Higher Education Studies; Vol. 5, No. 4; 2015
ISSN 1925-4741 E-ISSN 1925-475X
Published by Canadian Center of Science and Education

Using Student Ability and Item Difficulty for Making Defensible Pass/Fail Decisions for Borderline Grades

Boaz Shulruf[1], Phil Jones[1] & Rolf Turner[2]

[1] Office of Medical Education, University of New South Wales, Sydney, Australia

[2] Department of Statistics, University of Auckland, Auckland, New Zealand

Correspondence: Boaz Shulruf, Office of Medical Education, UNSW Medicine, University of New South Wales, Sydney, Australia. Tel: 612-9385-2693. E-mail: b.shulruf@unsw.edu.au

Received: May 17, 2015 Accepted: June 17, 2015 Online Published: July 21, 2015

doi:10.5539/hes.v5n4p106 URL: http://dx.doi.org/10.5539/hes.v5n4p106

Abstract

The determination of Pass/Fail decisions over Borderline grades, (i.e., grades which do not clearly distinguish between the competent and incompetent examinees) has been an ongoing challenge for academic institutions. This study utilises the Objective Borderline Method (OBM) to determine examinee ability and item difficulty, and from that reclassifying each Borderline grade as a Pass or Fail. Using the OBM, examinees' Borderline grades from a clinical examination were reclassified into Pass or Fail. The predictive validity of this method was estimated by comparing the examination original and reclassified grades to each other and to subsequent clinical examination results. The new model appeared as more stringent ($p<.0001$) than the original decisions. Implications for educators and policy makers are discussed. The OBM2 is found to provide a plausible solution for decision making over borderline grades in non-compensatory assessment systems.

Keywords: borderline grades, board of bxaminers, examinations, decision making

1. Introduction

1.1 Standard Setting and Decision Making in Higher Education

One of the most challenging tasks in clinical assessments is the Pass/Fail decision for borderline performance (Kramer et al., 2003; Patrício et al., 2009; Schoonheim-Klein et al., 2009; Shulruf, Turner, Poole, & Wilkinson, 2013; Wood, Humphrey-Murto, & Norman, 2006). This challenge is particularly difficult since many types of clinical examination include a "Borderline performance" in their marking sheets (Boursicot, Roberts, & Pell, 2007; Roberts, Newble, Jolly, Reed, & Hampton, 2006; Schoonheim-Klein et al., 2009; Wilkinson, Newble, & Frampton, 2001). As many clinical examinations are of high stake (Shumway & Harden, 2003), it is essential to make an accurate Pass/Fail decision which does not fail a competent student and does not pass an incompetent student, particularly given evidence that borderline students tend to remain underachieving throughout their studies (Pell, Fuller, Homer, & Roberts, 2012). The General Medical Council (2009a, 2009b) also expressed its concerns about assessment and standard setting practices in medical programmes in the United Kingdom. To address this critical issue a plethora of standard setting methods have been introduced and implemented in a range of clinical examinations (Boulet, De Champlain, & McKinley, 2003; Jalili, Hejri, & Norcini, 2011; Shulruf et al., 2013; Wass, Vleuten, Shatzer, & Jones, 2001; Wilkinson et al., 2001). Nonetheless despite this range of methods concerns about reliability, validity and acceptability remain (Ben-David, 2000; Brannick, Erol-Korkmaz, & Prewett, 2011) particularly within the context of clinical assessment where clinical examiners tend to avoid failing students and trainees (Cleland, Knight, Rees, Tracey, & Bond, 2008; Dudek, Marks, & Regehr, 2005; Morton, Cumming, & Cameron, 2007; Rees, Knight, & Cleland, 2009).

Most standard setting methods determine a Pass/Fail decision for Borderline grades by identifying a cutoff score within the borderline range by 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 cutoff score should be (Cusimano & Rothman, 2003; Geisinger & McCormick, 2010; Hurtz & Auerbach, 2003; Kaufman, Mann, Muijtjens, & Vleuten, 2000; Kramer et al., 2003; Verheggen, Muijtjens, Van Os, & Schuwirth, 2008; Wass et al.,

2001; Wayne et al., 2005), whereas the Borderline Group and Regression methods use only the test scores without any additional post examination judgment (Boursicot et al., 2007; Shulruf et al., 2013; Smee, 2001; Wilkinson, Frampton, Thompson-Fawcett, & Egan, 2003). Methods based on experts' judgment are susceptible to judgment bias and to date no consensus has been reached on an optimal way to achieve high reliability without recruiting a large number of experts (Ben-David, 2000; Brannick et al., 2011; Chang, Dziuban, Hynes, & Olson, 1996; Cizek & Bunch, 2007; Hurtz & Auerbach, 2003; Skorupski & Hambleton, 2005; Wayne et al., 2005).

Determining cutoff scores for Borderline grades by a combination of objective and subjective scores is commonly used since many clinical examinations employ marking sheets which include scores for particular tasks (deemed as "objective") as well as a global rating ("subjective") (Ben-David, 2000; Cizek & Bunch, 2007; Roberts et al., 2006; Wilkinson et al., 2003; Wilkinson et al., 2001). Such methods, although deemed to be reasonably objective, include an inherent flaw where the same sum of all the "objective" scores i.e. specific skills could be classified as "Pass", "Borderline" or "Fail" on the subjective score (global performance) for different examinees (Boursicot, Roberts, & Pell, 2006; Boursicot et al., 2007; Cizek & Bunch, 2007; Kellow & Willson, 2008; Norcini, 2003; Shulruf et al., 2013; Wood et al., 2006). It is therefore not surprising that significant inconsistencies have been found when these two types of standard setting methods were compared (Kaufman et al., 2000; Kramer et al., 2003; Lagha, Boscardin, May, & Fung, 2012; Wayne et al., 2005).

