Can borderline-regression method be used to standard set OSCEs in small cohorts?

Introduction:

Absolute methods of standard setting are more suitable for OSCEs. However, previous studies have theorized that borderline regression method (BRM) is not reliable for small sample sizes.

Materials and methods:

OSCE results for the 2017-2019 cohorts were analysed to compare BRM versus Modified Angoff. We reported on whether the stations in multiple cohorts were sufficiently equivalent to aggregate the results together to calculate which method of standard setting was more reliable. We finally used the bootstrapping method to compare the accuracy of BRM for small versus simulated larger cohorts.

Results:

BRM was a valid method for standard setting OSCEs in this dataset when station quality was sufficiently high. However, a large gap between the Angoff and BRM in some of the OSCEs was observed, which could be explained by poor use of the grading scale. Model fit statistics were generally adequate even with low sample sizes. Using the bootstrap of datasets, the error rate was much higher for low quality stations but was not an issue in high quality ones.

Discussion:

This study adds to the evidence that well-designed OSCEs can use BRM for small cohorts. However, there is a need for the institutions to properly assess their stations and their assessors, before embarking into using this method, to prevent from having to remove stations.

Conclusions:

This analysis suggests that BRM is an acceptable replacement for Angoff standard setting in small cohorts where there is a range of candidates undertaking the assessments and there are well-designed OSCES with well-trained examiners.

Key words: Standard setting, OSCE, borderline-regression, Modified Angoff

Introduction:

Performance-based examinations are assessments developed to consider the clinical competencies of healthcare students. OSCEs (Objective Structured Clinical Examinations) are clinical competency tests where the students rotate between a number of test stations with a task to be completed in 5 to 10 minutes. In each station the student's performance is observed by an examiner using a checklist or rating scales[1].

Standard-setting is the process of defining or judging the level of knowledge and skill requirement to meet a typical level of performance and then identifying a score on the examination score scale that corresponds to that performance standard. This process is designed to translate a conceptual definition of competence to an operational version called the passing score [2]. It is a judgmental process and thus needs to be defensible and controllable[3].

Standard setting (SS) methods can be divided into two categories: 1) Relative or norm-referenced (where the standard of the test is based on the performance of the students by establishing which percentage of students will pass the assessment) and 2) Absolute or criterion-referenced (where the passing score is defined by how many items or tasks are required to be performed correctly in order to pass)[2,4]. When assessing OSCEs, absolute methods are more suitable as we are assessing a specific task that has been very well defined[5].

Within the absolute methods we can distinguish two procedures: rational (test-centered) and empirical (examinee-centered). When using rational methods the judges focus on exam content. Examples of this method are Angoff and Ebel[6]. When Angoff SS method is used, a group of expert examiners make a judgement on each station and estimate the marks that a borderline candidate (a student who would just pass, not an average student) would get. An average of the scores is generated[7] and this determines the pass score for each station. The addition of all the pass scores for each individual station will determine the pass score of the examination. The modified Angoff method follows the same structure described above, however there is discussion amongst judges when there are large discrepancies of marks. This discussion might prompt changing the scores of the outliers and an agreed pass score for each station is agreed[6].

The modified Angoff method has the advantages that it is founded on the content of the test[5], however it has been reported that standards may be set unrealistically high for this type of examination[5,8] and that ideally items in the checklist should be independent of each other. This is not the case for OSCEs, where items are usually nested within and dependent on each other[8]. This method is also expensive, time-consuming and organizationally complex[5].

Empirical methods on the other hand determine the standard by judging the performance of each individual candidate relative to a performance standard based on some external criteria or on the overall test performance. They provide a more holistic approach to SS and is therefore more appropriate for performance-based assessments[5]. Examples of this method are contrasting groups and borderline regression method (BRM). We will focus on the latter as it has been proven to be more reliable when setting standard in smaller cohorts[9]. To calculate the SS pass mark, examiners assess the students using a checklist and will also give each candidate a global rating (this is the clinical impression that the examiner has of the candidate's performance on that station regardless of the marks they got on the checklist). These global ratings are categorized into unsatisfactory, borderline, satisfactory, highly satisfactory or excellent. This method then regresses all of the candidates' checklists scores onto their global ratings to produce a linear equation. Where the model is well-fitting, this method has proven that by inserting the midpoint of the global rating scale corresponding to the borderline group into this equation, a corresponding predicted checklist score can be determined. This score becomes the pass mark for that station[10]. As this regression line is fit through cohort performance data, it could be argued that it has a normative component or that it is a hybrid method[9]. Additions of these pass marks give the SS mark for the OSCE examination, including the patient partner scores.

