


RESEARCH ARTICLE

Open Access



Borderline grades in high stakes clinical examinations: resolving examiner uncertainty

Boaz Shulruf^{*} , Barbara-Ann Adelstein[†], Arvin Damodaran[†], Peter Harris[†], Sean Kennedy[†], Anthony O'Sullivan[†] and Silas Taylor[†]

Abstract

Background: Objective Structured Clinical Exams are used to increase reliability and validity, yet they only achieve a modest level of reliability. This low reliability is due in part to examiner variance which is greater than the variance of students. This variance often represents indecisiveness at the cut score with apparent confusion over terms such as “borderline pass”. It is amplified by a well reported failure to fail.

Methods: A borderline grade (meaning performance is neither a clear pass nor a clear fail) was introduced in a high stakes undergraduate medical clinical skills exam to replace a borderline pass grade (which was historically resolved as 50%) in a 4 point scale (distinction, pass, borderline, fail). Each Borderline grade was then resolved into a Pass or Fail grade by a formula referencing the difficulty of the station and the performance in the same domain by the student in other stations. Raw pass or fail grades were unaltered. Mean scores and 95%CI were calculated per station and per domain for the unmodified and the modified scores/grades (results are presented on error bars). To estimate the defensibility of these modifications, similar analysis took place for the P and the F grades which resulted from the modification of the B grades.

Results: Of 14,634 observations 4.69% were Borderline. Application of the formula did not impact the mean scores in each domain but the failure rate for the exam increased from 0.7 to 4.1%. Examiners and students expressed satisfaction with the Borderline grade, resolution formula and outcomes. Mean scores (by stations and by domains respectively) of students whose B grades were modified to P were significantly higher than their counterparts whose B grades were modified to F.

Conclusions: This study provides a feasible and defensible resolution to situations where the examinee's performance is neither a clear pass nor a clear fail, demonstrating the application of the resolution of borderline formula in a high stakes exam. It does not create a new performance standard but utilises real data to make judgements about these small number of candidates. This is perceived as a fair approach to Pass/Fail decisions.

Background

Ensuring competence in clinical skills is central to medical and health professions education. The Objective Structured Clinical Examination (OSCE) was introduced to increase the reliability and validity of clinical skills assessment in medical education [1]. Since being introduced, the OSCE has become established as one of the leading assessment tools in medical schools and across

many health professions education programs [2–7]. Previous studies provide extensive evidence supporting the reliability and validity of the OSCE [3, 8–14]. To improve reliability examiners are commonly provided with a predetermined checklist to use when marking an examinee's performance using categories from Fail to Distinction and at times numeric marks are attached to these categories [15]. Among the possible categories is a Borderline grade which describes a level of performance that is neither clear pass nor clear fail [14, 16, 17].

Definitions of the Borderline grade vary [18–20], for example ‘Students who possess just enough knowledge

* Correspondence: bshulruf@unsw.edu.au

[†]Barbara-Ann Adelstein, Arvin Damodaran, Peter Harris, Sean Kennedy, Anthony O'Sullivan and Silas Taylor contributed equally to this work. Faculty of Medicine, University of New South Wales, Sydney, Australia



to reject F-responses' [20] or 'a minimally competent graduate is one who meets the standard by the smallest possible margin' [21]. That perhaps explains why Cizek described a borderline proficiency as 'an abstraction in terms of performance' [22]. The intangible nature of the borderline grade poses a challenge to assessors, since different individuals may have different understandings regarding what constitutes a borderline grade. By definition, clear pass and clear fail grade decisions mean that there is no ambiguity and therefore examiners are more likely to agree with each other. Nonetheless, even in these circumstances, the literature suggests that the impact of examiners on the assessment outcome is critical. For example, examiners reported that they felt less confident when awarding a fail grade than when giving a pass grade [23]. Biases, such as gender and culture, have been found to have affected examiners' judgements [24], as has examiners' familiarity with the examinees [25]. A recent study on an OSCE used in an Exercise Physiology program found that the examiners accounted for 24.1% of the variance in technical skills scores, whereas students accounted only for 4.9% of the variance [26]. A comprehensive meta-analysis estimated that OSCEs achieve an overall low reliability (<.60) and suggested that an OSCE '*does not guarantee reliable scores and accurate decisions about medical students*' [27]. This substantial evidence suggests that examiners' biases are unavoidable when OSCEs are employed. Although not explicit in the literature, the biases discussed are more likely to affect students performing at the borderline level than when their performance is a clear pass or clear fail.

To date, the literature addresses the borderline grade issue by suggesting methods for defining cut-scores for the entire OSCE examination or for specific stations. Among these are the Angoff method the Borderline Groups Method, the Borderline Regression Method and the Contrasting Groups Method [28–37].

The AMEE guide no. 49 'How to measure the quality of the OSCE: A review of metrics' favours the Borderline Regression Method (BRM) [38, 39] since it "*uses all the assessment interactions between assessors and candidates, and these interactions are real*" [2] and is "*objectively based on predetermined criteria, using a large number of assessors and generates a wide range of metrics*" [2].

Guided by the principles suggested by G Pell, R Fuller, M Homer and T Roberts [2], a new method (The Objective Borderline Method, henceforth: OBM) for addressing challenges raised by borderline grades was introduced [40–43]. The OBM utilises the institutional predetermined criteria for clear pass and clear fail and uses all assessment interactions between assessors and candidates [2] to determine whether a borderline grade should be reclassified as pass or fail, which is argued to improve the reliability of the OSCE. This study reports

the application of the OBM principles, when the OBM2 [42] was applied to a high stakes OSCE undertaken at the end of the second year of the Medicine program at UNSW Medicine, UNSW Sydney, Australia.