1.2 Modern Test Theory and Decision Making

Modern test theories such as item response theory (IRT) have become more commonly used in medical education (Boulet et al., 2003; Downing, 2003). IRT methods have been mostly used for improving the quality of written test items rather than determining Pass/Fail cutoff scores for clinical examinations (Downing, 2003; Schuwirth & Vleuten, 2010), or helping to calibrate test items before applying other commonly used standard setting methods (Boulet et al., 2003; Ferdous & Plake, 2008; Grosse & Wright, 1986; MacCann & Stanley, 2006; Wang, Wiser, & Newman, 2001). The most advanced standard-setting method that uses IRT framework for determining Pass/Fail cutoff score is the Bookmark method (Buckendahl, Smith, Impara, & Plake, 2002; Karantonis & Sireci, 2006; Lewis, Mitzel, & Green, 1996; Peterson, Schulz, & Engelhard Jr., 2011). Nonetheless, the Bookmark method has been criticized mainly for being resource intensive and the use of arbitrary value (.67 probability of success) to establish the point that is used to rank order items for the judges' booklets, which may explain why it has not been widely used for setting cutoff scores in clinical examinations (Karantonis & Sireci, 2006; Lewis et al., 1996). It is also noted that although findings suggest that the Bookmark is preferable over Angoff method (Peterson et al., 2011), concerns about judges' ability to make reliable decisions despite that additional information remain (Davis-Becker, Buckendahl, & Gerrow, 2011; Deunk, Van Kuijk, & Bosker, 2014).

1.3 The Objective Borderline Method: An Alternative Method for Decision Making over Borderline Grades

In response to the abovementioned challenges, Shulruf et al. (2013) recently introduced the Objective Borderline Method (OBM) which is based upon a measure of difficulty of the examination in question, formed from the initial results that the students' actually obtained on this examination. There are two separate but related and fairly natural measures of difficulty available (Raykov & Marcoulides, 2011; Sax & Reade, 1964). One seeks to combine these two measures into a single measure. There are innumerable ways in which this might be done. A simple, plausible and perhaps intuitively way is to think of the two initial measures (which are both numbers between 0 and 1 and generated from observed proportions of those two categories) as being notionally the probabilities of success on two independent tests or experiments. The measure is formed simply as the product of these two probabilities and may thus be conceptualized as the probability of success on both tests. It must be emphasized here that these two independent tests are conceptual only. However, they serve as a useful heuristic guide to our thinking in constructing the combined measure. This combined measure is just an index (similar to other indices e.g. BMI) that its validity is determined only by its usefulness. This combined probability or index is by no mean the probability of the occurrence of any actual event (Shulruf et al., 2013).

Explicitly the initial results consist of a number of Fail, Borderline, and Pass grades. The first notional test consists of drawing a grade at random from the collection of all Fail and Borderline grades. "Success" is considered to be drawing a Borderline grade, and the first measure of difficulty is the probability of drawing a Borderline grade (i.e. the observed proportion of Borderlines among the pool of Borderlines and Fails). The second notional test consists of drawing a grade at random from the collection of all Borderline and Pass grades. Here "success" is considered to be drawing a Pass grade (i.e. the observed proportion of Passes among the pool

of Passes and Borderlines). Each of these two tests (achieving Borderline rather than Fail and achieving Pass rather than borderline) is a common measure of difficulty when a test includes two categories (Raykov & Marcoulides, 2011; Sax & Reade, 1964). If the numbers of Fail, Borderline and Pass Grades are n_P, n_B, n_F respectively then the probability of success on the first notional test is $P_{r1} = n_B / (n_F + n_B)$ and the probability of success on the second notional test is $P_{r2} = n_P / (n_B + n_p)$. The combined measure of difficulty is then $P_r = P_{r1} \times P_{r2} = (n_B / (n_F + n_B) \times (n_P / (n_B + n_p)$. Figure 1 schematically illustrates the how the OMB index is calculated.

The OBM utilizes P_r in such a way that it assigns conceded Pass to the proportion of Borderline grades equals to P_r. and conceded Fail to the remaining Borderline grades.

It is acknowledged that like all standard setting models the OBM is derived from some arbitrary premises (Cizek, 1993). However, Shulruf et al. (2013) demonstrated that the OBM is at least as effective as other standard setting methods (e.g. the Regression and the Borderline Groups methods).