This method has proven to have more benefits and to be more reliable than other borderline group methods and as such is the one that the literature suggests to be the best for performance-based examinations[1,2,5,10] Nonetheless Dwyer et al[11] concluded that modified Angoff method can be used to set acceptable and credible cut-scores, however they found no difference when comparing it with other methods such as the borderline regression method.

Schoonheim-Klein et al[1] suggested that only 50 students would suffice to obtain a reliable pass/fail standard within the borderline regression method, however the studies they compared their data to, had sample sizes of 50 or more candidates. Recently Homer et al[12] published a paper where he studied small cohort of PA (Physician Associate) students and compared this with Angoff scores; he concluded that BRM produced more realistic range of cut-scores compared to more delimited results when using Angoff.

The aims of our study were to evaluate the impact of sample size *via* simulation and compare the accuracy of borderline regression for the small *versus* simulated cohorts. We also wanted to assess reliability of borderline regression method by using the root mean square error (RMSE) as well as to compare standard setting pass mark from the borderline regression method with the current system we are using which is modified Angoff method. Finally we also wanted to analyse if adding the candidates marks of the same station used in multiple years was sufficiently equivalent to allow for use in standard setting, and if so how much data was likely to be necessary before new stations could use borderline regression for standard setting

Materials and methods:

Results data from the OSCE assessments carried out at the Institute of Dentistry at the University of Aberdeen from 2017 to 2019 were anonymised and analysed. This included OSCEs undertaken by the dental students in years 2, 3 and 4 of their Bachelors in Dental Surgery (BDS) undergraduate programme. BDS 2 students undertook 10 stations, BDS 3 12 stations and BDS 4 13 stations. Each station had a total of 25 marks and used a combination of checklists with rating scales to assess student's general clinical ability, professionalism and communication skills, amongst other domains. Each station also had a five-point global grade (1= unsatisfactory, 2= borderline, 3= satisfactory, 4=highly satisfactory and 5=excellent). Each station was marked by the same assessor, except in year 4 were 3 out of the 13 stations were marked by 3 assessors each which had been previously calibrated.

- Borderline Regression vs Modified Angoff:

The first step was to aggregate the data and check suitability for borderline regression. This involved checking the data for homogeneity (that same stations in different years

performed similarly, independent of the cohort of students), that the full range of global grades were being used and if the associations were linear. This was done by carrying out significance tests for means/standard deviations between years, with no significant difference found. Once the output data from the regression had been developed, we used the following formula to calculate our pass score[2]:

Pass score = (median of ratings x Variable 1) + Intercept

Median of ratings refers to the global ratings available (which is 5 in our case), variable 1 is the slope of the regression line and Intercept is where X meets Y on the regression line on a scatter plot chart.

The passing score was then adjusted by the standard error of estimation (labelled "standard error" in the summary output).

Following this initial evaluation of the data, we run borderline regression for all available stations and assessments and reported on the pass rates for borderline regression *versus* modified Angoff to inform us of the practical impact of any changes made.

We also reported on whether the stations in multiple cohorts were sufficiently equivalent to allow for use in standard setting, and how much data was likely to be necessary before new stations could use borderline regression for standard setting by doing the same analysis described above, where we added the scores in all the cohorts and measured if the passing score was more reliable.

The following metric were calculated and analysed to assess the quality of the OSCEs[13]:

- Cronbach's alpha to measure internal consistency (0.7 or higher is usually regarded as acceptable).
- Coefficient of determination R^2: is the proportional change in the dependent variable (checklist score) due to change in the independent variable (global rating). This allows to determine the degree of linear correlation between the checklist and the global rating at each station (R^2>5 will indicate a reasonable relationship between the two variables).
- Inter-grade discrimination: slope of the regression line which indicates the average increase in checklist mark corresponding to an increase of one grade on the global rating scale (should be in the order of a tenth of the maximum available checklist mark which is 25 at the Institute of Dentistry).

All OSCE stations had already been SS using the modified Angoff and we used these previous data to compare with the results we obtained using the model described above and checked the reliability of each method.

The reliability of the Angoff method was assessed using Generalisability theory by estimating the RMSE (Root Mean Square Error) of the pass/fail standard of each cluster and the total OSCE according to the following formula:

$$\mathrm{RMSE}_{\mathrm{Angoff}} = \sqrt{\frac{\sigma^2_{j:i}}{\sum N_{j:i}}}$$

where $\sigma^2_{j:i}$ i was the estimated variance component for the effect of judges nested in stations, the error, and $N_{j:i}$ was the number of judges providing Angoff estimates for the standard of the station (*i*).

