International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2019): 7.583

Procedure of Standard Setting Using Borderline Regression Methods in Objective Structured Clinical Examination of Undergraduate Medical Education

Shahid Hassan

School of Medicine at International Medical University, Kuala Lumpur Malaysia

Abstract: ***Standard setting with borderline regression method (BRM) is a practical and defensible process to determine the cut-off point in Objective Structured Clinical Examination (OSCE). For standard setting method in OSCE, two different scores, the checklist scores and the global rating scores were recorded. These scores were used to apply standard setting using BRM to calculate the pass/fail cut-point in five stations OSCE in an undergraduate programme. The traditionally set pass marks at 50% wasreestablished to be 54.98% using BRM. To assess the quality of OSCE, the auto-generated correlation of determination (R^2) ranging from 19.9% to 56.6% and slope of regression line "B" coefficient ranging from 4.00 to 10.32 points among the 5-stations OSCEwas also considered. Root Mean Square Error (RMSE) was determined as an absolute indicator of reliability in a composite scaled score of multi-dimensional OSCE stations. BRM was found a less time consuming and practical method of standard setting for OSCE stations. It also has the advantage of providing data for quality assurance. RMSE is more appropriate to determine the fitness of model than Cronbach's alpha.$R^2$and the intergrade discrimination provides the other relative indices of reliability in multi stations OSCE of students' clinical performance.***

Keywords: Objective structured clinical examination, Standard setting, Borderline regression method, Absolute and relative reliability

1. Introduction

Standard setting method is a judgmental process that results in pass-fail standards in a systematic, reproducible, credible and defensible manner [1, 2, 3]. For standard setting method of OSCE stations with patient encounters, the examiners directly observed students and their ratingof clinical performance using two different scores.The first rating of clinical performanceis the marks assigned to students as checklist score converted to percentage grade(0-100%) and the second score may base on global rating using a Likert scale ranging 3-5 categories such as 1= Fail, 2= Borderline, 3=Pass, 4 = Good, and 5= Excellent.

In BMR the evaluators give their global rating based on students' overall clinical competence gauged through the learning outcome of relevant station and the students' observed performance. The evaluatorsare advised to avoid summing up the checklist score of the candidate at a station in order to have global rating score independent of any influence, whatsoever reflected in checklist score. Practiced otherwise, a global clinical rating often unintentionally becomes a simple conversion of a checklist score into global rating score. The total test score calculated by summing up the station checklist scores is produced as graph against the global rating score for standard setting. The BRMto set a standard (cut-off passing mark)is different from Borderline Group Method (BGM) generally used [4, 5, 6]. For each station, a linear regression model is usedin which the student's checklist scores and global rating scores are considered as dependent and independent scores respectively.In statistical modeling regression analysis is a set of statistical process for estimating a trend between a dependent variable at Y-axis and an independent variable at X-axis. Regression analysis is a form of predictive modeling technique, which investigate the relationship between a response and a predictor variable for strength and direction.

BRM is more reliable and valid than the borderline group method (BGM). Reliability may be inflated if the global rating and the checklist rating are marked close to each other. In case of OSCE it is more like a composite score forming a scale rather than a continuous score. Therefore, to talk about reliability of OSCE we talk of other reliability indicators and not Cronbach's alpha alone.RMSE is more appropriate to determine the fitness of model than Cronbach's alpha. RMSE is the standard deviation of the residuals (prediction errors) from line of fit, which tells us about how far from the regression line data points are. RMSE is a measure of how spread out these residuals are or how concentrated the data is around the line of best fit.RMSE error is the frequently used measure of difference between values predicted by a model (estimator) and the values actually observed. In RMSE, using a model the residuals (errors) is first calculated. Residuals are squared, added up and divided by the total number of sample in population or it may take the square root of variance. The general formula for RMSE = Σ $[(r^1+ r^2+r^1+r^3+r^4 \ldots r^n)/N$

Another statistical output to determine the reliability isR^2, the amount of variation along the Y-axis that can be explained by the variable along the X-axis. It is also the square of correlation coefficient (r) and is called Coefficient of Determination. R^2 tells how well regression line estimates an actual value or it is the amount of variation of dependent variable explained by the independent variable. The auto-generated $R^2$is considered the other relative indicator of reliability with intergrade discrimination found as line of slope in regression equation in multi stations OSCE to determine the students' clinical performance. RMSE is considered the absolute indicator of OSCE.