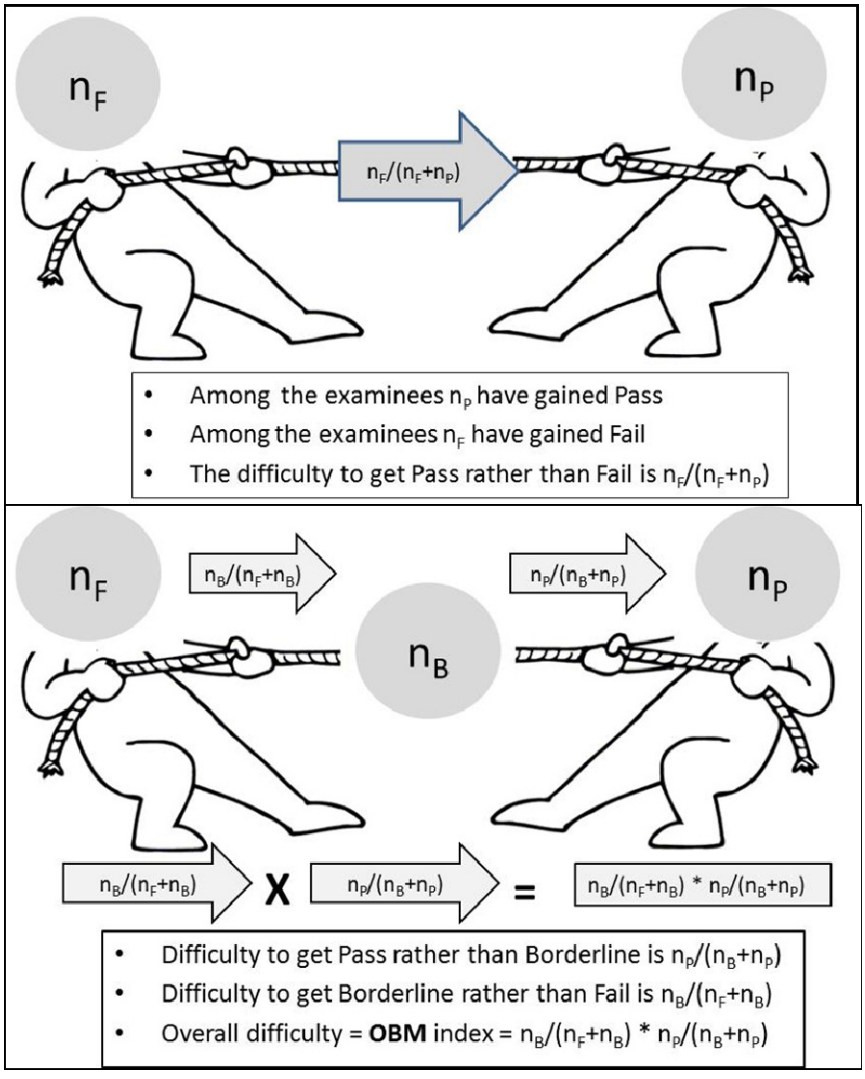

Figure 1. The principle of the OBM: combining indices of difficulty

1.4 The Overarching Objectives of the Current Study

The current study introduces a modification to the OBM model which enables making pass/fail decisions for any types of marks (continuous or categorical) as long as marks can be initially classified into three categories: Pass,

Borderline and Fail and the number of Passes is greater than zero. It offers a practical and theoretically defensible method to determine which of the Borderline grades, within a categorical set of grades, should be considered as Pass and which should be considered as Fail. The improved model is named the Objective Borderline Method 2 (OBM2) as it uses two measures (examinee ability and item difficulty) to determine whether a Borderline grade should be reclassified as Pass or Fail. Unlike OBM, the OBM2 *does not establish a cut-score* but it determines whether a Borderline grade should be Pass on Fail on a case by case basis. Such a solution may betefit any panel of examineres who need to make pass/fail decision over borderline grade for non-compensatory assessment systems, where high score in one domain cannot compensate for a low score on another. A thorough review that took place in the preparation of this study failed to identify any such method.

1.5 The Objective Borderline Method 2 (OBM2)

The original OBM estimates the combined probability (P_r) of being successful in two notional tests based on the counts of Passes, Borderlines, and Fails (n_P n_B n_F respectively) for a set of examination scores of a group of examinees. OBM2 uses the same approach as the OBM but at the item rather than the examination level, assuming all items are unidimensional (Hattie, 1985).

Consequently, when a group of examinees are assessed using a set of unidimensional items and their performance is classified as Pass, Borderline or Fail, it is possible to calculate two different combined probabilities (i.e. P_r) for each Borderline grade. The first is based on the particular examinee's grades across all items (referred to as the examinee's P_r and denoted as P_e); the second P_r is based on the grades for a particular item across all examinees (referred to as the item's P_r and denoted as P_i). Analogous to Item Response Theory (Kolen & Brennan, 2004), P_e is a measure of examinee's ability and P_i is a measure of item difficulty. The relationship between these two probabilities (P_e and P_i) can be used to determine whether the Borderline grade should be conceded Pass or Fail. This relationship is expressed by a decision index (P_d), which is the quotient

$P_d = (P_e) / ((P_e) + (P_i))$.

When $P_d \geq .5$ it means that the examinee's ability is greater than or equal to item difficulty hence the Borderline grade should be conceded Pass. Note that when $P_d = .5$ the Pass/Fail decision cannot be determined by this index. In this case the decision must be determined by a pre-specified policy.

The current study aims to estimate the validity of the OBM2 by examining what the consequences would be if Borderline grades of medical students' clinical examination were reclassified as Pass or Fail using the OBM2.

2. Methods

2.1 Data

The UNSW Medicine program is a 6 year undergraduate entry program organized into three phases, each comprised of two academic years. At the end of each phase, students must pass a clinical skills examination before progressing to the next phase. This study used data from the Phase 1 and Phase 2 clinical examinations from five cohorts of students.