For the borderline regression method the RMSE of the pass/fail standard of a cluster and the total OSCE was defined by the following formula:

$$\mathrm{RMSE}_{BR} = \sqrt{\frac{1}{M^2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{M} \left\{ s^2_{\mathrm{regr},i} \cdot \left(1 + \frac{(R_0 - \mathrm{MN}_{R,i})^2}{[(n_0 - 1)/n_0] \cdot \mathrm{SD}^2_{R,i}}\right)\right\}}$$

where M was the number of stations in the cluster, n_o the number of students who attended the stations of the cluster, $s2_{regr;i}$ was the standard error of estimate of the regression of the check list score on the global rating, and $MN_{R,i}$ and $SD_{R,i}$ the mean and standard deviation of the global score for the *i*th station in the cluster, respectively, R_o the pass/fail cut off for the global rating and n the (hypothetical) number of students for which $RMSE_{BR}$ was estimated[1,14]. Both methods were then compared using a t-test analysis to see which one was more reliable.

- Bootstrapping:

To be able to answer the question of whether the small sample size in each year prohibits the use of borderline regression on a single cohort's worth of data, we evaluated the impact of sample size *via* simulation, whereby we generated hypothetical large cohorts using our own data as an input and compare the accuracy of borderline regression for the small *versus* simulated cohorts. If the model fits were

similar, this would suggest that borderline regression even in small cohorts could be used, even for a new OSCE station, without building up a historical record of performance. Data was simulated using *R* based on the assessment results in the cohort.

Bootstrapping is a form of simulation. From an initial starting point of our small cohort it is possible to create much larger simulated datasets. In this process, each student has an equal likelihood of being "drawn" from the original dataset and added to the bootstrapping sample. The sampling process is repeated a given number of times (e.g. 1000), and "real" students appears in the bootstrap dataset multiple times (Figure 1). This process approximates an acceptably realistic large sample of actual students.

Once the bootstrap dataset is available, it is then possible to rerun the BRM on both the original real dataset, and the bootstrap dataset. The larger the difference between the two model fit statistics, the less confident we can be in the reliability of the real dataset[15].

Results:

Following from aggregating all the data and carrying out borderline regression on all datasets, we generated nine reports on the psychometric characteristics of the stations. This involved:

1) A report of internal consistency (alpha) to see if the assessment was reliable enough to make defensible judgments
2) A station-by-station report of how much variance was explained by the association between global judgments and scores (with an R^2 of over 0.5 being preferred)
3) A comparison of a coerced linear (straight line) model vs. an unrestricted (curved line) model. The comparison or linear and non-linear models is a visual inspection. Since BRM assumes a linear association between the two variables, where the model is permitted to be non-linear it should still approximate a straight line. If the line is visible curved (especially if e.g. it approximates a u-shaped curve) this indicates serious issues with the model fit. (Figure 2)

4) A comment on whether the data is minimally adequate to allow for borderline regression (if less than 4 global grades were used)

Table 1 summarises the results, including the reliability (alpha) and the number of stations with undesirable characteristics such as (1) being such a poor fit it should be removed from the analysis (2) having low R^2 (3) having significant nonlinearity and (4) using fewer than four global judgment categories. We have also reported the standard for BRM and Angoff, as well as the difference between the two.

- Station stability:

A significant number of stations had been reused across multiple sittings, with no meaningful changes to the structure of the station.

We linked BRM standards and R^2 values across diets. Typically, stations fell into one of two categories. For stations with high quality indicators (use of four or more judgments, R^2 typically over 0.5) the BRM standards were consistent over time. These stations had a standard deviation of around 1 mark. This suggests the judgments are reliable, and that the vast majority of students who passed one instance of the station would pass the others.

However, stations which tended to use fewer than four judgments and have low R^2 values behaved very erratically, meaning candidates could realistically pass in one instance and fail in another, even though the station did not change.

- Bootstrapping:

Bootstrap datasets of 20, 150, and 1,000 were used to simulate small (equivalent to the real sittings in this study), medium (a big UK medical school), and large (unrealistic but statistically desirable) cohorts.

The results confirmed that the error rates are much higher for small datasets and where station quality is low, this produces very divergent model fit statistics (and so BRM standards).

By error rates we refer to general problems of misfit. Where error rates are high, the borderline regression model produces implausible valuables (such as extremely high or low passing rates, or has so few scores in some categories that the uncertainty around the predicted values is very high).

This means that for such stations students may pass one instance, yet fail another instance, despite the stations being virtually identical. However, this was less of an issue for high quality stations.

