Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination

Sara Mortaz Hejri[1], Mohammad Jalili[2], Arno M. M. Muijtjens[3], Cees P. M. Van Der Vleuten[4]

[1]Department of Medical Education; School of Medicine, Tehran University of Medical Sciences, Tehran, Iran, [2]Department of Emergency Medicine, Department of Medical Education; School of Medicine, Tehran University of Medical Sciences, Tehran, Iran, [3]Department of Medical Education and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University; Maastricht, The Netherlands, [4]Department of Educational Development and Research, School of Health Professions Education, Maastricht University; Maastricht, The Netherlands

Background: One of the methods used for standard setting is the borderline regression method (BRM). This study aims to assess the reliability of BRM when the pass-fail standard in an objective structured clinical examination (OSCE) was calculated by averaging the BRM standards obtained for each station separately. **Materials and Methods:** In nine stations of the OSCE with direct observation the examiners gave each student a checklist score and a global score. Using a linear regression model for each station, we calculated the checklist score cut-off on the regression equation for the global scale cut-off set at 2. The OSCE pass-fail standard was defined as the average of all station's standard. To determine the reliability, the root mean square error (RMSE) was calculated. The R^2 coefficient and the inter-grade discrimination were calculated to assess the quality of OSCE. **Results:** The mean total test score was 60.78. The OSCE pass-fail standard and its RMSE were 47.37 and 0.55, respectively. The R^2 coefficients ranged from 0.44 to 0.79. The inter-grade discrimination score varied greatly among stations. **Conclusion:** The RMSE of the standard was very small indicating that BRM is a reliable method of setting standard for OSCE, which has the advantage of providing data for quality assurance.

Key words: Borderline regression method, objective structured clinical examination, reliability, standard setting

How to cite this article: Mortaz Hejri S., Jalili M, Muijtjens AMM, Van der Vleuten CPM. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci 2013;18:887-91.

INTRODUCTION

The pass-fail standard is a cut-score on a test that indicates the minimal adequate level of competence and defines students who performed satisfactorily. Although standards may be set through arbitrary decisions, standard setting is a judgmental process that results in defensible pass-fail standards in a systematic, reproducible, and defensible manner.[1-3] Many studies on standard setting methods have been conducted in the area of written assessments. However, recent studies have been focused on setting cut-scores for performance tests like objective structured clinical examinations (OSCEs).[4-11]

Standard setting procedures can be categorized as either exam-centered, in which the content of the test is reviewed by the expert judges (e.g., Angoff method) or examinee-centered, where expert decisions are based on the actual performance of the examinees.[2,3,12,13] One of these latest methods is the borderline regression method (BRM). In the BRM, a rater evaluates student's performance at each station by completing a checklist and a global rating scale. The checklist marks from all examinees at each station are then regressed on the attributed global rating scores, providing a linear equation. The global score representing borderline performance (e.g., 2 on the global performance rating scale) is substituted into the equation to predict the pass-fail cut-score for the checklist marks.[5]

There are several advantages to this method: It is based on actual performance of all examinees, it uses the judgments of expert examiners, and it is not time consuming.[5,8,14] Yet, another important advantage of BRM is that it can be used to generate metrics to evaluate the quality of an OSCE. These include the R^2 coefficient, the adjusted value of R^2, and the inter-grade discrimination.[15]

Considering the above mentioned advantages of the BRM, it is important to prove that it is a reliable procedure for standard setting. Earlier studies have calculated the precision for a single application of the BRM (average checklist score vs. average global

Address for correspondence: Dr. Mohammad Jalili, 7th floor, Tehran University of Medical Sciences, Ghods Street, Keshavarz Blvd, Tehran, Iran.
E-mail: mjalili@tums.ac.ir
Received: 12-01-2013; **Revised:** 28-01-2013; **Accepted:** 28-04-2013

score).[6,10] The aim of this study is to assess the reliability of BRM as a standard setting method for a pre-internship OSCE, where the overall OSCE pass-fail standard was calculated by averaging the BRM standards obtained for each station separately.