Context

The UNSW Medicine program is a six-year undergraduate entry program awarding two degrees, namely Bachelor of Medical Sciences (BMed) and Doctor of Medicine (MD) [44]. This modular program consists of three phases, each of two years. Students undertake barrier examinations at the end of each phase.

Year 2 students sit a clinical examination at the end of Phase 1. This focuses on three domains: generic communication skills; clinical communication skills (medical history-taking); and physical examination skills [45]. Previous to the borderline method being introduced (OBM2, described in detail in the next section), examiners were able to offer one of four categorical grades: Fail (F); Borderline Pass (P-); Clear Pass (P); and Exceeded Expectations/ Distinction (P+), and numeric scores were generated from these grades as follows: (F) = 3; (P-) = 5; (P) = 7; (P+) = 9 (out of 10). When student performance was uniformly outstanding, that is they received only P+ grades, examiners were at liberty to upgrade the numeric score from 9 to 10 for details see: [46].

The motivation for seeking an improved method for making defensible Pass/Fail decisions for the borderline grades arose from concerns expressed by course and program leaders that clinical examiners were overly lenient and failed to fail poorly performing students, rather tending to award P- rather than F grades. It was noted that examiners' free text comments often indicated unsatisfactory performances, while the grade awarded was a P- (borderline pass), observations which echoed concerns reported in the literature [47, 48]. In addition, concerns were expressed regarding the nature and meaning of the P- grade. The P- was described as a 'Borderline Pass', yet some examiners perceived it as a 'pass with conditions'. Under this previous marking method (henceforth, 'traditional method') accumulated criteria results of two P-'s and no fails was graded as an overall pass, but a student with three or more P-'s failed the station. This had a logical flaw since the P- (or Borderline Pass grade) was neither numerically nor descriptively defined as a Fail. Finally, the Phase 1 OSCE generally had a failure rate of less than 1%, significantly lower than other barrier examinations in the Medicine program.

The introduction of the OBM2 [42, 43] to the literature provided an opportunity for the program directors to improve the pass/fail decision making for the OSCE's in the Medicine program and it was conducted across all clinical examinations in the UNSW Medicine program in 2016. Implementation of OBM2 focused on the

abovementioned concerns. The Borderline Pass (P-) grade was replaced with a 'Borderline' grade, which indicates (as per the examiner instructions) that the examiner is unable to conclude whether the student performance for a particular assessment criterion (item) was a clear pass or a clear fail. Using of the Borderline grade permitted examiners to award an undetermined pass/fail assessment where appropriate, and when insufficient evidence was available to make a clear decision, prevented examiners being forced to do so. The use of the undetermined 'Borderline' grade (B) thus aimed to achieve two distinct goals: mitigate examiner's marking bias [25, 47], and reducing examiner anxiety in difficult cases, which at times might encourage examiners to give students the "benefit of the doubt" and award them with an unjustifiable 'pass'.

Implementation of the OBM2 included a revision of assessment guides and alterations to examiner training, which occurred throughout the 2016 academic year and the OSCE occurred in November 2016. Four relevant Faculty governance committees, each including student representatives, independently discussed and approved the implementation of the OBM2. A contingency plan was prepared should the implementation of the OBM2 fail. OBM2 implementation replaced all the examiner awarded B grades with either an F or P as determined by the OBM2 algorithm.

The current study focuses on the implementation of OBM2 [42] in this clinical skills examination (OSCE) undertaken at the end of Phase 1. The next section (*'The Objective Borderline Method'*) describes the OBM2 in detail.

The objective borderline methods (OBM and OBM2)

The OBM

The OBM [40] is a method that yields an index from two independent proportions of responses to examination items when the possible responses are: clear pass and above (P); clear fail (F); and borderline (B), an undetermined grade. The two relevant proportions are: (1) the proportion of P grades among all the non-F grades; and (2) the proportion of B grades among all the non-P grades.

If the number of P grades is p ; the number of F grades is f ; and the number of B grades is b then:

The proportion of the B grades among all the non-P grades is: $\text{Pr}(B) = b/(f + b)$.

The proportion of the P grades among all the non-F grades is: $\text{Pr}(P) = p/(b + p)$.

The OBM index is the multiplication of these: $\text{OBM index} = \text{Pr}(P) \times \text{Pr}(B) = [p/(b + p)] \times [b/(f + b)]$.

Thus the OBM index presents: the difficulty of not getting an F grade (i.e. getting a B grade) given a P mark is not achievable; and the difficulty of getting a P grade given all grades are above clear fail ($>F$ grade). Multiplication of proportions is an acceptable practice for yielding indices

derived from observations. [49]. Importantly, although $\text{Pr}(P)$ and $\text{Pr}(B)$ are related, they remain independent since a particular *proportion* of P grades among the P and the B grades *cannot* determine the *proportion* of the B grades among the B and the F grades (and vice versa). The OBM does not apply when there are no B grades, since no decision is required. The OBM applies when examination marks are on a continuous scale yet uncertainty exists regarding the cut-score separating pass from fail. In order to apply the OBM there is a need to determine the minimum score for clear pass and the maximum score for clear fail, whereas the scores that are neither clear pass nor clear fail are defined as borderline. Since the OBM is a multiplication of two proportions each of which is a sub-group within a group (i.e. $\text{Pr}(B) = b/(f + b)$; $\text{Pr}(P) = p/(b + p)$), the OBM index is always ≤ 1 . Upon introduction the OBM was used to determine the proportion of borderline grades that should be re-classified to Pass; and 1-OBM index determined the proportion of borderline grades that should be reclassified to Fail [40]. Using this classification a cut-score could be estimated (the lowest borderline grade that was reclassified to Pass). B Shulruf, R Turner, P Poole and T Wilkinson [40] demonstrated that the cut-scores generated by the OBM were highly correlated when compared with cut-scores generated by other methods. However, known standard setting methods, including the OBM, still left the question of determining individual student borderline results unresolved.