Discussion:

BRM is typically a valid method of standard setting OSCEs in this dataset. Indeed, when the overall item statistics are acceptable and the number of stations with major issues low, the standards can be very close – within 5 marks of Angoff methods. Two factors explain the sometimes-large gaps between the two standard setting methods. Firstly, a combination of poor use of the grading scale and a low overall score can give an extremely harsh or generous BRM standard. This is usually due to the use of only three of the five global categories. Monitoring such stations can address these issues.

Secondly, there is a tendency for some stations to exhibit quite high non-linearity, which is often an issue with examiner training, experience, and happiness with the marking scheme. In a few stations, candidates received an unsatisfactory global judgment, but a high score using the checklist; this complicates the standard setting process. This issue can be solved in future diets with more focused exam training in the use of global judgement and also with examiner feedback post assessment; as suggested by Wong et al, providing structured feedback to examiners might reduce the harshness and leniency of their results[16].

Station stability scores across multiple cohorts is encouraging as they argue for the long-term stability of station standards even for small cohorts. However, they also re-emphasise the problem of poor station quality: if station quality is low, standards will be very unstable, and judgments will not be defensible. Therefore it is critical that a psychometric analysis is carried out after each OSCE assessment to identify low quality stations and removed them or adjust them routinely before re-suing the station in another cohort.

These results confirm the previous work done by Homer et al[15] where they concluded that cohorts of less than 50 candidates have high standard error of measurer and other metrics such as R^2 start to become unstable. However our study adds to the evidence[12] that a proactive strategy of station curation and monitoring would be

sufficient to make BRM an acceptable replacement for Angoff standard setting in small cohorts. However, if even a few stations have very poor metrics the BRM scores will be seriously affected. The decision to remove stations needs to have careful consideration as this will influence the blueprint of that assessment and the overall course, and therefore reverting to Angoff methods might then be required.

Conclusion:

This study suggests that BRM is an acceptable replacement for modified Angoff standard setting method in small cohorts where there is a range of candidates undertaking the assessment (this might not be possible with re-sitting assessments), although if there are a few number of stations with poor metrics, the BRM scores will be seriously affected.

There is a need for the institutions to properly assess their stations and their assessors, before embarking into using this method, to make sure they are robust enough to use this method, to prevent from having to remove stations following from an OSCE assessment, which would compromise the blueprint of the examination.

In order to mitigate these problems, institutions might consider having more stations in each OSCE assessment, a readiness to remove poorly functioning stations (bearing in mind this could compromise the assessment blueprint) or to revert to using alternative standard setting methods when necessary.

Conflict of Interest: The Authors declare no conflict of interest.

Data availability:

The data that support the findings of this study are available from the corresponding author, RML, upon reasonable request.

Bibliography:

(1) 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):162-171.

(2) McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach 2014;36(2):97-110.

(3) Cusimano M. Standard setting in medical education. Acad.Med. 1996;71(Suppl 10):S112-S120.

(4) Norcini J. Research on standards for professional licensure and certifications examinations. Evaluation & the Health Professions 1994;17:160-177.

(5) Kramer A, Muijtjens A, Fansen K, Dusman H, Tan L, Van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Medical education 2003;37:132-139.

(6) Horner M, Darling J. Setting standards in knowledge assessments: comparing Ebel and Cohen via Rasch. Medical teacher 2016;38(12):1267-1277.

(7) Friedman ben-David M. AMEE Guide No 18: Standard setting in student assessment. Medical teacher 2000;22(2):120-130.

(8) 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. Advances in Health Sciences Education 2006;11:173-183.

(9) Tavakol M, Dennick R. Post-examination interpretation of objective test data: monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Medical teacher 2012;34:e161-175.

(10) Wood T, Humphrey-Murto S, Normal G. Standard setting is a small scale OSCE: a comparison of the modified borderline-group method and the borderline regression method. Advances in Health Sciences Education 2006;11:115-122.

(11) Dwyer T, Wright S, Kulasegaram KM, Theodoropoulos J, Chahal J, Wasserstein D, et al. How to set the bar in competency-based medical education: Standard setting after an Objective Structured Clinical Examination (OSCE). BMC Med Educ 2016;16(1).

(12) Homer M, Fuller R, Hallam J, Pell G. Setting defensible standards in small cohort OSCEs: Understanding better when borderline regression can 'work'. Medical teacher 2020;42(3):306-315.

(13) 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;32:802-821.