METHODS

In this study, a 14-station OSCE was administered to 105 medical students prior to internship phase at Tehran University of Medical Sciences in 2010. The fourteen 4 min stations represented different domains of clinical skills relevant to clerkship experience. Five stations using the written questions were excluded from the analysis. In the following part of the paper, we will use the term OSCE to indicate the nine-station performance-based subtest. In the nine stations with patient encounters, the examiners directly observed student's performance and gave two scores: The checklist score (percentage correct, 0-100) and the global rating score (1: Fail, 2: Borderline, 3: Sufficient, 4: Good, and 5: Excellent). The raters were instructed to give the global score based on their overall impression of the examiner's candidates and not to convert the checklist score into a global rating. To make this even harder to occur, the raters were not supposed to sum up the checklist scores of the candidate in that station. The total test score was calculated by averaging the station checklist scores. The global rating was only used for standard setting purpose.

The BRM was applied to establish a standard. For each station, we used a linear regression model in which the student's checklist scores and global scores were considered as dependent and independent variables, respectively. Then we calculated the checklist score cut-off on the regression equation for the global scale cut-off set at 2. The corresponding pass-fail standard for the OSCE (PFS_{OSCE}) was defined as the average of the nine station cut-scores. The percentage of students passing the OSCE accordingly is indicated as the pass rate.

To assess the quality of OSCE, the following metrics were calculated for each station: The R^2 coefficient (the squared linear correlation between the checklist score and the global rating score), and the inter-grade discrimination (the slope of the regression line).

To determine the reliability of the PFS_{OSCE}, the root mean square error (RMSE) of the estimated standard was calculated: The lower the RMSE, the more reliable the standard is. For this purpose, the regression-based method to calculate the precision for a single application of the BRM (OSCE average checklist score vs. OSCE average global score) presented in Muijtjens *et al*. was extended.[6] The extension provides an estimate of the RMSE for the current situation where the OSCE standard is obtained by averaging the checklist cut-off scores that were obtained by applying BRM for each station separately.[10]

Assuming that the error in the checklist cut-off scores is independent over the M stations of the OSCE for the error in the OSCE checklist standard it holds:

$$RMSE_{OSCE} = \sqrt{\frac{1}{M^2}\cdot\frac{1}{n}\sum_{i=1}^{M}\left\{s^2_{regr,i}\left(1+\frac{(G_0-Mean_{G,i})}{[(n-1)/n]SD^2_{G,i}}\right)\right\}}$$

Where, M is the number of stations, n is the number of candidates attending the OSCE, $s_{regr,I}$ is the standard error of estimate of the regression (estimate of the standard deviation (SD) of the residual error in the regression) for the i^{th} station, $Mean_{G,i}$ and $SD_{G,I}$ are the mean and SD of the student's global scores G_i for the i^{th} station, respectively, and G_0 is the cut-off value of the global score, which is identical for all stations.

For each station separately, say for station i, the corresponding RMSE can be obtained on the basis of the expression above with some plausible modifications: Dropping the summation leaving only the i^{th} term, and setting M equal to one.

RESULTS

For each of the nine stations in the OSCE Figure 1 shows the scatter plot of the checklist score versus the global score for the 105 candidates attending the OSCE. Each circle indicates the result of a candidate. However, it should be noted that the scores of some students may be identical and will result in coinciding circles in the plot. This is clearly demonstrated in the panel of the splinting station [Figure 1, second row, second column] where the circle at the point (global score = 1, checklist score = 0) represents 65 candidates having the same result. Each panel presents the linear regression of checklist score versus global score (solid line), the pass-fail cut-off value for the global score (equal to two, vertical broken line), and the corresponding BRM pass-fail cut-off value for the checklist score (horizontal broken line). The lower right panel (Total) shows the scatter plot for the mean global score and mean checklist score (total test score), where the mean is taken by averaging a candidates scores over the nine stations of the OSCE. The broken line indicates the OSCE checklist standard, which is obtained by averaging the BRM cut-off scores of the nine stations in the OSCE.