Each of the Phase 1 and Phase 2 clinical skills examinations comprises six standardised stations (for more details on the curriculum and clinical assessments see: McNeil, Hughes, Toohey, & Dowton, 2006). The students are assessed in nine criteria encompassing generic communication skills, clinical history skills and physical examination skills. A standard grading sheet is used at each station with additional specific descriptors relevant to the station's tasks. A common 4-point grading system is used for each criterion: Fail, Borderline, Pass and Exceptional. The examiners do not provide a global grade for the station. A Pass/Fail decision for each station is based on the proportion of Fail and Borderline grades—failing a station is a result of at least two Fail grades or a combination of one Fail grade and more than two Borderline grades. A Pass/Fail decision for the examination is based on the number of failed stations—students must pass at least three stations. Each grade is also converted to a numerical score (with Borderline representing 50% of maximum score); a Fail decision is also made if a student's total numerical score is <50%. Students who fail the Phase 1 clinical skills examination are offered a supplementary examination after a period of remediation. Students who fail the supplementary examination are excluded from the program.

2.2 Sample

Test data were available from 1,136 students who sat the Phase 1 clinical examination. Of these students, 42 did not progress to the Phase 2 clinical examination and their grades in Phase 2 clinical examination were considered in our analysis as Fail. This inclusion was based on data not presenting in this study suggesting that the discontinuation of those students was due to dissatisfactory performance in their clinical and non-clinical studies in Phase 1. Thus, this analysis includes all 1,136 students (Y2004, N=210; Y2005, N=229; Y2006, N=226; Y2007, N=238; Y2008, N=233). Demographic data such gender, age or ethnicities were not included in the dataset and the analysis as they were not deemed relevant to the model discussed.

2.3 Analysis

The first analysis employed factor analysis of raw examination scores within each station to ensure unidimensionality of the items (Hattie, 1985). Then, within each station the decision index (P_d) was calculated leading to the assignment of Pass/Fail to each Borderline grade that was originally given to a student for a performance criterion within a station. Next, based on the OBM2 reclassification of grades, pass fail decisions for the whole clinical examination (all six stations) were calculated in the way described above (students must pass at least three stations and total score from all stations must exceed 50%).

The last stage compared the predictive validity of the original grades in the Phase 1 clinical examination with the reclassified grades derived by the OBM2. The sensitivity, specificity, positive and negative predictive values, and accuracy (overall fraction correct) of the Phase 1 grades for predicting performance in the subsequent Phase 2 clinical examination were measured (see Table 1) (Bossuyt, 2011).

Table 1. Definition of true positive, true negative, false positive and false negative (adapted from Bossuyt, 2011)

		Clinical examination results Phase 2	
		Pass	Fail
Pass/Fail Decision for Borderline grades in clinical examination Phase 1	Pass	True Positive (TP)	False Positive (FP)
	Fail	False Negative (FN)	True Negative (TN)
Sensitivity	TP/(TP+FP)		
Specificity	TN/(TN+FP)		
Positive predictive value	TP/(TP+FP)		
Negative predictive value	TN/(TN+FN)		
Accuracy	(TP+TN)/(TP+TN+FP+FN)		

3. Results

3.1 Suitability of the Data

The results indicate that the data did not fully meet the criteria for unidimensionality as in some stations the items were loaded on two factors. Nonetheless, Table 2 suggests that within each station there is only one meaningful underlying factor since none of the factor loadings in any of the six stations met the criteria for two discrete factor structure (Pett, Lackey, & Sullivan, 2003) and the variance explained by the first factor was between 28 and 35 percent whereas the second factor explained no more than 6%. Thus, it was decided to carry on with the analysis, particularly given the high level of internal consistency within each station (Cronbach's alpha= .80, .80, .77, .82, .79, .82 for Stations 1 to 6 respectively).

Table 2. Factor Matrix (ML)

Station		1		2		3		4		5		6	
Factor		1	2	1	2	1	2	1	2	1	2	1	2
Assessment criterion	1	.651	-.338	.641	-.174	.686	-.374	.682		.749	-.301	.669	-.058
	2	.599	-.256	.616	-.343	.601	-.090	.673		.652	-.190	.640	-.153
	3	.571	-.125	.586	.070	.545	.165	.616		.609	-.013	.598	-.236
	4	.569	.158	.576	-.280	.540	.270	.599		.551	.246	.580	.192
	5	.561	.340	.563	.037	.538	-.237	.587		.549	.085	.576	-.125
	6	.556	-.098	.562	.292	.498	.110	.540		.479	.381	.574	-.225
	7	.552	.418	.536	.289	.471	.287	.538		.459	-.049	.547	-.097
	8	.550	.071	.533	.268	.468	.185	.531		.437	.094	.544	.329
	9	.491	-.067	.495	.090	.433	.321	.527		.404	.459	.543	.401
Variance explained (%)		32.3	5.9	32.4	5.3	28.7	5.9	34.9		30.6	6.2	34.5	5.2

The average percentage of Borderline grades that were reclassified as Pass (by criterion by station) was 25.8% (range 0.0-58.8%). The comparison of the Pass/Fail decisions of the Phase 1 clinical examination across the original grades indicates that the OBM2 model was more stringent than the original decision, yet the decisions made by the OBM2 had high level of agreement with the "original decisions" (decisions made by the board of examination within the institute) (Accuracy=.88) (Table 3).

3.2 Models Comparison

Table 3. Distribution of final pass fail grades by decision model

		Original Decision		
		Fail	Pass	Accuracy*
OBM2	Fail	9	132	0.88
	Pass	0	995	
	Total			1136

Note. *Accuracy= overall fraction correct (proportion of agreement out of all grades)

The quality of the OBM2 was estimated by comparing the overall clinical examination grades in the Phase 2 clinical examination with the overall outcomes of the Phase 1 clinical examination as calculated in two ways: by the original method and by the OBM2 model.