The OBM2

Logically, for each assessment criterion one must either pass or fail but never both, since not passing means failing, and not failing means passing. Thus a borderline mark or grade (used here interchangeably) means that the examiner could not make a clear decision, most probably due to insufficient information provided to them during the examination. The OBM2 was thus introduced as a *pass/fail decision making* process to reclassify the indecisive borderline grades to the most likely decisive grade, either pass or fail. The OBM2 is not a tool for generating a cut-score but is a tool for making defensible decisions when there is uncertainty.

The OBM generates two indices: the 'Difficulty' of an item, when all responses to a single item given by all students are considered, and an index of student 'Ability', when considering all responses to all items within a construct given by a single student. The OBM2 uses these two OBM indices to make pass/fail decisions for B grades and works by these two OBM indices being calculated for each B grade. If $\text{Ability} < \text{Difficulty}$ the B grade is reclassified as an F grade, otherwise if $\text{Ability} \geq \text{Difficulty}$ then the B grade is reclassified as P grade.

It is important to understand that although inspired by Item Response Theory (IRT), the OBM2 is by no means

a form of IRT, nor is it an alternative to IRT models. The OBM2 is used only in relation to a particular type of examination consisting of three types of grades (Fail, Borderline, Pass and above) and its only purpose is facilitating the pass/fail decisions when borderline grades are awarded. Similarity to IRT is evident in that item difficulty and student ability are measured on the same scale, and this is where the OBM2 and IRT are most comparable. The OBM2 is applicable only when items underlie a single construct, in the case of our study, each of the three domains: generic communication skills; clinical communication skills (medical history-taking); and physical examination skills [45]. Previous studies demonstrated that, to yield a high level of accuracy, items need to be loaded on a single factor and yield at least a moderately acceptable level of reliability (Cronbach's $\alpha > .60$). [42, 50].

Table 1 demonstrates how the OBM2 is applied. This example is taken from responses to assessment criteria related to physical examination in an OSCE conducted at one of UNSW's clinical examination sites. The OSCE consists of six stations, in which each had 12 assessment criteria related to the three domains [45] and there are 24 examinees. For each item one can be awarded an F, B, P or P+ grade which were converted to a numeric score F (=3), B (=5), P (=7) or P+ (=9) for analysis. This produces a "raw" score. OBM indices (Ability and Difficulty) were calculated for each item and each student when applicable (if no B grade was obtained, no OBM index was calculated). Then for each B grade a comparison between Ability and Difficulty was made as described above. The arrows on the right hand-side of the 5's (=borderline) indicate whether the 5 is modified to 7 (\uparrow) or to 3 (\downarrow). The grades in this demonstration yield good internal consistency (Cronbach's $\alpha = .749$). Readers may scrutinise the table to see how the OBM2 works across students and items. This table is readily constructible using Excel™ and readers may test it using their own data.

A simulation study using one of the OBM versions [41] demonstrated that on average the accuracy of the pass/fail decisions made by the OBM was about 70% which is equivalent to effect size = 1 [51]. A more recent study [43], which used real data and applied the OBM2 (the OBM version presented here), yielded 77% accuracy which is equivalent to effect size = 1.4 [51].

This study sought to identify the impact of the implementation of the OBM2 on examination results in a high stakes OSCE early in the Medicine Program, and further, to assess the validity and defensibility of the application of OBM2 to a high stakes OSCE in a Medicine program.

The study was approved by the UNSW Human Research Ethics Advisory (HREA) Panel ref.: HC15421.

Methods

Data

This study used data of Phase 1 final OSCE examination (end of Year 2) in the medicine program ($N = 271$).

All original grades generated in the OSCE were modified to the respective 'raw' scores (F = 3; B = 5, P = 7, P+ = 9 or 10 if all grades in the station were P+).

The OBM2 was applied to the 'raw' scores such that the B grades (score = 5) were modified to either F(=3) or P(=7).

Statistical analysis

Descriptive statistics were employed to report the distribution of the grades before and after the reclassification. Mean scores and 95%CI were calculated per station and per domain for the unmodified and the modified scores/grades (results are presented on error bars). Similar analysis took place for the P and the F grades which resulted from the modification of the B grades (results are presented on error bars). For comparison of success rate, the OSCE grades were calculated twice. First time using the 'traditional method' [45] and then using the OBM2.

Anecdotal feedback from students, examiners and other staff engaged in the OSCE's was received through committees' discussions. Quotes could not be provided since these were not covered by the ethics approval but summaries of feedback were added to the discussion as complementary contextual information.

Results

The OSCE yielded 14,634 grades of which 687 (4.69%) were Borderline (B grade) (Table 2). The application of the OBM2 reclassified 355 (51.7%) of the Borderline grades to Fail and 332 (48.3%) to Pass.

Passing the OSCE required achieving a mean score ≥ 5 in all three domains and all six stations. The implementation of the OBM2 increased the number of students who failed the OSCE from 2 (0.7%) using the 'traditional method' to 11 (4.1%). Figures 1 and 2 demonstrate that the reclassification did not have any significant impact on the mean scores across domains and separately across stations. Nonetheless, Figs. 3 and 4 demonstrate that the mean scores (by stations and by domains respectively) of students whose B grades were modified to P were significantly higher than their counterparts whose B grades were modified to F.