Performance of students in the pre-internship OSCE resulted in a mean total test score of 60.78 (SD = 8.04). The Pass-Fail Standard of the OSCE was 47.37. The RMSE of the

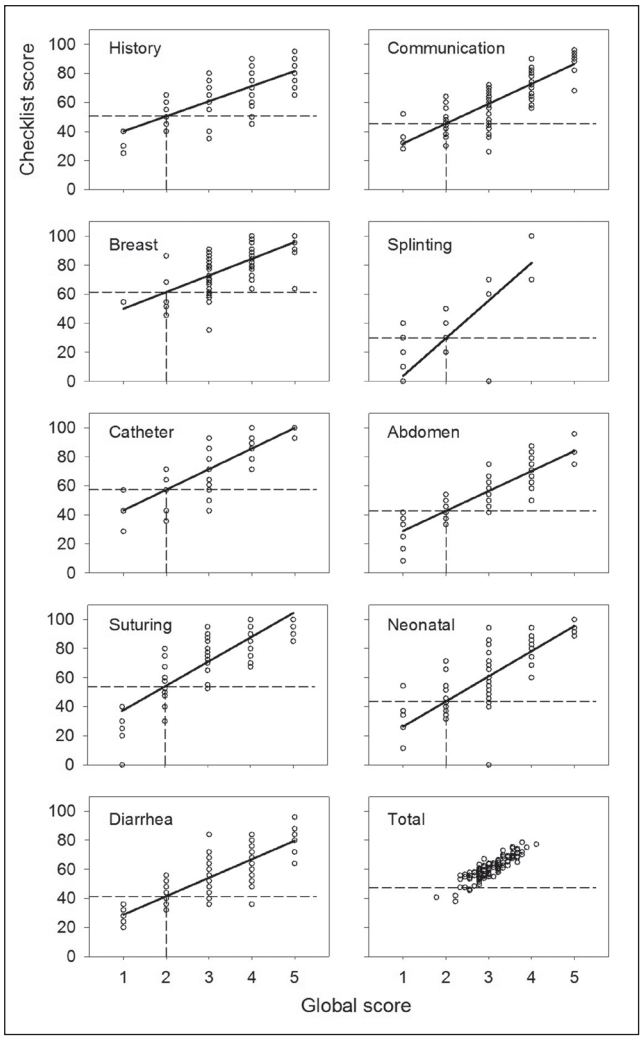

Figure 1: Scatter plots of the checklist score versus the global score for the nine stations in the in the pre-internship objective structured clinical examination (OSCE) with 105 candidates. Each panel presents the linear regression of checklist score versus global score (solid line), the pass-fail cut-off value for the global score (equal to 2, vertical broken line), and the corresponding pass-fail cut-off value for the checklist score (horizontal broken line) according to the borderline regression method (BRM). The lower right panel (total) shows the scatterplot of the mean global and checklist scores over the nine stations for the 105 candidates, the broken line indicating the pass-fail cut-off score for the mean checklist score (total score); the latter cut-off score was obtained by averaging the BRM cut-off scores of the nine stations in the OSCE

standard was 0.55, which is very small compared to the SD of the total test score amounting to 8.04, thereby indicating that the standard is sufficiently reliable. The percentage of students passing the whole exam was 95.2% [see lower right panel of Figure 1]. Descriptive statistics including the Pass-Fail standards and the corresponding RMSE and pass rate for each station are presented in Table 1. The mean student checklist scores and standard deviation for each station are also displayed. As shown in the table, the lowest and highest pass rates were obtained in splinting (19.0%) and breast examination (89.5%) stations, respectively. The most accurate BRM standard was found for the abdominal examination station (RMSE equal to 0.98) while the least accurate standard was found for the breast examination station (RMSE equal to 2.27).

The degree of linear correlation (R^2) between the checklist score and the overall global rating ranged from 0.44 to 0.79, with the highest value pertaining to the abdominal examination station, and falling below the threshold of 0.5 in only one station (breast examination). The slope of the regression line varied greatly among stations. In splinting station, for instance, an increase of more than 25 points in the checklist score was required to produce a one-point increment in the global rating scores [Table 1].

DISCUSSION

BRM as a standard setting method is much more convenient and less resource consuming compared to other procedures like Angoff. Furthermore, owing to the fact that global grade is awarded in addition to the checklist score; BRM has the advantage of generating a number of indices that are useful in measuring the quality of the OSCEs. Considering the fact that BRM is widely used as a standard setting method, assessing its reliability is of paramount importance. The focus of this study was to evaluate the reliability of the BRM, using the RMSE for a pre-internship OSCE, where the OSCE pass-fail standard was calculated by averaging the BRM standards obtained for each station separately.

