertainty and increases a sense of fairness;
7. the OBM sets up the pass/fail cut-off score based on the relevant measure (the estimated probability that a Borderline score will be reclassified as Pass), which removes the impact of extreme scores on the cut-off score;
8. the OBM does not assume a normal distribution of scores and may be applied with any score distribution;
9. unlike some other standard setting methods (e.g. Contrasting Groups) there is no need for a minimum number of examinees to fail in order for the OBM to establish a cut-off score;
10. with the OBM, each score is uniquely classified to one category and it is impossible that the same score would be classified (even by different examiners) as both Pass and Fail, as in Fig. 2, because only Borderline scores are reclassified. This is an underlying principle of standard setting methods. Finally:
11. the OBM tolerates a range of Borderline bands, as shown in Fig. 1, which enhances the confidence in its stability and validity.

Nonetheless, some may argue that a shortcoming of the OBM is that different cut-off scores might be established for different populations sitting the same examination. However this is not a major shortcoming, since the Borderline grades are all deemed to be possibly pass and possibly fail with the cut-off score based on the *probability* of Borderline grades being a Pass for each population of examinees. Given the advantages of the OBM this minor shortcoming should not impede educators using it.

It is acknowledged that establishing the validity of a standard setting method is not an easy task nor could it ever be completely achieved (Cizek and Bunch 2006; Kane 2001; Messick 1995a, b; Schuwirth et al. 2011; Schuwirth and van der Vleuten 2012). However, the OBM seems a valid method for pass/fail standard setting, as it meets several relevant and desirable criteria as agreed in Ottawa conference 2010 (Schuwirth et al. 2011):

- We have established a clear statistical rationale for the OBM, as well as the theoretical justification for a new and better method for pass/fail standard setting within and beyond medical education theory (recommendations 2, 3, 6 and 8).

- We suggest that the methodology used in this study is optimally able to answer the research question rather than be superior to all others (recommendation 9).
- Finally we used a set of defensible arguments (recommendation 12) to support the validity of the OBM. We set up the validity check within an appropriate context i.e. OSCE examinations for medical students and compared different standard-setting methods using the same data (recommendation 13).

We acknowledge that consequential validity (Messick 1995b) was out of the scope of this study and recommend that future research addresses that issue.

Study limitations and further research

Our study has some limitations. The most important is that we used data that had been tested before, hence we could not determine Pass, Fail and Borderline ranges from scratch. To circumvent this issue, we used different ways to establish the Pass, Borderline and Fail ranges. If comparisons of different methods for establishing pass/fail cut-off scores is done prospectively, care needs to be taken to avoid the ethical, if not legal, implications of having slightly different cut-off scores. We found the differences to be minimal to negligible. Another limitation is that we used data from only one cohort of students, which may limit the generalisability of the study. However, since we tested the OBM over 16 independent OSCE stations, and all results appeared to be similar, it is reasonable to assume that the OBM would perform in a similar way in other types of assessments.

Another limitation of the OBM is that if there are no Passes among the examinees population, the OBM will indicate that none of the borderline scores should be reclassified as Pass. Again, this is an unlikely scenario and is an issue for most other standard-setting methods.

Further research, is recommended to confirm the robustness and utility of the OBM in other settings and to identify any unknown limitations of the OBM. We have, though, already demonstrated that the range of the borderline scores has a negligible impact on the cut-off score. Furthermore, before carrying out this empirical study on real data, we extensively tested the OBM on hundreds of simulated datasets, testing different distributions and different borderline ranges. None of the simulations, which determined up to 60 % of the scores as borderline grades provided unreasonable results. Nonetheless, it is important to identify whether or not there is any range, score distribution, and proportion of Borderline scores that severely destabilises the model.

Last but not least, this study tests the validity of the OBM using OSCE data. Based on the theoretical background presented earlier on, we assume that the OBM could be used for any test that results in a score. Testing the OBM in different context or types of assessments (for example long cases, mini-CEXs, short answer questions, or even a set of MCQ questions) is not within the scope of this study and might be scrutinised in further research.

Conclusions

We have described a new method, the OBM, to determine pass/fail cut-off scores, that yielded has cut-off scores very similar to the mBGM and the Regression methods. Yet, the OBM has several advantages over other methods in that it combines objectivity, realism, theoretical robustness and, no less importantly, is simple to use.

Appendix 1: Practical instructions and an example

How to use the OBM for setting up an objective pass/fail cut-off score.