Table 4. Distribution of Phase 2 outcomes by Phase 1 outcome by type of decision

		Phase 2	
		Pass	Fail
Original decision	Pass	1056	71
	Fail	4	5
OBM2	Pass	945	50
	Fail	115	26

Table 5. Quality indices OBM2 vs. original decisions

Index	OBM2	Original decision
Specificity	.342	.066
Sensitivity	.892	.996
False Positive	.044	.063
False Negative	.101	.004
Accuracy	.855	.934

The results indicate that the original decision yielded accuracy of .93 and sensitivity of .99 but specificity of only .07. The *OBM2* model was less accurate but as a more stringent model it yielded the higher level of specificity (.34). 71 (6.6%) students passed the Phase 1 clinical examination based on the original decision but failed in Phase 2. In comparison, the *OBM2* model passed only 50 (4.6%). The cost of increasing the specificity was that the *OBM2* model resulted in failing 115 (11%) of students in the Phase 1 clinical examination who were later successful in the Phase 2 clinical examination.

4. Discussion

The main objective of this study was to utilise the recently introduced Objective Borderline Model (OBM) (Shulruf et al., 2013) for supporting pass/fail decisions for students who performed at the borderline level in their clinical examination. This was achieved by modifying the OBM to incorporate two measures (examinee ability and item difficulty) for determining whether a Borderline grade should be reclassified as Pass or Fail. In order to provide robust evidence, this study followed the relevant recommendations for research on assessment from the Ottawa 2010 Conference (Schuwirth et al., 2011): (a) basing the research on robust scientific theory (recommendation 7, 8, and 9); (b) taking the modern approach for validity by looking at consequential validity rather than merely comparing one method with another (recommendations 12, 13); (c) adopting the Item Response Theory (IRT) conceptual framework in the development of a new method (recommendation 18). We note that recommendation 18 was only partially followed as OBM2 applies only one feature analogous to IRT which is the comparison of student ability with item difficulty and in no way it is suggested that IRT models were applied in this study/model.

The OBM and OBM2 introduce a new concept in the field of standard setting by "legitimising" the category of a Borderline grade. The underlying assumption is that a Borderline grade is one of which indicates that the examinee's assessed performance could not clearly be classified as either Pass or Fail and this is a category by its own right (Jalili et al., 2011; Norcini, Shea, & Kanya, 1988). Furthermore, the OBM2 is a plausible solution for making decisions when the data suggest uncertainty (Draper, 2005; Ramsey, 1926). Nonetheless the OBM2 is not a standard setting method in the sense that it does not set any cut-score but only provides evidence-based indication whether a borderline grade should be conceded Pass or Fail.

The underlying assumption of previous standard setting methods is that there is an inevitable misclassification of examinees' proficiency where some truly proficient examinees are mistakenly classified as not proficient (False Negative) and others who truly did not reach the appropriate proficiency level are mistakenly classified as proficient (False Positive) (Cizek, 2012; Cizek & Bunch, 2007). The OBM and OBM2 methods address this concept of misclassification by determining the range of Borderline grades as the range where the level of competency could not be classified *without any doubts* as either clear Pass or clear Fail (Shulruf et al., 2013). This method of classification applies to the determination of the Pass/Fail scores (the definition of in/competency) and the actual classification of examinee's performance by the examiners (Kane, 1994). Skorupski and Hambleton (2005) for example, demonstrated that the majority of panelists engaged in the item mapping standard-setting method reported having difficulty distinguishing between performance categories.

Thus enabling examiners or judges to use a "Borderline" category and accepting the uncertainty of such a category might be an appropriate approach rather than forcing them to make a decision based on limited information. The actual decision whether a Borderline grade should be reclassified as Pass or Fail would then be decided by all data points available which deemed to be more reliable.

The question of where one should set the cutoff point—employing a stringent policy by granting the final Pass for the clear Pass (minimizing the number of False Positives) or taking a more lenient policy and granting a final Pass to those who did not clearly fail (minimizing False Negative) needs to be decided. This could be decided either by

an agreed panelists' opinion who apply judges-based standard setting or by policy makers who decide which test-based (panelist free) standard setting are to be used (Kane, 2013). Since each test-based standard setting applies different mathematical procedure, the results are expected to be somewhat different even if applied on the very same data (Wood et al., 2006). Consequently no standard setting method, including the OBM/OBM2, could be absolutely objective.

In this study we investigated the impact of the OBM2 with respect to a policy that aims to maximize specificity and minimise the number of False Positives. The results clearly demonstrate that both indices would have been improved (in accordance with our pre-determined policy) had OBM2 been implemented. Note that progression to Phase 2 was based on the original decisions (decisions made by the Board of Examination) and thus this comparison is somewhat problematic. However, this limitation would apply to any study using real data which resulted in similar decision making. It is noteworthy, however, that had the OBM2 been used, the specificity would have increased from 7% to 43% resulting with a trade-off of a drop in the sensitivity from 99% to 89% which overall, based on our view, is a preferable outcome for the chosen policy.