2. Material and Method

Five stations manned OSCE with a simulated patient and an examiner was administered as a pilot project to 136 students in their preclinical phase ofundergraduate students. Students were assigned marks (out of 10) using a checklist converted to percentage grade score. Examiner also provided a global rating on 5 points Likert-scale. The checklist cut-off score was calculated by a simple regression model (equation, Y = a + bx) regressed on global scale set at 2 defined forborderline students. The corresponding pass-fail score (PFS) as standardfor the OSCE in a standard setting method wascalculated by averaging the all OSCE stations cut-scores after individual station checklist score regressed on global score. The percentage of students passing the OSCE accordingly was indicated as the pass rate also obtained against a standard set score obtained using the borderline regression method.

Microsoft Excel worksheet was developed with independent variable along the X-axis in the first column and the dependent variable along the Y-axis in the second column. This arrangement is important for the statistical calculation of cut-off score and the RMSE subsequently. To assess the fitness of regression model as how good the model predicts the PFS of OSCE was statistically analyzed, examining the auto-generated correlation of determination"R^2 '' and the slope of the regression line as the intergrade discrimination"B". Besides, Root Mean Square Error (RMSE) was also calculated using formula inMicrosoft Excel. The step to calculate cut-off score and the RMSE in Microsoft Excel was as followed.

Borderline Regression Method:

1. The data was collectedas dependent variable (checklist score along Y-axis) and independent variable (global score along X-axis).

2. A scatter plot was created and a linear line of best fit was added to this graphic output.

3. The pass markwas identified as the cut-off point where the line of best fit is crossed bya vertical line drawn up from the borderline judgment set at 2 (borderline students' score) of global rating.

4. Alternatively, the pass-fail cut-off score could also be calculated from the regression equation (Y = a + bx) auto-generated in scatterplot with an additional click or running the regression statistics in data analysis option."Y" in the equation indicates the cut-off score, "a "the line of intercept, where the line of slope cuts the Y-axis. "b" indicates the line of slope multiplied by "x" value provided by borderline global rating set prior to calculation of cut-off point.

5. The fitness of regression model (equation) was evaluated by examining the correlation of determination R^2, which ranges from 0 to 1.$R^2$is the proportion of squared sum of regressionand squared sum of total as SSR/SST respectively, whereas SST = SSR+SSE (squared sum of error) (see figure 1).
6. The slope of regression linewas determined as the amount of change in number of checklist score as dependent variableagainst 1-unit increase in global score as independent variable called intergrade discrimination.

3. Result

The standard set passing marksfor each OSCE station was establishedto be 46.43% using BRMmean of all 5 OSCE stations shown as "total" in scatterplot matrix (see figure 2) and total test score average (see table). This however, showed an improvement of passing ratewith 76.47% (BRM) versus 67.05% (BGM) and 74.26% traditionallyselected passing marks of 50% to determine PFS of OSCE (see table). Borderline group method (BGM) was calculated using average of all checklist score of borderline students, which is considered less robust. To assess the fitness of regression model in OSCE, correlation of determination "R^2," and the slope of the regressionline (intergrade discrimination) "B" was also determined. 38.5% to 65.00% of BRM cut-score variation in different OSCE stationswere explained by global score rating decided by the experts."B" coefficient showed an increased 1 unit of global rating causing an increase in checklist score by 4.00 to 10.32 points among the 5-stations OSCE. This suggests small intergrade discrimination in favor of checklist list cut-point validity.