Table 1: Scores, standards, pass rates, and metrics for the OSCE stations and the total test

Station number	Station name	Students' mean score (SD)	BRM standard (RMSE)	Pass rate (%)	R^2	Inter-grade discrimination
1	History taking	66.36 (13.80)	50.49 (1.68)	87.6	0.55	10.34
2	Communication skill	62.59 (16.60)	45.42 (1.61)	84.8	0.64	13.67
3	Breast examination	79.67 (13.92)	61.45 (2.27)	89.5	0.44	11.46
4	Splinting	11.24 (19.25)	29.74 (1.51)	19.0	0.72	25.90
5	Foley catheterization	77.82 (16.95)	57.41 (1.41)	84.8	0.75	14.20
6	Abdominal examination	56.15 (15.58)	42.79 (0.98)	77.1	0.79	13.76
7	Suturing	78.50 (21.08)	53.94 (1.84)	87.6	0.73	16.85
8	Neonatal examination	60.46 (21.29)	43.64 (1.76)	74.3	0.64	17.31
9	Approach to diarrhea	54.25 (17.33)	41.52 (1.41)	73.3	0.63	12.72
	Total test	60.78 (8.04)	47.37 (0.55)	95.2		

OSCE=Objective structured clinical examination; BRM=Borderline regression method; RMSE=Root mean square error

Overall, the low RMSE of the total OSCE cut-score shows a high reliability of the standard setting procedure. The results are comparable with several other studies, which employed a similar technique to assess the reliability of the BRM [Table 2]. Overall, the standard error is approximately half a point on a percentage scale. For taking decisions, we might multiply this standard error with 1.96 for a confidence level of 95%. That means that the BRM produces a standard that could be ±1% range on the checklist scoring scale. If we had set our pass mark 1% lower, our pass rate would have been the same (95.2%). If we had set our pass mark 1% higher, the pass rate would have been 93.3%. That means that the noise caused by the BRM leads to an approximate 1.9% of shifts in pass/fail decisions. With an increasing number of examinees and/or increasing number of stations these results might even improve, because the RMSE would decrease (and the reliability would increase).[10]

The relatively low RMSE of the BRM standard for the abdominal examination station is consistent with the strong correlation expressed by the high R^2 for this station. It is due to the spread of points over the whole range of the two score scales (checklist and global) in combination with a fairly strong relation between the two. It indicates that the station is of adequate difficulty and sufficiently sensitive to tap performance differences consistently from both perspectives. The opposite situation is found for the breast examination: Low R^2 and high RMSE. This point merits further explanation: With this station, global scores are mainly concentrated at levels three and four and within each of these levels the checklist scores are widely spread. These characteristics indicate that this station lacks discriminative power, and the validity of the checklist and/or the global score is questionable.

Generally, in all except one station, higher overall global ratings corresponded with higher checklist scores, giving rise to greater values of R^2 coefficient (0.55-0.79). This is similar to the study conducted by Homer and Pell, in which at each station, the two variables always showed a significant positive correlation, varying in size from 0.659 to 0.865.[16] As shown in Table 1, station three (breast examination) is less satisfactory in this regard, with an R^2 value of 0.44. The main problem with this station is a wide-spread of checklist scores for each global grade [Figure 1]. This unsatisfactory relationship demonstrates some degree of non-linearity. Pell *et al*. suggest that in this situation, other methods rather than linear regression model may provide a better explanation.[15] In our case, adding quadratic and/or a cubic term does not change the fitted relation considerably, and hardly increases the R^2 (linear + quadratic: $R^2 = 0.440$, linear + quadratic + cubic: $R^2 = 0.443$). We think this kind of low correlation between global and checklist score indicates that one of the two measures or both are unreliable and/or invalid or they regard very different aspects of performance.

On the other hand, we should be cautious when interpreting the R^2 values because if raters automatically translated checklist score into a corresponding global score, the R^2 would have artificially been inflated.[15] Other psychometric indicators of quality should be used to identify possible problems.[15] As an example, station four, which had a high failure rate also showed an unacceptable inter-grade discrimination. Although no clear guidance on "ideal" value for inter-grade discrimination exists, Association for Medical Education in Europe guide no. 49 recommends this value should be "of the order of a 10th of the maximum available checklist mark".[15] Hence, we considered values below 20 as tolerable (the maximum checklist score was 100). For the splinting station, the distribution of the points in the scatter plot is not adequate for a reliable regression result: The large majority of points are concentrated at the lower left and only a few very influential points at the upper right support the steep regression line. The extreme skewedness of the score distribution is also indicated by the very low mean value for this station: 11.24. Obviously, the station is too difficult or the candidates were not adequately trained for the skills required for this station. In summary, although considering a station to be flawed solely based on the high number of failures is an incorrect assumption,[15] scrutiny of station performance may inform curriculum effectiveness.

