The Objective Borderline method (OBM): A probability-based model for determining an objective Pass/Fail cut-off score for Borderline grades

Boaz Shulruf[1, 2]**,** Rolf Turner[2], Phillippa Poole[2], Tim Wilkinson[3]

University of New South Wales[1]; University of Auckland[2]; University of Otago[3]

Corresponding Author: Associate Professor **Boaz Shulruf** , UNSW Medicine
b.shulruf@unsw.edu.au

Abstract

The decision to pass or fail a student is the most critical in 'high stakes' examinations. This study introduces and validates a new probabilistic-based standard-setting method for determining the pass/fail cut-off score from borderline grades named the Objective Borderline method (OBM). The OBM sets up the cut-score based on the distribution of the Pass, Borderline and Fail grades within a given examination. 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). We used 'Objective Structured Clinical Examination' (OSCE) results from one medical school in New Zealand to establish the pass/fail cut-off scores by the abovementioned three methods. The results indicate that the pass/fail cut-off scores generated by the OBM were similar to those generated by the more established methods ($0.840 < r < 0.998$; $p<.0001$). 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 is simple to use. Moreover, although demonstrated within clinical assessment context, some simulated studies (unpublished) demonstrated that the OBM is context free and is applicable across almost any context with very few limitations.

The Objective Borderline method (OBM): A probability-based model for determining an objective Pass/Fail cut-off score for Borderline grades

Introduction

One of the most challenging tasks in educational assessments is making Pass/Fail decision for borderline performance (Kramer et al., 2003; Patrício et al., 2009; Roberts, Newble, Jolly, Reed, & Hampton, 2006; Schoonheim-Klein et al., 2009; Shulruf, Turner, Poole, & Wilkinson, 2013; Wood, Humphrey-Murto, & Norman, 2006). Making a wrong decision in passing a borderline student could have negative consequences in perpetuating weaknesses in applied knowledge and performance, which in high stake context may literally be life-threatening (for example see: Hays, Sen Gupta, & Veitch, 2008). On the other hand, failing a competent examinee has adverse consequences for the student and is a loss for the society.

Most standard setting methods determine a Pass/Fail decision for Borderline grades by identifying a cut-off score within the borderline range, using statistical/mathematical calculations deemed to be objective (Ben-David, 2000; Cizek, 2012; Cizek & Bunch, 2007). Among the most commonly used methods are the Nedelsky, Ebel, Angoff, Hofstee, Borderline Group, and Regression methods (Ben-David, 2000; Cizek, 2012; Cizek & Bunch, 2007). Nedelsky, Ebel, Angoff and Hofstee methods use expert panels to estimate what a cut-off score should be (Cusimano & Rothman, 2003; Hurtz & Auerbach, 2003; Kaufman, Mann, Muijtjens, & van der Vleuten, 2000; Kramer et al., 2003; Verheggen, Muijtjens, Van Os, & Schuwirth, 2008; Wass, van der Vleuten, Shatzer, & Jones, 2001; Wayne et al., 2005). The Borderline Groups and Regression methods, however, apply statistical techniques to the test scores to set up the cut-off score, without any further judgement (Boursicot, Roberts, & Pell, 2007; Smee, 2001; Wilkinson, Frampton, Thompson-Fawcett, & Egan, 2003). Methods based on experts' judgment are vulnerable to judgment bias and to date no consensus has been reached to determine an optimal way for achieving high test reliability without employing a large number of panellists (Ben-David, 2000; Brannick, Erol-Korkmaz, & Prewett, 2011; Cizek & Bunch, 2007; Hurtz & Auerbach, 2003; Wayne et al., 2005).

The current study introduces a new pass/fail decision-making method for Borderline grades named the *Objective Borderline Model* (**OBM**). The OBM addresses some of the shortcomings of the currently available methods, particularly by minimising judges' biases and improving the objectivity in the decision-making process. Specifically, the OBM is a method that determines the pass/fail cut-off score by using the proportions (used to form probabilities) of Pass, Borderline and Fail grades, to set up a defensible cut-score.

The fundamental underlying assumption of the OBM are: (1) the test examiners may clearly decide three ranges of examination scores (Nedelsky, 1954): clear Pass (P); clear Fail (F); and Borderline grade (B). The Borderline grade 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; and (2) 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 validate the OBM and demonstrate its feasibility and practicality. This is done by comparing the pass/fail cut-off scores generated by the

OBM with those yielded from the Regression and the Borderline Group (BGM) Methods.