It is evident that the OBM2 model is more stringent than the original decisions made by the Board of Examination for determining Pass/Fail and would increase the number of students failing the clinical skills examination. However, this is expected and perhaps even desirable. Clinical examiners tend to avoid failing students and trainees particularly as they give borderline students the benefit of the doubt (Cleland et al., 2008; Dudek et al., 2005; Morton et al., 2007; Rees et al., 2009). Such practice is argued to have the potential for major adverse impact on medical practice (Albanese, 1999). Moreover, in their comprehensive review of sources of bias in clinical performance rating, Williams, Klamen, and McGaghie (2003) summarized compelling evidence suggesting that the tendency for leniency is pretty much embedded in clinical assessment practices with little to no impact of training on such examiners' bias. Consequently, applying the OBM2 model where a Borderline grade is made a legitimate and well defined category (when neither clear Pass nor clear Fail may confidently granted) would help examiners avoiding the leniency bias and minimize passing incompetent examinees (Kane, 1994).

Currently, most grading criteria describe what a competent examinee should demonstrate but fail to define what constitutes borderline performance. Even when the borderline performance is defined, the descriptors are vague, indecisive and poorly correlate with the checklist scores (Pell, Fuller, Homer, & Roberts, 2010). The description of clear Pass and clear Fail criteria aligned to *measureable* teaching and learning objectives would provide transparent expectations to examinees.

An important contribution of the OBM2 is that it provides Pass/Fail decisions for Borderline grades when the grading system does not use continuous scales but only ordinal categories (e.g. Fail, Borderline, Pass, Excellent). This is an advantage of the OBM2 compared to other standard setting methods, which do not have that capacity, particularly when the use of a continuous scale makes little sense if any. Moreover, many of the previous standard setting models assume that the categories (for example in OSCE stations) are points on an interval scale (Boursicot et al., 2007; Cizek & Bunch, 2007; Kramer et al., 2003) although that assumption receives little support from the statistical and educational measurement literature, particularly when the number of the ordinal categories is fewer than six (Agresti, 2010; Torra, Domingo-Ferrer, Mateo-Sanz, & Ng, 2006).

How acceptable is the OBM2? Although difficult to judge at this stage, there are some indications that it should be easily acceptable. First, it is very easy to use and requires no mathematical/statistical skills (see Appendix 2). Second, the OBM2 does not add any cost to current programs, except the time required revising the Pass/Fail criteria as described above, which is negligible compared to other methods using panels of experts (Cizek & Bunch, 2007). Third, the analogy between OBM2 and IRT (both use item difficulty and examinee ability) might be appealing. This analogy is a major advance in the field because by definition a Borderline grade by itself includes very little information, only that it is neither clear Pass nor clear Fail. All other information relevant to the examinee's performance on the "borderline item" is embedded in their performance on all other items encompassed within the same dimension (Cronbach, 1951). Furthermore, the inclusion of item difficulty in the OBM2 method acts as a correction for examiner's leniency/stringency bias. The Pass/Fail decision is determined by the comparison of two calculated probabilities. As clearly observed in [Equation 2], the harder the item the smaller the P_i thus the greater the P_d (and vice versa for an easier item). This correction enhances the fairness of the examination, and removes concerns over examiners' bias, which has its most critical impact on borderline performance.

Obviously the OBM2 raises some challenges which need to be addressed. The OBM2 compares two calculated measures conceptualized as probabilities. However, a small number of items may affect its resolution, which if not fine enough might impair its effectiveness. In this study there were nine items (criteria) for each station which yielded 25 unique values of P_{ri}, providing sufficient resolution for calculating (P_d). Six items, however, would yield only 12 unique values of P_{ri}. Hence we recommend that the OBM2 should be used only when six or more items are included. Nonetheless, further investigation is needed to determine the impact of the number of items on the OBM2 outcomes.

More limitations are related to the study itself. This study used test data from five past clinical examinations (five cohorts). The grading sheets defined Fail criteria based only on curriculum objectives which were deemed to suffice for this pilot study. Since no practical decision was made based on the OBM2, this deviation from the suggested practice is minor. It is therefore, recommended that further studies take a prospective approach to ensure that Pass and Fail criteria are defined as described earlier in this paper. The other minor limitation is that the 42 students who did not progress to Phase 2 were deemed to have failed the clinical examination in that Phase. This decision was made as most of these students did not continue due to poor performance in Phase 1. Since no information on performance in Phase 2 was available, any imputation of data to the clinical examination results of Phase 2 would anyway be based on performance in Phase 1. Therefore, given the low number of failures in the programme, it was believed that including those students in the analysis and assigning them a Fail outcome in the Phase 2 clinical examination would be the most plausible approach.

An important feature of the OBM2 is that it is applicable to non-compensatory assessment systems, where high score in one domain cannot compensate for a low score on another. No previous study was found in the literature to address decision making over borderline grades for such assessment systems. It is acknowledged that this is the first step only and further research may yield better formulae/indices to support decision making over borderline grades either within or beyond the OBM framework.

5. Conclusion

Michael Kane's definition of validity provides some important insight into this the research on standard setting: "Validity is a property of the interpretations assigned to test scores, and these interpretations are considered valid if they are supported by convincing evidence" (Kane, 2013, p. 56). Like all other methods, the OBM2 has advantages and shortcomings and they have been discussed above in detail. Whether the evidence provided in this study to support the validity of the OBM2 is sufficiently convincing is left to the readers to judge. Nonetheless, unless *empirically* proved otherwise, the OBM2 is a *plausible* method for supporting pass/fail decision making for borderline grades, particularly when a non-compensatory assessment system is applied and the risk of passing incompetent examinees who received Borderline grades is of a major concern.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*. Hoboken, N.J.: Wiley. http://dx.doi.org/10.1002/9780470594001

Albanese, M. (1999). Rating educational quality: Factors in the erosion of professional standards. *Academic medicine: Journal of the Association of American Medical Colleges*, *74*(6), 652-658. http://dx.doi.org/10.1097/00001888-199906000-00009