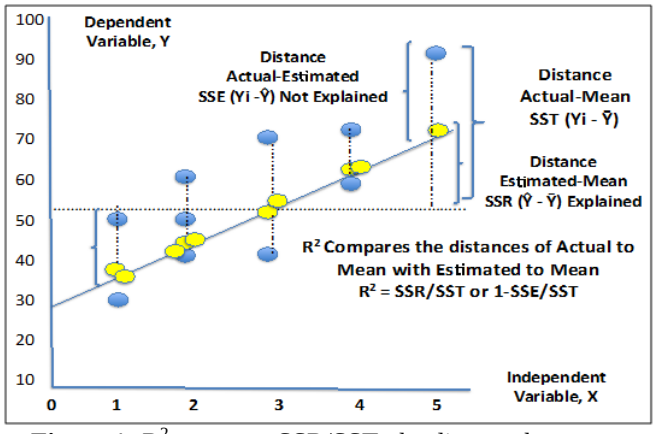

Figure 1: R^2compares SSR/SST, the distance between estimated score from the mean score (SSR) with actual score from mean score (SST) and SST=SSR+SSE (sum of the square error)

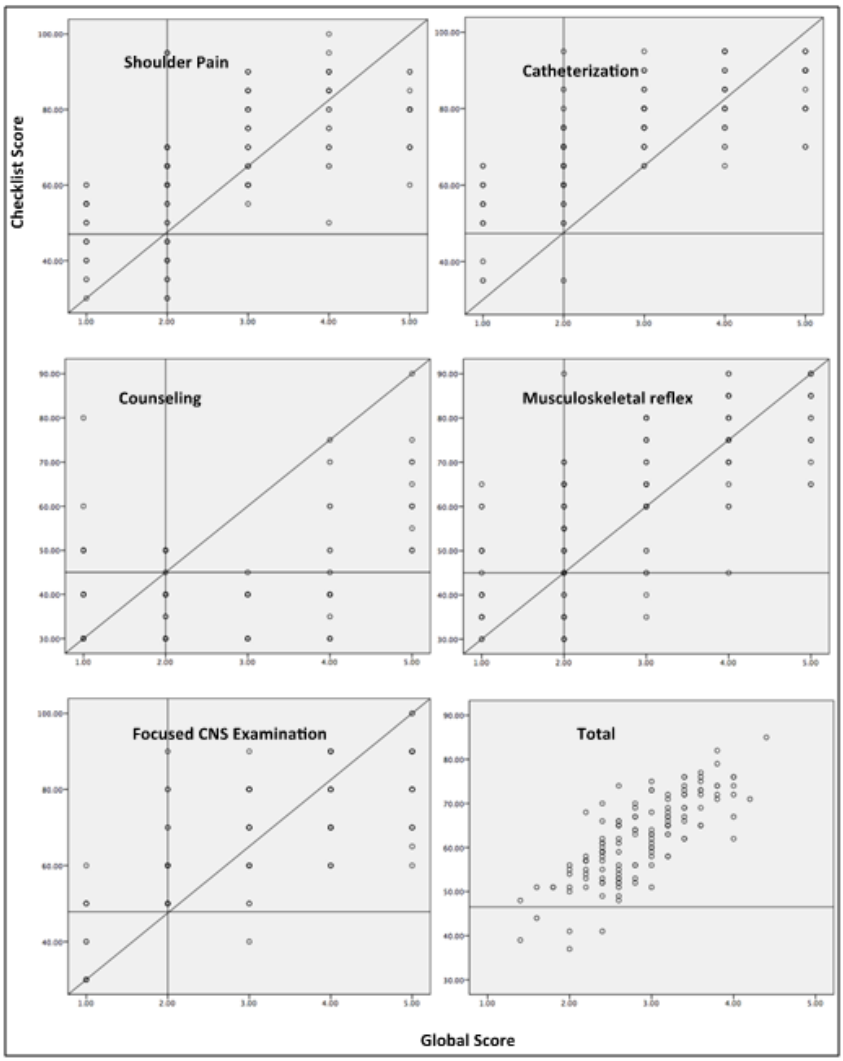

Figure 2: Scatter plots of the checklist score versus the global score for the five stations OSCE in preclinical phase of 136 candidates. Each panel presents checklist score regressed against global score

Table 1: Scores (mean and standard deviation), BGM standards, BRM metrics (cut-score, RMSE, $R^2$and B) and pass rate of OSCE in preclinical phaseof undergraduate programme.