Discussion

The results of this study clearly demonstrate that the application of the OBM2 to the Phase 1 OSCE at UNSW did impact student outcomes, but in a desirable direction. The most important finding is that the implementation of the OBM2 increased the failure rate from 2 (0.7%) to 11 (4.1%). This major change, more than fivefold, was more

Table 1 A demonstration of how the OBM2 is calculated

Items Student No.	1	2	3	4	5	6	7	8	9	10	11	12	F	B	P	Ability: $b/(b + f)*p/(p + b)$
1	5↓	5↑	7	7	9	9	9	7	7	7	7	7	0	2	10	0.833
2	9	9	7	7	9	9	7	9	9	9	9	9	0	0	12	
3	7	7	7	7	7	7	7	3	5↓	5↓	7	7	1	2	9	0.545
4	5↓	5↓	7	7	7	7	7	3	7	7	3	7	2	2	8	0.400
5	7	7	7	7	7	7	7	7	9	9	7	7	0	0	12	
6	9	5↑	7	7	7	9	7	7	7	7	7	7	0	1	11	0.917
7	7	9	7	7	9	9	7	7	9	9	7	7	0	0	12	
8	7	7	7	7	7	7	7	7	7	9	7	7	0	0	12	
9	7	7	7	7	7	7	7	9	9	9	7	7	0	0	12	
10	7	7	7	5↓	5↓	7	7	7	7	7	5↑	7	0	3	9	0.750
11	7	7	7	7	7	7	7	7	5↓	5↓	5↑	7	0	3	9	0.750
12	7	9	7	7	9	7	7	7	9	7	7	7	0	0	12	
13	7	7	5↓	5↓	7	9	7	7	7	7	7	7	0	2	10	0.833
14	7	7	7	7	7	7	7	7	7	7	7	9	0	0	12	
15	7	9	7	5↓	7	7	7	3	5↓	7	7	7	1	2	9	0.545
16	7	5↑	7	7	9	9	7	7	7	7	7	7	0	1	11	0.917
17	7	7	9	7	7	7	7	7	7	7	7	7	0	0	12	
18	7	7	7	7	7	7	7	7	5↑	5↓	7	7	0	2	10	0.833
19	7	7	9	7	7	7	9	7	7	7	7	7	0	0	12	
20	7	9	7	7	7	9	7	7	7	7	9	9	0	0	12	
21	7	7	7	7	7	7	7	7	7	7	7	7	0	0	12	
22	7	7	9	9	7	7	7	9	7	7	7	7	0	0	12	
23	7	7	9	9	7	7	7	7	9	9	7	9	0	0	12	
24	9	9	7	7	7	9	7	7	7	7	7	7	0	0	12	
F	0	0	0	0	0	0	0	3	0	0	1	0				
B	2	4	1	3	1	0	0	0	4	3	2	0				
P	22	20	23	21	23	24	24	21	20	21	21	24				
Difficulty: $b/(b + f)*p/(p + b)$	0.917	0.833	0.958	0.875	0.958			0.833	0.875	0.609						

Arrows indicate if the mark is to be modified up or down based on the calculation of Ability and Disability score

* meand multiplication

Table 2 Grade distribution before and after the implementation of the OBM2

Grade	Score	Prior re-classification		Post re-classification	
		N	%	N	%
F	3	83	0.57	438	3.00
B	5	687	4.69		
P	7	10,338	70.64	10,670	72.91
P+	9	3211	21.94	3211	21.94
P++*	10	315	2.15	315	2.15
Total		14,634	100.00	14,634	100.00

P++ when all grades in a station are P+ Score = 10 (as defined by the program assessment guideline)

reflective of failure rates in other barrier examinations in the Medicine program. Nonetheless, in comparison to the literature a failure rate of 4.1% is within the *lower range* previously reported in clinical examinations [29, 52, 53]. Consequently this outcome (the increase in the failure rate) is perceived as a positive outcome as it increases the confidence that fewer incompetent students passed the OSCE [23, 54, 55] without any expression of dissatisfaction from any stakeholders, particularly students. Moreover, the high level of satisfaction expressed by students, examiners and program leaders throughout the implementation of the OBM2 adds credibility and support to the acceptability of the OBM2 as a method which properly resolves the indecisive borderline grades.

The findings from this study demonstrate that the reclassification made by the OBM2 was appropriate. Figures 3 and

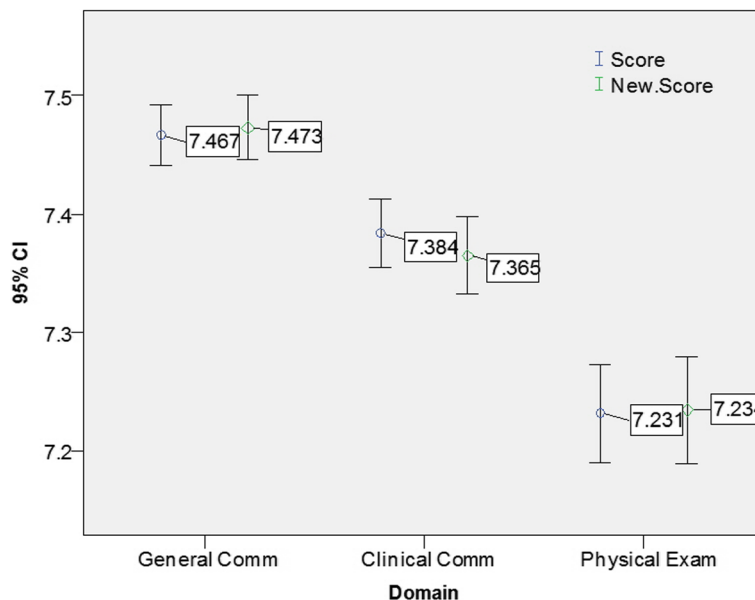


Fig. 1 Mean score by domains prior and post reclassification

4 demonstrate that B grades that were reclassified to P were associated with higher overall performance whereas reclassification of B to F grades was associated with a lower level of overall performance either across domains within each station or across stations within each domain.