Ben-David, M. (2000). AMEE guide no. 18: Standard setting in student assessment. *Medical Teacher*, *22*(2), 120-130. http://dx.doi.org/10.1080/01421590078526

Bossuyt, P. (2011). Defining biomarker performance and clinical validity. *Journal of Medical Biochemistry*, *30*(3), 193-200. http://dx.doi.org/10.2478/v10011-011-0028-0

Boulet, J., De Champlain, A., & McKinley, D. (2003). Setting defensible performance standards on osces and standardized patient examinations. *Medical teacher*, *25*(3), 245-249. http://dx.doi.org/10.1080/0142159031000100274

Boursicot, K., Roberts, T., & Pell, G. (2006). Standard setting for clinical competence at graduation from medical school: A comparison of passing scores across five medical schools. *Advances in Health Sciences Education*, *11*(2), 173-183. http://dx.doi.org/10.1007/s10459-005-5291-8

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. http://dx.doi.org/

10.1111/j.1365-2923.2007.02857.x

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

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of angoff and bookmark standard setting methods. *Journal of Educational Measurement*, *39*(3), 253-263. http://dx.doi.org/10.1111/j.1745-3984.2002.tb01177.x

Chang, L., Dziuban, C., Hynes, M., & Olson, A. (1996). Does a standard reflect minimal competency of examinees or judge competency? *Applied Measurement in Education*, *9*(2), 161. http://dx.doi.org/10.1207/s15324818ame0902_5

Cizek, G. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, *30*(2), 93-106. http://dx.doi.org/10.1111/j.1745-3984.1993.tb01068.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.

Cleland, J. A., Knight, L. V., Rees, C. E., Tracey, S., & Bond, C. M. (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical Education*, *42*(8), 800-809. http://dx.doi.org/10.1111/j.1365-2923.2008.03113.x

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. http://dx.doi.org/10.1007/BF02310555

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. http://dx.doi.org/10.1097/00001888-200310001-00028

Davis-Becker, S., Buckendahl, C., & Gerrow, J. (2011). Evaluating the bookmark standard setting method: The impact of random item ordering. *International Journal of Testing*, *11*(1), 24-37. http://dx.doi.org/10.1080/15305058.2010.501536

Deunk, M. I., Van Kuijk, M. F., & Bosker, R. J. (2014). The effect of small group discussion on cutoff scores during standard setting. *Applied Measurement in Education*, *27*(2), 77-97. http://dx.doi.org/10.1080/08957347.2014.880441

Downing, S. M. (2003). Item response theory: Applications of modern test theory in medical education. *Medical Education*, *37*(8), 739-745. http://dx.doi.org/10.1046/j.1365-2923.2003.01587.x

Draper, D. (2005). *Thinking about uncertainty: An introduction to probability and statistics*. University of California, Santa Cruz.

Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to fail: The perspectives of clinical supervisors. *Academic Medicine*, *80*(10), S84-S87. http://dx.doi.org/10.1097/00001888-200510001-00023

Ferdous, A., & Plake, B. (2008). Item response theory–based approaches for computing minimum passing scores from an angoff-based standard-setting study. *Educational and Psychological Measurement*, *68*(5), 778-796. http://dx.doi.org/10.1177/0013164407312605

Geisinger, K., & McCormick, C. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*(1), 38-44. http://dx.doi.org/10.1111/j.1745-3992.2009.00168.x

General Medical Council. (2009a). *Quality assurance of basic medical education: Report on dundee medical school, university of dundee*. General Medical Council.

General Medical Council. (2009b). *Quality assurance of basic medical education: Report on st george's medical school, university of london*. General Medical Council.

Grosse, M. E., & Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation & the Health Professions*, *9*(3), 267-285. http://dx.doi.org/10.1177/016327878600900301

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and ltenls. *Applied Psychological*

Measurement, *9*(2), 139-164. http://dx.doi.org/10.1177/014662168500900204

Hurtz, G., & 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. http://dx.doi.org/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. http://dx.doi.org/10.1111/j.1365-2923.2011.04073.x

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461. http://dx.doi.org/ 10.3102/00346543064003425

Kane, M. (2013). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53-88). London: Routledge.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice*, *25*(1), 4-12. http://dx.doi.org/ 10.1111/j.1745-3992.2006.00047.x

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. http://dx.doi.org/10.1097/00001888-200003000-00018

Kellow, T., & Willson, V. (2008). Setting standards and establishing cut scores on criterion-referenced assessments some technical and practical considerations. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 14-28). Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781412995627.d4

Kolen, M., & Brennan, R. L. (2004). *Item response theory methods Test equating, scaling, and linking*. Springer New York. http://dx.doi.org/10.1007/978-1-4757-4310-4

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. http://dx.doi.org/10.1046/j.1365-2923.2003.01429.x

Lagha, R. A., Richter, B., Christy, K., May, W., & Fung, C. C. (2012). A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic Medicine*, *87*(8). http://dx.doi.org/10.1097/ACM.0b013e31825cea4b

Lewis, D., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A bookmark approach*. Paper presented at the The 1996 Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

MacCann, R. G., & Stanley, G. (2006). The use of Rasch modeling to improve standard setting. *Practical Assessment Research and Evaluation*, *11*(2), 1-17. http://pareonline.net/getvn.asp?v=11&n=2