(14) Mortaz-Hejri S, Jalili M, Muijtjens A, Van der Vleuten C. Assessing reliability of the borderline regression method as a standard setting procedure for objective structures clinical examination. Journal of Research in Medical Sciences 2013;18(10):887-891.

(15) Homer M, Pell G, Fuller R, Patterson J. Quantifying error in OSCE standard setting for varying cohort sizes: a resampling approach to measuring assessment quality. Medical teacher 2016;38(2):181-188.

(16) Wong W, Roberts C, Thistlethawaite J. Impact of Structured Feedback on Examiner Judgements inObjective Structured Clinical Examinations (OSCEs) UsingGeneralisability Theory. Health Professions Education 2020;6:271-281.

Figure 1: Schematic representation of the bootstrapping process.

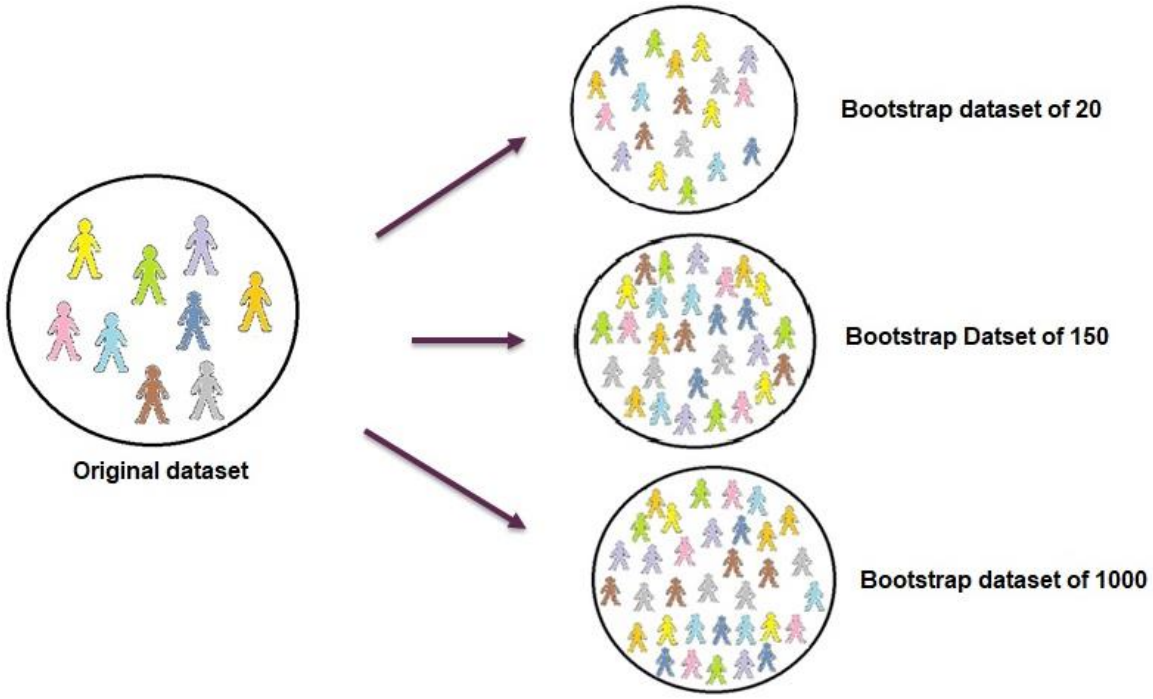

Table 1: Summary of BRM and Angoff

Assessment	Alpha	Stations to remove	Stations $R^2 < .5$	Nonlinear stations	Stations < 4 judgments	BRM standard	Angoff standard	Difference
BDS2 - 2017	.56	0	5	0	4	101.44	105.1	-3.66
BDS3 - 2017	.64	0	6	0	5	166.92	178.0	-11.08
BDS4 - 2017	.67	0	4	1	4	120.53	123.5	-2.97
BDS2 - 2018	.82	0	3	0	3	132.78	134.6	-1.82
BDS3 -2018	.45	0	4	1	4	NA* (148.09)	160.0 (146.62)	NA (1.47)
BDS4 - 2018	.47	1	6	0	0	142.33	141.6	0.73
BDS2 - 2019	.80	0	5	3	4	121.16	123.5	-2.34
BDS3 - 2019	.75	0	3	0	2	136.76	160.5	-23.74
BDS4 - 2019	.005	2	7	0	6	136.62**	170**	-33.38

*Note: One station failed to produce any kind of regression line, but there is no obvious reason why this should be. This means a pass score cannot be calculated. The values in brackets are with stations 5 and 7 removed from the analysis and Angoff standard.

**A station was marked without a global judgment: this has been removed from the Angoff