No	Station Name	Mean Score (SD)	BGM Standard (No of St.)	BRM/SRME Standard (No of St.)	R^2	B	Pass Rate (%) BGM/BRM (At 50%)
1	History taking (Shoulder pain)	66.06 (16.43)	58.57 (99)	59.90/0.761 (99)	.428	8.76	72.79/72.79 (83.82%)
2	Procedural Skills (U. Catheterization)	73.86 (13.63)	65.14 (96)	65.00/0.490 (107)	.551	8.42	70.59/78.68 (97.06%)
3	Communication skills (Counseling)	40.95 (11.96)	37.33 (86)	38.50/0.455 (86)	.190	4.00	63.23/63.23 (24.26%)
4	Examination skills (M. Skeletal reflex)	60.84 (17.48)	50.41 (87)	52.5/0.381 (87)	.566	10.32	63.97/63.97 (72.06%)
5	Examination skills (Focus neurogenic)	70.18 (15.23)	60.36 (88)	59.00/0.374 (115)	.499	9.01	64.71/84.56 (94.85%)
	Total Test Score Average	62.38 (14.95)	54.41 (91.20)	54.98/0.492 (98.80)	-	-	76.06/72.65 (74.41)

BGM: Borderline group method, BRM: Borderline regression method, R^2: Correlation of determination and B: Intergrade discrimination.

4. Discussion

BRM is an examinee-centered standard setting method categorized by experts' decision based on examinee's actual performance. There are several advantages of BRM such as examinees' reliable performance, experts' opinion, less time consuming and taking each and every score into account to decide on PFS on OSCE [7, 8, 9]. However, reliability may be inflated if the global rating and the checklist rating are marked close to each other. BRM can also be used to review the appropriateness of OSCE checklist. BRM, though more reliable and valid than the BGM has been evaluated for reliability using RMSE [10] before implementing the BRM to decide on PFS.Cronbach's alpha as the test of consistency in a multi-dimensional scale of scores may be questioned for reliability?Whether Cronbach's alpha is the right choice with or without principle component analysis or RMSE in evaluation of OSCE? Author believes the right choice for establishing the reliability of OSCE is RMSE and we need to go beyond alpha for reliability of OSCE with RMSE as the absolute test of reliability and R^2(correlation of determination)and B (intergrade discrimination) as the other relative statistics of reliability in OSCE.Root Mean Square Error (RMSE) offers an efficient method of assessing the reliability of BRM addressing the issues associated with Cronbach's alpha such as issues of dimensionality with OSCE stations and dependence on number of stations or students associated with OSCE. RMSE represent the goodness of line of fit or to what extent this linear fit represents the data (score) well. A small value of RMSE indicates that BRM is a reliable method of setting standard for OSCE. This has the advantage of providing data for quality assurance in post-examination evaluation of assessment. However, with an increasing number of examinees and/or increasing number of stations the RMSE would decrease and the reliability would increase [11].

Another important advantage of BRM is that regression analysis can also generate metric that may help to evaluate the quality of OSCE [12]. The auto-generatedstatistics are correlation coefficient (R), correlation of determination (R^2), slope of regression line as intergrade discrimination coefficient and equation of regression required to calculate RMSE, all related to reliability of OSCE. R^2 value tells about percentage variation of checklist score explained by globalrating. In investing reliability, R^2 is generally interpreted as the percentage of a score achieved in a test that can be explained by a benchmark score of global expert rating on students' clinical performance. An R^2 of 100% means that all scores of checklist scores are completely explained by the global rating of experts as an independent variable that one may be interested in.In investing, a high R^2, between 85% and 100%, indicates that the checklist of students' clinical performance is aligned with expert's opinion reflected in global rating score. A cut-off score with a low R^2, at 50% or less, indicates the checklist score is not aligned with global rating score. R^2 gives an estimate of the relationship between score of a dependent variable and scores of an independent variable. However, it doesn't tell whether a chosen model is good or bad, nor will it tell whether the data and predictions on a given data (OSCE) is biased. A high or low R-squared isn't necessarily good or bad, as it doesn't convey the absolute reliability of the model, nor whether a rightly chosen regression. One can get a low R^2 for a good model, or a high R-square for a poorly fitted model, and vice versa. Therefore, R^2 alone for reliability of OSCE may not be a good choice unless RMSE shows a low value.