The OBM2 provides a feasible and defensible solution to a relatively overlooked problem: *how to properly assess an examinee's performance which could not be clearly identified as either pass or fail*. All known standard setting methods aim to establish a cut-score which

determines whether the examinees passed or failed the examination (or crossed the boundary for other classifications e.g. pass-distinction). These methods assume that a borderline grade describes a particular level of performance, either just pass (borderline-pass) or just fail (borderline-fail) or in the middle between pass and fail i.e. borderline [38, 56, 57]. In other words the assumption is that a borderline grade describes a discrete level of performance. Logically, this assumption is problematic. One either meets performance criteria or not; it

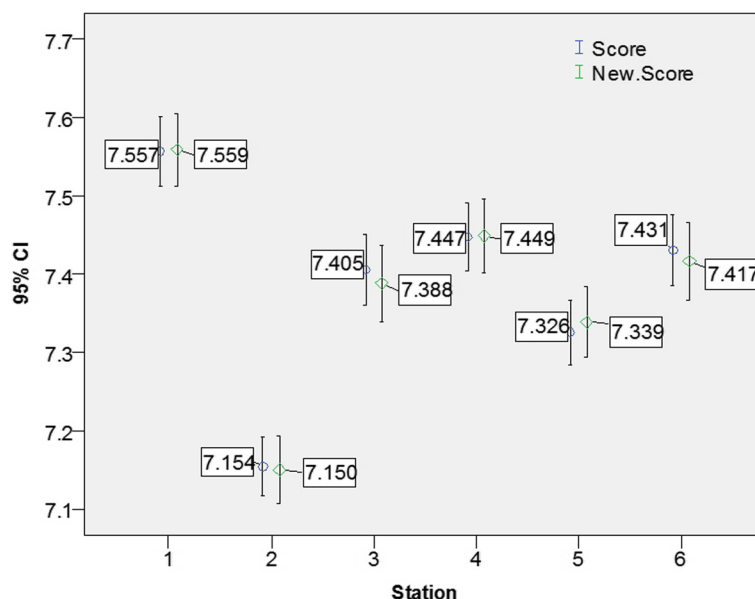


Fig. 2 Mean score by station prior and post reclassification

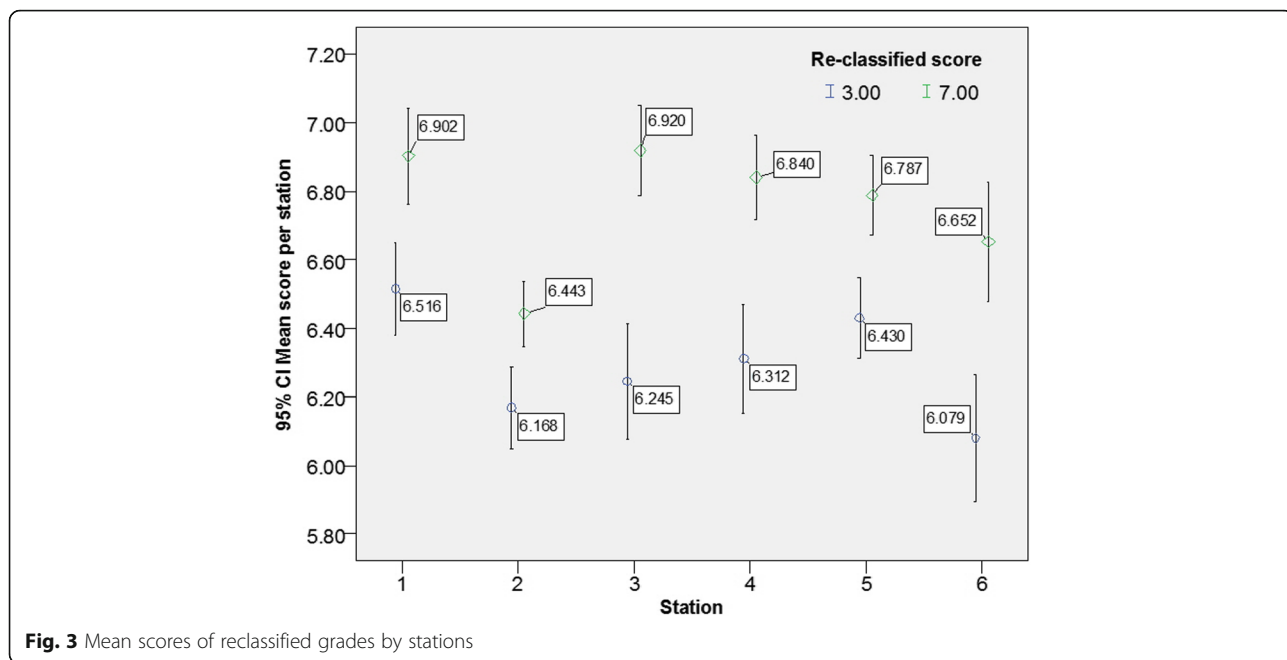


Fig. 3 Mean scores of reclassified grades by stations

is impossible to both meet and not meet the performance criteria at the same time. Similarly, it is impossible to neither meet nor fail to meet the performance criteria at the same time. Thus, borderline-pass means meeting the performance criteria yet at a lower level and borderline fail means failing to meet the performance criteria but only just. A borderline grade could therefore indicate that there is insufficient information to determine whether the examinee met or did not meet the performance criteria. OBM2 provides a simple yet defensible

method for making this determination. When the information is insufficient to determine the examinee’s level of performance, the OBM2 utilises all the ‘real’ assessment interactions between assessors and examinees, to objectively (i.e. without any additional judgement, based on predetermined criteria and using a large number of examiners) determine the most likely level of performance, i.e. pass or fail [2].