The Objective Borderline Model (OBM)

When there is a collection of grades achieved by a group of students in a single examination, each could be classified as either '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 uncertainty as to whether the grade should be Pass or Fail. It is also 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. Assume that the number of the Fails grades is n_F , the number of the Pass grades is n_P , and the number of the Borderline grades is n_F. Then, the probability of a grade taken from the pool of the Fail and Borderline grades to be Borderline is $(P_{r1})= (n_B /(n_F + n_B)$. Similarly the probability of a grade taken from the pool of the Borderline and Pass grades to be Pass is $(P_{r2})= (n_P /(n_B + n_p)$. P_{r1} and P_{r2} are the mathematical expressions of the difficulty to cross each of the respective thresholds (Fail-Borderline and Borderline-Pass), which is similar to the item difficulty when there are only two outcome categories Pass/Fail (Schuwirth & van der Vleuten, 2010).

Since P_{r1} and P_{r1} are, by assumption, independent, the probability that both conditions are met is simply the product $P_r= (P_{r1}) \times (P_{r2}) = (n_B /(n_F + n_B) \times (n_P /(n_B + n_p)$.

Probabilities and proportions are practically interchangeable (DasGupta, 2010), thus reclassifying the top Borderline scores greater than P_r as Pass is justifiable since this proportion of Borderline grades is equal to the probability of being successful in crossing the two thresholds (Fail-Borderline and Borderline-Pass). Therefore, the cut-score should be the lowest Borderline score that was reclassified into Pass.

Note that the OBM utilises the probabilities P_{r1} and P_{r2} for purpose of decision making only. Those probabilities cannot be used to predict scores or grades of any individual or groups since at the time when the OBM was applied the grades had already been known, but a decision about reclassification of the Borderline grades was yet to be made.

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 & Bunch, 2007; Zieky & Livingston, 1977) which was also used by Wilkinson (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 .

The second method was the Regression method (Wood et al., 2006). In the regression model, 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.

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 established a set of ranges by varying the distance from the cut-off score as defined by the mBGM (See Table 1).

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.
05.SD	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.

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

Table 2 shows the proportion (%) of borderline scores identified by each method and the cut-off scores identified for each station classified by method and borderline range.

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 (Table 2). The correlations among the mean cut-off scores of the models was high ($.984 \leq r \leq .998$ $p < .0001$).

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 very different paradigm and statistical method.

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

	Objective Borderline Model (OBM)											
Station	%	1SD	%	0.5SD	%	2SE	%	1SE	%	mBGM	%	Regression
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

Discussion

This study aimed to describe and measure the 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: the modified BGM method and the Regression method (Cizek & Bunch, 2007; Wilkinson et al., 2001; Wood et al., 2006; Zieky & Livingston, 1977). The results indicate that the OBM is as good as the other methods in that it generated very similar cut-off scores to the mBGM and the Regression methods (Table 2). 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 criterion for pass/fail decision within the education assessment context and thus does not overrides the examiners' decision (Nedelsky, 1954). Since this study demonstrated 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.

In comparison to the other standard setting methods, the OBM uses a probabilistic rather than compromised model to sets up the Pass/Fail cut-score. Using a probability model is deemed to be more realistic and modest than other methods to solve a problem of uncertainty, yet it 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 trusts the examiners (or curriculum writers) to set up standards for Pass and Fail and does not let a panel of experts to override those standards; (3) the OBM does not consider hypothetical groups of examinees to set up cut-off scores but rather uses data from the actual examination; (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 high level of statistics, and can be readily 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 in advance how the cut-off score is calculated, which minimises uncertainty and increases a sense of fairness; (7) the OBM does not assume a normal distribution and is not affected by extreme scores; and (8) the OBM tolerates a range of Borderline bands, as shown in Table 2, 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 *possible* Pass and possible Fail with the cut-off score based on the *probability* of Borderline grades to be Pass for each population of examinees. Given the advantages of the OBM this minor shortcoming should not impede educators using it.

Also as indicated in the Methods section the OBM uses probabilities in an unusual way. These probabilities are 'theoretical' as the real scores and grades are already known. However, this is way of modelling applies to all other data-driven standard setting methods (e.g. Regression, Bookmark, BGM) as all those methods use the examination *known* scores to reclassify (to Pass or Fail) the scores within the Borderline range.

It is acknowledged that measuring the validity of test standards is not an easy task nor could it be completely achieved (Cizek & Bunch, 2007; Messick, 1995a, 1995b; Schuwirth et al., 2011). However, it is believed that this study has met most of the relevant desirable criteria to suggest that the OBM is a valid method for pass/fail standard setting as appeared were agreed in Ottawa conference 2010 (Schuwirth et al., 2011) . We established a clear a robust theoretical and statistical rationale for the OBM (recommendations 2, 3, 6 & 8); 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 student; and we compared different standard-setting methods using the same data (recommendation 13). We acknowledge that consequential validity (Messick, 1995b) was not within the scope of this study and we recommend that future research will address that issue.