McNeil, P., Hughes, C., Toohey, S., & Dowton, B. (2006). An innovative outcomes-based medical education program built on adult learning principles. *Medical Teacher*, *28*(6), 527-534. http://dx.doi.org/10.1080/01421590600834229

Morton, J., Cumming, A., & Cameron, H. (2007). Performance-based assessment in undergraduate medical education. *The Clinical Teacher*, *4*(1), 36-41. http://dx.doi.org/10.1111/j.1743-498X.2007.00138.x

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

Norcini, J., Shea, J., & Kanya, T. (1988). The effect of various factors on standard setting. *Journal of Educational Measurement*, *25*(1), 57-65. http://dx.doi.org/10.2307/1435024

Patrício, M., Julião, M., Fareleira, F., Young, M., Norman, G., & Vaz Carneiro, A. (2009). A comprehensive checklist for reporting the use of osces. *Medical Teacher*, *31*(2), 112-124. http://dx.doi.org/10.1080/01421590802578277

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2010). How to measure the quality of the osce: A review of metrics—AMEE guide no. 49. *Medical Teacher*, *32*(10), 802-811. http://dx.doi.org/10.3109/0142159X.2010.507716

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2012). Is short-term remediation after osce failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in osce assessments. *Medical Teacher*, *34*(2), 146-150. http://dx.doi.org/10.3109/0142159X.2012.643262

Peterson, C., Schulz, E. M., & Engelhard, Jr. G. (2011). Reliability and validity of bookmark-based methods for standard setting: Comparisons to angoff-based methods in the national assessment of educational progress. *Educational Measurement: Issues and Practice*, *30*(2), 3-14. http://dx.doi.org/10.1111/j.1745-3992.2011.00200.x

Pett, M., Lackey, N., & Sullivan, J. (2003). *Making sense of factor analysis*. London: Sage Publications, Inc.

Ramsey, F. (1926). Truth and probability. In R. Braithwaite (Ed.), *Foundations of mathematics and other logical essays* (pp. 56-198). London: Kegan, Paul, Trench, Trubner & Co. Ltd.

Raykov, T., & Marcoulides, G. (2011). *Introduction to psychometric theory*. London: Routledge.

Rees, C. E., Knight, L. V., & Cleland, J. A. (2009). Medical educators' metaphoric talk about their assessment relationships with students: You don't want to sort of be the one who sticks the knife in them. *Assessment & Evaluation in Higher Education*, *34*(4), 455-467. http://dx.doi.org/10.1080/02602930802071098

Roberts, C., Newble, D., Jolly, B., Reed, M., & Hampton, K. (2006). Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher*, *28*(6), 535-543. http://dx.doi.org/doi:10.1080/01421590600711187

Sax, G., & Reade, M. (1964). Achievement as a function of test difficulty level. *American Educational Research Journal*, *1*(1), 22-25. http://dx.doi.org/10.3102/00028312001001022

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. http://dx.doi.org/10.1111/j.1600-0579.2008.00568.x

Schuwirth, L., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., . . . Wagner-Menghin, M. (2011). Research in assessment: Consensus statement and recommendations from the ottawa 2010 conference. *Medical Teacher*, *33*(3), 224-233. http://dx.doi.org/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. http://dx.doi.org/10.1002/9781444320282.ch14

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. http://dx.doi.org/10.1007/s10459-012-9367-y

Shumway, J., & Harden, R. (2003). AMEE guide no. 25: The assessment of learning outcomes for the competent and reflective physician. *Medical Teacher*, *25*(6), 569-584. http://dx.doi.org/10.1080/0142159032000151907

Skorupski, W., & Hambleton, R. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education*, *18*(3), 233-256. http://dx.doi.org/10.1207/s15324818ame1803_3

Smee, S. (2001). Setting standards for an objective structured clinical examination: The borderline group method gains ground on angoff. *Medical Education*, *35*, 1009-1010. http://dx.doi.org/10.1046/j.1365-2923.2001.01047.x

Torra, V., Domingo-Ferrer, J., Mateo-Sanz, J. M., & Ng, M. (2006). Regression for ordinal variables without underlying continuous variables. *Information Sciences*, *176*(4), 465-474. http://dx.doi.org/10.1016/j.ins.2005.07.007

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. http://dx.doi.org/10.1007/s10459-006-9035-1

Wang, N., Wiser, R., & Newman, L. (2001). *Use of the Rasch IRT model in standard setting: An item mapping method*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *THE LANCET*, *357*(9260), 945-949. http://dx.doi.org/10.1016/S0140-6736(00)04221-5

Wayne, D., Fudala, M., Butter, J., Siddall, V., Feinglass, J., Wade, L., & McGaghie, W. (2005). Comparison of two standard-setting methods for advanced cardiac life support training. *Academic Medicine*, *80*(10), S63-S66. http://dx.doi.org/10.1097/00001888-200510001-00018

Wilkinson, T., Frampton, C., Thompson-Fawcett, M., & Egan, T. (2003). Objectivity in objective structured clinical examinations: Checklists are no substitute for examiner commitment. *Academic Medicine*, *78*(2), 219-223. http://dx.doi.org/10.1097/00001888-200302000-00021

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. http://dx.doi.org/10.1111/j.1365-2923.2001.01041.x

Williams, R. G., Klamen, D. A., & McGaghie, W. C. (2003). Special article: Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching and Learning in Medicine*, *15*(4), 270-292. http://dx.doi.org/10.1207/S15328015TLM1504_11

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. http://dx.doi.org/10.1007/s10459-005-7853-1

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/3.0/).