The example below includes scores from OSCE station 1 used in the study. The borderline grades are defined as all scores between 6.75 and 7.25 inclusive, in total 12 scores.

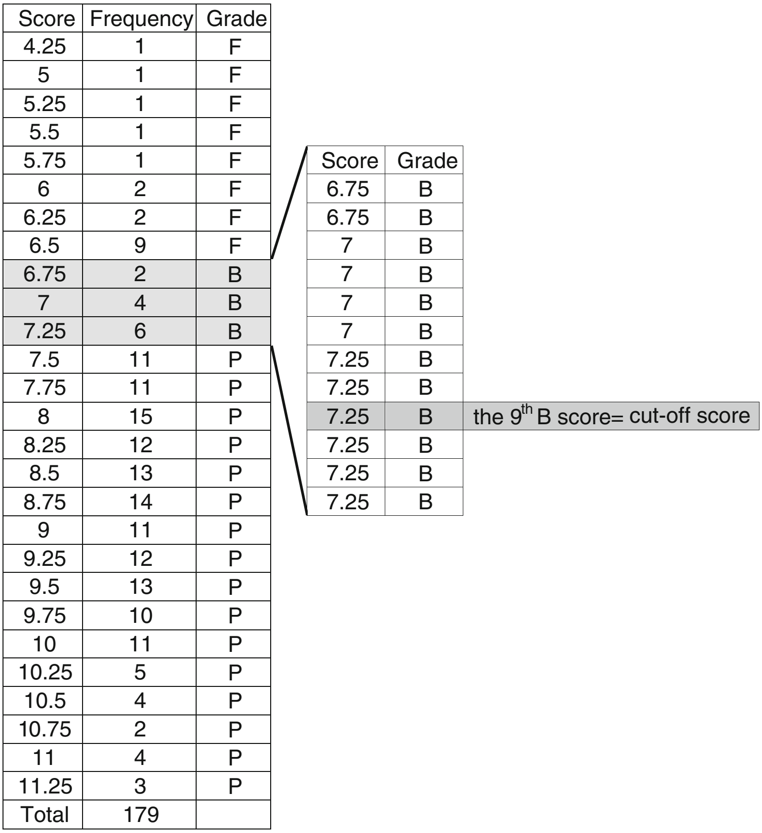
The distribution of scores and classification for ‘Pass’ (P), ‘Borderline’ (B) and ‘Fail’ (F) is presented in Table 4. Overall there are 179 scores distributed as follow, where F, B, and P stand for the number of grades in each category:

Number of fails (F)	F = 22
Number of borderlines (B)	B = 12
Number of passes (P)	P = 145

Use Eq. 1 to identify the score which will be used as the pass/fail cut-off.

Equation 1 the quantile determining the cut-off score

Table 4 Example OSCE scores and grades (station 1)



Score	Frequency	Grade
4.25	1	F
5	1	F
5.25	1	F
5.5	1	F
5.75	1	F
6	2	F
6.25	2	F
6.5	9	F
6.75	2	B
7	4	B
7.25	6	B
7.5	11	P
7.75	11	P
8	15	P
8.25	12	P
8.5	13	P
8.75	14	P
9	11	P
9.25	12	P
9.5	13	P
9.75	10	P
10	11	P
10.25	5	P
10.5	4	P
10.75	2	P
11	4	P
11.25	3	P
Total	179	

Score	Grade	
6.75	B	
6.75	B	
7	B	
7	B	
7	B	
7	B	
7.25	B	
7.25	B	
7.25	B	the 9th B score= cut-off score
7.25	B	
7.25	B	
7.25	B	

$$1 - \left(\frac{\mathrm{B}}{(\mathrm{B} + \mathrm{F})}\right) * \left(\frac{\mathrm{P}}{(\mathrm{P} + \mathrm{B})}\right) \tag{1}$$

Place F = 13, B = 12 and P = 145

$$1 - \left(\frac{12}{(12 + 22)}\right) * \left(\frac{145}{(145 + 12)}\right) = 0.674 = 67.4\,\%$$

The 67.4th percentile among the borderline scores determines the cut-off score.

Since there are 12 borderline (B) scores (see the shaded rows in Table 4), the cut-off score is the score above the 67.4 % of the borderline scores which is the 8.09th (12 × 0.674 = 8.09) or rounded up the 9th borderline score. In Table 4 we see that the 9th borderline score is 7.25 which is the minimum pass score for this OSCE examination for this population of examinees.