An important feature of the OBM2 is that it does not create new performance standards nor does it set any

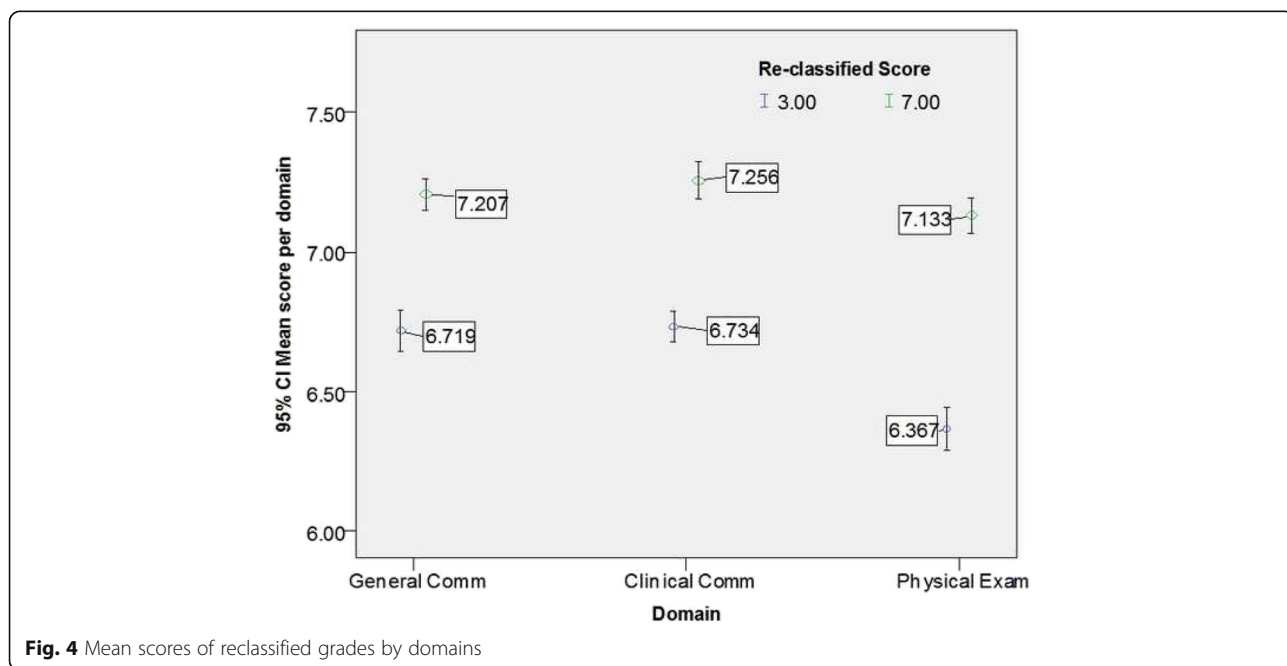


Fig. 4 Mean scores of reclassified grades by domains

cut-scores. The OBM2 utilises the very same predetermined performance criteria as advised to students, teachers and examiners prior to the examination taking place. This adds fairness since no examinee could be negatively impacted by the OBM2 once the performance criteria have been clearly met; nor would one unjustifiably be granted a pass once the performance criteria have clearly not been met. The OBM2 applies only when the examiner cannot clearly determine the level of the examinee's observed performance.

An additional advantage of the OBM2 is that it considers the difficulty of each item (i.e. performance criterion). That means that when the OBM2 is applied, the pass/fail decision is more stringent when it is made for an easy item and more lenient when made for a difficult item. This provides an additional layer of fairness to the pass/fail determinations made by the OBM2, and this was noted in student feedback throughout the implementation of the OBM2.

In earlier studies, the OBM and OBM2 were successfully applied to either simulated data or historical data (that is, data that had been generated with no prior intention to apply such methodology) [40–43]. A recent study using OSCE data generated within the OBM framework (borderline = indecisive mark), demonstrated high efficacy of the OBM2 when applied to a common OSCE setting where the pass/fail decisions were made based on assessment marks within a station [50]. The current study demonstrates that the OBM2 is applicable even when pass/fail decisions need to be made based on assessment marks across stations.

Nonetheless, despite the consistent positive outcome yielded from the 'live' implementation of the OBM2 in the UNSW Medicine program, there remains a need for further research to better clarify the OBM and OBM2 limitations, particularly but not limited to assessment data that include very small number of students.

Conclusions

In conclusion, this study demonstrated that the OBM2 is a simple, feasible and defensible method to reclassify indecisive grade (i.e. borderline) to either pass or fail. The OBM2 was found acceptable by all stakeholders and is now fully implemented in a well-established undergraduate Medicine program in Australia. Future studies may provide better insight into the OBM2 including its limitations and advantages.

Abbreviations

OBM: Objective Borderline Method; OSCE: Objective Structured Clinical Examination

Acknowledgements

Not applicable.

Funding

No funding was obtained for this study.

Availability of data and materials

Anonymised data are stored on UNSW secure server. Data are available by request conditional to the approval of UNSW ethics committee.