Slope of regression line, besides an indicator of direct linear relationship between independent and dependent variable also tells about inter-grade discrimination as how much increase in checklist points is incurred with 1-unit increase in global rating score. Small inter-grade discrimination value in case of BRM is considered better than large value since global rating has been held more valid than the checklist rating by many researchers in published literature [13, 14, 15, 16].There is no clear guidance about the value for inter-grade discrimination however, Association for Medical Education in Europe guide No. 49 recommends this value of the order of a 10^{th} of the maximum available checklist mark score [17].

5. Conclusion

Borderline Regression Method is a judgmental process of determining pass-fail standards of examinees score in a systematic, reproducible, and defensible manner. This is less time consuming method of standard setting for manned OSCE stations with students, simulated patients and evaluators encounter. It also has the advantage of providing data for quality assurance with auto-generated statistical data of correlation of determination (R^2) and intergrade discrimination besides, RMSE as the absolute reliability coefficient of OSCE.

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. Thousand Oaks, CA: Sage Publications, Inc; 2007. Standard setting: A guide to establishing and evaluating performance standards for tests; pp. 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] Kane M. Choosing between examinee-centered and test-centered standard-setting methods. Educ Assess. 1998; 5:129–45.
[6] Liu M, Liu KM. Setting pass scores for clinical skills assessment. Kaohsiung J Med Sci. 2008; 24:656–3.
[7] 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.
[8] Wood TJ, Humphrey-Murto SM, Norman GR.

Standard setting in a small scale OSCE: A comparison of Modified Borderline-Group Method and the Borderline Regression Method. Adv Health Sci Educ Theory Pract. 2006; 11:115–22.

[9] 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.

[10] Sara Mortaz Hejri, Mohammad Jalili, Arno M. M. Muijtjens, and Cees P. M. Van Der Vleuten. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci. 2013 Oct; 18(10): 887–891.

[11] 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.

[12] 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.

[13] Bunmi S. Malau-Aduli, Sue Mulcahy, Emma Warnecke, Petr Otahal, Peta-Ann Teague, Richard Turner, Cees Van der Vleuten. Inter-Rater Reliability: Comparison of Checklist and Global Scoring for OSCEs. 2012. Vol.3, Special Issue, 937-942; in SciRes (http://www.SciRP.org/journal/ce).

[14] Joong Hiong Sim, Yang Faridah Abdul Aziz, Anushya Vijayananthan, Azura Mansor, Jamuna Vadivelu, Hamimah Hassan. A Closer Look at Checklist Scoring and Global Rating for Four OSCE Stations: Do the Scores Correlate Well? Volume 7 Issue 2 2015 DOI: 10.5959/eimj.v7i2.341 www.eduimed.com.

[15] Kaitlin Turner, Maegan Bell, Lindsay Bays, Carmen Lau, Clara Lai, Tetyana Kendzerska, Cathy Evans, Robyn Davies. Correlation between Global Rating Scale and Specific Checklist Scores for Professional Behaviour of Physical Therapy Students in Practical Examinations. Education Research International Volume, 2014, Article, ID, 219512, 6, pages, http://dx.doi.org/10.1155/2014/219512.

[16] Sajesh Kalkandi Veettil, Kingston Rajiah**.** Impact of Task-based Checklist Scoring and Two Domains Global Rating Scale in Objective Structured Clinical Examination of Pharmacy Students. Indian Journal of Pharmaceutical Education and Research. Vol 50 Issue 1 Jan-Mar, 2016.

[17] 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