Note that six of the borderline grades have scores of 7.25 and it is impossible to rank them, hence everyone who achieved 7.25 or more will receive a pass grade in this station (similar to any other standard-setting method when a number of borderline scores could not be ranked).

References

Behuniak, P., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. *Educational and Psychological Measurement, 42*(1), 247–255. doi:10.1177/0013164482421031.

Ben-David, M. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher, 22*(2), 120–130.

Bowers, J., & Shindoll, R. (1989). *A comparison of the angoff, beuk, and hofstee methods for setting a passing score* (pp. 1–38). Iowa: ACT Organisation.

Cizek, G. J., & Bunch, M. (2006). *Standard setting: A guide to establishing and evaluating performance standards on tests*. London: Sage Pubns.

Cusimano, M. (1996). Standard setting in medical education. *Academic Medicine, 71*(10), S112–S120.

Ebel, R. L. (1972). *Essentials of educational measurement*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Halpin, G., & Halpin, G. (1987). An analysis of the reliability and validity of procedures for setting minimum competency standards. *Educational and Psychological Measurement, 47*, 977–983.

Harden, R. M., & Gleeson, F. A. (1979). Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education, 13*(1), 39–54. doi:10.1111/j.1365-2923.1979.tb00918.x.

Hays, R., Sen Gupta, T., & Veitch, J. (2008). The practical value of the standard error of measurement in borderline pass/fail decisions. *Medical Education, 42*, 810–815.

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement, 63*(4), 584–601. doi:10.1177/0013164403251284.

Jalili, M., Hejri, S. M., & Norcini, J. J. (2011). Comparison of two methods of standard setting: The performance of the three-level Angoff method. *Medical Education, 45*(12), 1199–1208. doi:10.1111/j.1365-2923.2011.04073.x.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*(4), 319–342. doi:10.1111/j.1745-3984.2001.tb01130.x.

Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L., & Van Der Vleuten, C. (2003). Comparison of a rational and an empirical standard setting procedure for an OSCE. *Medical Education, 37*(2), 132–139. doi:10.1046/j.1365-2923.2003.01429.x.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: Manual for setting standards of performance eonducational and occupational tests*. Princeton, NJ: Educational Testing Service.

Messick, S. (1995a). Standards of validity and the validity of standards in performance asessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8. doi:10.1111/j.1745-3992.1995.tb00881.x.

Messick, S. (1995b). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*(1), 3–19. doi:10.1177/001316445401400101.

Norcini, J. (2003). Setting standards on educational tests. *Medical Education, 37*(5), 464–469. doi:10.1046/j.1365-2923.2003.01495.x.

Norcini, J., & Shea, J. A. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10*(1), 39–59. doi:10.1207/s15324818ame1001_3.

Prideaux, D., Roberts, C., Eva, K., Centeno, A., Mccrorie, P., Mcmanus, C., et al. (2011). Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33*, 215–223.

Schoonheim-Klein, M., Muijtjens, A., Habets, L., Manogue, M., van der Vleuten, C., & van der Velden, U. (2009). Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *European Journal of Dental Education, 13*(3), 162–171.

Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., et al. (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 conference. *Medical Teacher, 33*(3), 224–233. doi:10.3109/0142159X.2011.551558.

Schuwirth, L., & van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education, 46*(1), 38–48. doi:10.1111/j.1365-2923.2011.04098.x.

Wayne, D. B., Fudala, M. J., Butter, J., Siddall, V. J., Feinglass, J., Wade, L. D., et al. (2005). Comparison of two standard-setting methods for advanced cardiac life support training. *Academic Medicine, 80*(10), S63–S66.

Wilkinson, T., Newble, D., & Frampton, C. (2001). Standard setting in an objective structured clinical examination: Use of global ratings of borderline performance to determine the passing score. *Medical Education, 35*, 1043–1049.

Woehr, D. J., Arthur, W., & Fehrmann, M. L. (1991). An empirical comparison of cutoff score methods for content-related and criterion-related validity settings. *Educational and Psychological Measurement, 51*(4), 1029–1039. doi:10.1177/001316449105100423.

Wood, T., Humphrey-Murto, S., & Norman, G. (2006). 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, 11*(2), 115–122. doi:10.1007/s10459-005-7853-1.

Zieky, M. J., & Livingston, S. A. (1977). *Basic skills assessment. Manual for setting standards on the basic skills assessment tests*. New Jersey Princeton, NJ: Educational Testing Service.