Authors' contributions

BS Generated the concept, designed the study, undertook the statistical analysis and contributed to the interpretation of the results and the writing. BA contributed to data collection, the interpretation of the data and to the Writing, AD contributed to data collection, the interpretation of the data and to the Writing, PH contributed to data collection, the interpretation of the data and to the Writing, SK contributed to data collection, the interpretation of the data and to the Writing, AO contributed to data collection, the interpretation of the data and to the Writing, ST contributed to data collection, the interpretation of the data and to the Writing. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The study was approved by Human Research Ethics Advisory (HREA) Panel G: Health, Medical, Community and Social ref. # HC15421.

Consent for publication

The Ethics approval does not require acquisition of participants' consent. The data were generated via a regular educational assessment activity, irrespective to the study. Thus, there are no 'participants' in the study since the study used administrative data generated from student assessment records. These data, in their anonymised form, are approved by our Ethics Committee to be used for this study.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 January 2018 Accepted: 8 November 2018

Published online: 20 November 2018

References

- Harden R, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ*. 1979;13(1):39–54.
- Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: a review of metrics – AMEE guide no. 49. *Medical Teacher*. 2010; 32(10):802–11.
- Dong T, Swygert K, Durning S, Saguil A, Gilliland W, Cruess D, DeZee K, LaRochelle J, Artino A. Validity evidence for medical school OSCEs: associations with USMLE® step assessments. *Teaching and Learning in Medicine*. 2014;26(4):379–86.
- Redfern S, Norman I, Calman L, Watson R, Murrells T. Assessing competence to practise in nursing: a review of the literature. *Res Pap Educ*. 2001;17(1):51–77.
- Selim A, Ramadan F, El-Gueneidy M, Gaafer M. Using objective structured clinical examination (OSCE) in undergraduate psychiatric nursing education: is it reliable and valid? *Nurse Educ Today*. 2012;32(3):283–8.
- Davis M, Ponnampereuma G, McAleer S, Dale V. The objective structured clinical examination (OSCE) as a determinant of veterinary clinical skills. *Journal of veterinary medical education*. 2006;33(4):578–87.
- Simmons B, Egan-Lee E, Wagner S, Esdaile M, Baker LC, Reeves S. Assessment of interprofessional learning: the design of an interprofessional objective structured clinical examination (iOSCE) approach. *Journal of Interprofessional Care*. 2011;25(1):73–4.
- Vallevand A, Violato C. A predictive and construct validity study of a high-stakes objective clinical examination for assessing the clinical competence of international medical graduates. *Teaching and Learning in Medicine*. 2012;24(2):168–76.
- Najjar R, Docherty A, Miehle N. Psychometric properties of an objective structured clinical assessment tool. *Clinical Simulation in Nursing*. 2016;12(3):88–95.
- Eberhard L, Hassel A, Bäumer A, Becker F, Beck-MuBotter J, Börmicke W, Corcodel N, Cosgarea R, Eiffler C, Giannakopoulos NN, et al. Analysis of quality and feasibility of an objective structured clinical examination (OSCE) in preclinical dental education. *Eur J Dent Educ*. 2011;15(3):172–8.