There are some limitations in our study. First, generalizability of the results of the present study may be limited by the fact that it was based on one rather small sample of 105 students in a single test. However, this study confirms the results of Kramer *et al*. and Schoonheim *et al*.; thus, we believe that the findings of this study can be extended to a wider context.

Table 2: Comparison of RMSEs in different studies that assessed the reliability of BRM

Study	Number of stations	Number of trainees	Pass-fail standard %	RMSE
Kramer *et al*. (2003)	16	86	57.6	0.6
Hobma *et al*. (2004)	8*	88	41.6†	0.41
Schoonheim-Klein *et al*. (2009)	14	119	55.1	0.3
Current study	9	105	47.37	0.55

BRM=Borderline regression method; RMSE=Root mean square error. *Calculated by averaging 704 video recorded consultations of 88 GPs observed; †Calculated by adapting an original score of 2.5 on a global scale ranging from 0 to 6

Secondly, we used data only from nine out of 14 stations of the original OSCE. Finally, the main disadvantage to using RMSE approach in assessing reliability of BRM procedure is statistical complexity.

CONCLUSION

The current study confirms that using RMSE is an efficient method of assessing the reliability of BRM. It also proves that BRM is a reliable method of setting standard for OSCE and has the advantage of providing data for quality assurance.

ACKNOWLEDGMENTS

The authors would like to thank Azim Mirzazadeh MD, Director of the Education Development Office, School of Medicine, TUMS, and Ali Labaf MD, Director of the Clinical Skills Centre, School of Medicine, TUMS, for their aid with the design and implementation of the OSCE, and also for their constant support during this project.

REFERENCES

1. Cusinamo MD. Standard setting in medical education. Acad Med 1996:71:112-20.
2. Norcini JJ. Setting standards on educational tests. Med Educ 2003;37:464-9.
3. Cizek GJ, Bunch MB. Standard setting: A guide to establishing and evaluating performance standards for tests. Thousand Oaks, CA: Sage Publications, Inc.; 2007. p. 20-2.
4. Wilkinson TJ, Newble DI, Frampton CM. 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.
5. 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. Objective structured clinical examinations. Med Educ 2003;37:132-9.
6. Muijtjens AM, Kramer AW, Kaufman DM, van de Vleuten C. Using resampling to estimate the precision of an empirical standard setting method. Appl Meas Educ 2003;16:245-56.
7. Hobma SO, Ram PM, Muijtjens AM, Grol RP, van der Vleuten CP. Setting a standard for performance assessment of doctor-patient communication in general practice. Med Educ 2004;38:1244-5.
8. Wood TJ, Humphrey-Murto SM, Norman GR. Standard setting in a small scale OSCE: A comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Adv Health Sci Educ Theory Pract 2006;11:115-22.
9. Boursicot KA, Roberts TE, Pell G. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ 2007;41:1024-31.
10. 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;13:162-71.
11. Jalili M, Hejri SM, Norcini JJ. Comparison of two methods of standard setting: The performance of the three-level Angoff method. Med Educ 2011;45:1199-208.
12. Kane M. Choosing between examinee-centered and test-centered standard-setting methods. Educ Assess 1998;5:129-45.
13. Liu M, Liu KM. Setting pass scores for clinical skills assessment. Kaohsiung J Med Sci 2008;24:656-3.
14. Davison I, Cooper R, Bullock A. The objective structured public health examination: A study of reliability using multi-level analysis. Med Teach 2010;32:582-5.
15. Pell G, Fuller R, Homer M, Roberts T, International Association for Medical Education. How to measure the quality of the OSCE: A review of metrics-AMEE guide no. 49. Med Teach 2010;32:802-11.
16. Homer M, Pell G. The impact of the inclusion of simulated patient ratings on the reliability of OSCE assessments under the borderline regression method. Med Teach 2009;31:420-5.

Source of Support: This study was funded by Tehran University of Medical Sciences. **Conflict of Interest:** None declared.