Our study has a few limitations. The most important is that we used administrative data 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 made 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, yet to some students those changes in the cut-scores might be critical. 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. Last but not least, this study testes the validity of the OBM using OSCE data. Based on the theoretical background presented earlier on we assume that the OBM could be used within any educational context. Testing the OBM in different context or types of assessments (for example MCQ) is not within the scope of this study and should be investigated in future research.

References

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

Boursicot, K., Roberts, T., & Pell, G. (2007). Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education, 41*(11), 1024-1031. doi: 10.1111/j.1365-2923.2007.02857.x

Brannick, Michael T., Erol-Korkmaz, H. Tugba, & Prewett, Matthew. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical Education, 45*(12), 1181-1189. doi: 10.1111/j.1365-2923.2011.04075.x

Cizek, G. (2012). *Setting Performance Standards: Foundations, Methods, and Innovations* (2nd ed.). London: Routledge.

Cizek, G., & Bunch, M. (2007). *Standard Setting: A Guide to Establishing And Evaluating Performance Standards on Tests*. London: Sage Pubns.

Cusimano, M., & Rothman, A. (2003). The Effect of Incorporating Normative Data into a Criterion-Referenced Standard Setting in Medical Education. *Academic Medicine, 78*(10), S88-S90.

DasGupta, A. (2010). *Fundamentals of Probability: A First Course*. London: Springer.

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, Gregory M., & Auerbach, Meredith 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

Kaufman, D., Mann, K., Muijtjens, A., & van der Vleuten, C. (2000). A comparison of standard-setting procedures for an OSCE in undergraduate medical education. *Academic Medicine, 75*(3), 267-271.

Kramer, Anneke, Muijtjens, Arno, Jansen, Koos, Düsman, Herman, Tan, Lisa, & Van Der Vleuten, Cees. (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

Messick, Samuel. (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, Samuel. (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, Leo. (1954). Absolute Grading Standards for Objective Tests. *Educational and Psychological Measurement, 14*(1), 3-19. doi: 10.1177/001316445401400101

Patrício, Madalena, Julião, Miguel, Fareleira, Filipa, Young, Meredith, Norman, Geoffrey, & Vaz Carneiro, António. (2009). A comprehensive checklist for reporting the use of OSCEs. *Medical Teacher, 31*(2), 112-124. doi: doi:10.1080/01421590802578277

Roberts, Chris, Newble, David, Jolly, Brian, Reed, Malcolm, & Hampton, Kingsley. (2006). Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher, 28*(6), 535-543. doi: doi:10.1080/01421590600711187

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, Jerry, Gruppen, Larry, Kreiter, Clarence, Mennin, Stewart, Onishi, Hirotaka, . . . Wagner-Menghin, Michaela. (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher, 33*(3), 224-233. doi: doi:10.3109/0142159X.2011.551558

Schuwirth, L., & van der Vleuten, C. (2010). How to design a useful test: the principles of assessment. In T. Swanwick (Ed.), *Understanding Medical Education: Evidence, Theory and Practice* (pp. 195-207): The Association for the Study of Medical Education.

Shulruf, B., Turner, R., Poole, P., & Wilkinson, T. (2013). The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score for borderline grades in medical education programmes. *Advances in Health Sciences Education, 18*(2), 231-244. doi: 10.1007/s10459-012-9367-y

Smee, S. (2001). Setting standards for an objective structured clinical examination: the borderline group method gains ground on Angoff. *Medical Education, 35*, 1009-1010.

Verheggen, M., Muijtjens, A., Van Os, J., & Schuwirth, L. (2008). Is an Angoff Standard an Indication of Minimal Competence of Examinees or of Judges? *Advances in Health Sciences Education, 13*(2), 203-211. doi: 10.1007/s10459-006-9035-1

Wass, V., van der Vleuten, C., Shatzer, John, & Jones, Roger. (2001). Assessment of clinical competence. *THE LANCET, 357*(9260), 945-949.

Wayne, D. B., Fudala, Monica J., Butter, John, Siddall, Viva J., Feinglass, Joe, Wade, Leonard D., & McGaghie, William C. (2005). Comparison of Two Standard-setting Methods for Advanced Cardiac Life Support Training. *Academic Medicine, 80*(10), S63-S66.

Wilkinson, T., Frampton, Christopher M., Thompson-Fawcett, Mark, & Egan, Tony. (2003). Objectivity in Objective Structured Clinical Examinations: Checklists Are No Substitute for Examiner Commitment. *Academic Medicine, 78*(2), 219-223.

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.

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.