11. Artemiou E, Hecker K, Adams C, Coe JB. Does a rater's professional background influence communication skills assessment? *Journal of Veterinary Medical Education*. 2015;42(4):315–23.
12. Sandilands D, Gotzmann A, Roy M, Zumbo BD, De Champlain A. Weighting checklist items and station components on a large-scale OSCE: is it worth the effort. *Medical Teacher*. 2014;36(7):585–90.
13. van Vught A, Hettinga A, Denessen E, Gerhardus M, Bouwmans G, van den Brink G, Postma C. Analysis of the level of general clinical skills of physician assistant students using an objective structured clinical examination. *J Eval Clin Pract*. 2015;21(5):971–5.
14. Yousuf N, Violato C, Zuberi R. Standard setting methods for pass/fail decisions on high-stakes objective structured clinical examinations: a validity study. *Teaching and Learning in Medicine*. 2015;27(3):280–91.
15. Khan K, Ramachandran S, Gaunt K, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part I: an historical and theoretical perspective. *Medical Teacher*. 2013;35(9):e1437–46.
16. Pell G, Fuller R, Homer M, Roberts T. Is short-term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments. *Medical Teacher*. 2012; 34(2):146–50.
17. Cizek G, Bunch M. The Angoff method and Angoff variations. In: Cizek G, Bunch M, editors. *Standard Setting*. edn. Thousand Oaks, California: SAGE Publications, Inc; 2007. p. 81–96.
18. Cizek G, Bunch M. The contrasting groups and borderline group methods. In: Cizek G, Bunch M, editors. *Standard Setting*. edn. Thousand Oaks, California: SAGE Publications, Inc; 2007. p. 105–17.
19. Fuller R, Homer M, Pell G, Hallam J. Managing extremes of assessor judgment within the OSCE. *Medical Teacher*. 2016:1–9.
20. Nedelsky L. Absolute grading standards for objective tests. *Educ Psychol Meas*. 1954;14(1):3–19.
21. Burr S, Zahra D, Cookson J, Salih V, Gabe-Thomas E, Robinson I. Angoff anchor statements: setting a flawed gold standard? *MedEdPublish*. 2017;9(3):53.
22. Cizek G. An NCME instructional module on: setting passing scores. *Educ Meas Issues Pract*. 1996;15(2):20–31.
23. Tweed M, Thompson-Fawcett M, Wilkinson T. Decision-making bias in assessment: the effect of aggregating objective information and anecdote. *Medical Teacher*. 2013;35(10):832–7.
24. Woolf K, Haq I, McManus C, Higham J, Dacre J. Exploring the underperformance of male and minority ethnic medical students in first year clinical examinations. *Adv Health Sci Educ*. 2008;13(5):607–16.
25. Stroud L, Herold J, Tomlinson G, Cavalcanti R. Who you know or what you know? Effect of examiner familiarity with residents on OSCE scores. *Acad Med*. 2011;86(10 Suppl):S8–11.
26. Naumann F, Marshall S, Shulruf B, Jones P. Exploring examiner judgement of professional competence in rater based assessment. *Adv Health Sci Educ*. 2016;21(4):775–88.
27. Brannick M, Erol-Korkmaz T, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ*. 2011;45(12): 1181–9.
28. Rajiah K, Veettil S, Kumar S. Standard setting in OSCEs: a borderline approach. *Clin Teach*. 2014;11(7):551–6.
29. Wood T, Humphrey-Murto S, Norman G. Standard setting in a small scale OSCE: a comparison of the modified borderline-group method and the borderline regression method. *Adv Health Sci Educ*. 2006;11(2):115–22.
30. Roberts C, Newble D, Jolly B, Reed M, Hampton K. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher*. 2006;28(6):535–43.
31. Boursicot K, Roberts T, Pell G. Standard setting for clinical competence at graduation from medical school: a comparison of passing scores across five medical schools. *Adv Health Sci Educ*. 2006;11(2):173–83.
32. Kilminster S, Roberts T. Standard setting for OSCEs: trial of borderline approach. *Adv Health Sci Educ*. 2004;9(3):201–9.
33. Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, Van Der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. *Med Educ*. 2003;37(2):132–9.
34. Boulet J, De Champlain A, McKinley D. Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical teacher*. 2003;25(3):245–9.
35. Wilkinson T, Newble D, Frampton C. Standard setting in an objective structured clinical examination: use of global ratings of borderline performance to determine the passing score. *Med Educ*. 2001;35:1043–9.
36. Khan K, Gaunt K, Ramachandran S, Pushkar P. The objective structured clinical examination (OSCE): AMEE guide no. 81. Part II: Organisation & Administration. *Medical Teacher*. 2013;35(9):e1447–63.
37. Schoonheim-Klein M, Muijtjens A, Habets L, Manogue M, van der Vleuten C, van der Velden U. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *Eur J Dent Educ*. 2009;13(3):162–71.
38. Hejri S, Jalili M, Muijtjens A, Van der vleuten C. assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *Journal of Research in Medical Sciences*. 2013;18(10):887–91.
39. Boursicot K, Roberts T, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Med Educ*. 2007;41(11):1024–31.
40. Shulruf B, Turner R, Poole P, Wilkinson T. 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. *Adv Health Sci Educ*. 2013;18(2):231–44.
41. Shulruf B, Poole P, Jones P, Wilkinson T. The objective borderline method (OBM): a probabilistic method for standard setting. *Assess Eval High Educ*. 2014.
42. Shulruf B, Jones P, Turner R. Using student ability and item difficulty for standard setting. *Higher Education Studies*. 2015;5(4):106–18.
43. Shulruf B, Booth R, Baker H, Bagg W, Barrow M. Using the objective borderline method (OBM) to support Board of Examiners' decisions in a medical programme. *J Furth High Educ*. 2017;41(3):425–34.
44. UNSW Handbook 2017 <http://www.handbook.unsw.edu.au/2017/index.html>.
45. McNeil P, Hughes C, Toohey S, Dowton S. An innovative outcomes-based medical education program built on adult learning principles. *Medical Teacher*. 2006;28(6):527–34.
46. O'Sullivan A, Harris P, Hughes C, Toohey S, Balasooriya C, Velan G, Kumar R, McNeil P. Linking assessment to undergraduate student capabilities through portfolio examination. *Assess Eval High Educ*. 2012;37(3):379–91.
47. Hope D, Cameron H. Examiners are most lenient at the start of a two-day OSCE. *Medical Teacher*. 2015;37(1):81–5.
48. McManus I, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*. 2006; 6(1):42.
49. Shoukri M. *Measures of Interobserver agreement and reliability*. 2nd ed. London: CRC Press; 2010.
50. Shulruf B, Damodaran A, Jones P, Kennedy S, Mangos G, O'Sullivan A, Rhee J, Taylor S, Velan G, Harris P. Enhancing the defensibility of examiners' marks in high stake OSCEs. *BMC Medical Education*. 2018;18(10):1–9.
51. Coe R. It's the effect size, stupid: what effect size is and why it is important. In: *British Educational Research Association annual conference*. UK: Exeter. p. 2002.
52. Mortaz Hejri S, Yazdani K, Labaf A, Norcini J, Jalili M. Introducing a model for optimal design of sequential objective structured clinical examinations. *Adv Health Sci Educ*. 2016:1–14.
53. Lillis S, Stuart M, Takai N. New Zealand registration examination (NZREX clinical): 6 years of experience as an objective structured clinical examination (OSCE). *The New Zealand Medical Journal*. 2012;125(1361):74–80.
54. Rushforth H. Objective structured clinical examination (OSCE): review of literature and implications for nursing education. *Nurse Educ Today*. 2007; 27(5):481–90.
55. Harasym P, Woloschuk W, Cuning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ*. 2008;13(5):617–32.
56. Pell G, Roberts T. Setting standards for student assessment. *International Journal of Research & Method in Education*. 2006;29(1):91–103.
57. Cizek G, Bunch M. *Standard setting: a guide to establishing and evaluating performance standards on tests*. London: Sage Pubns; 2007.