Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: www.tandfonline.com/journals/imte20

Two-point regression standard setting, improving discrimination at both ends of the grading scale in OSCEs

Andrew M. Lunn, Christopher J. Harrison & John C. McLachlan

To cite this article: Andrew M. Lunn, Christopher J. Harrison & John C. McLachlan (25 Aug 2026): Two-point regression standard setting, improving discrimination at both ends of the grading scale in OSCEs, Medical Teacher, DOI: 10.1080/0142159X.2026.2721367

To link to this article: https://doi.org/10.1080/0142159X.2026.2721367

Published online: 25 Aug 2026.

MEDICAL TEACHER
https://doi.org/10.1080/0142159X.2026.2721367

MEDICAL TEACHER

Taylor & Francis by informa

OPEN ACCESS Check for updates

Two-point regression standard setting, improving discrimination at both ends of the grading scale in OSCEs

Andrew M. Lunn, Christopher J. Harrison and John C. McLachlan

School of Medicine and Dentistry, The University of Lancashire, Preston, UK

ABSTRACT

Background: Standard setting in Objective Structured Clinical Examinations (OSCEs) typically identifies a single cut score to distinguish competent from non-competent candidates. However, this approach does not help defensible differentiation between 'competent' and 'excellent' students.

Objective: This study introduces a Two-Point Regression model, an extension of the borderline regression method, to establish two cut scores, one for competence and one for excellence, enabling more nuanced classification of student performance.

Methods: A retrospective analysis was performed on 10 OSCEs delivered to UK medical students. For each station, linear regressions of global ratings (0–3) against checklist scores were used to generate the traditional pass cut score (cX = 1). Additional thresholds for excellence were generated at global thresholds eX = 2, 2.25, 2.5 and 3. Outcomes included: proportion graded as excellent (≥70% after scaling), mean scaled grades, and mean global scores of 'excellent' students. Statistical analyses included chi-square, ANOVA, and post-hoc testing as appropriate.

Results: Pass/fail classifications were unchanged across all methods. The Two-Point regression model substantially reduced the proportion of students achieving excellence from 47% to 17.2% when the excellence cut score was set to a global threshold of eX = 2.25, aligning with national benchmarks. Mean global scores of 'excellent' students increased in line with threshold choice, demonstrating strong internal consistency (R^2=0.9939).

Conclusion: The Two-Point Regression method offers an easily implementable and defensible approach for identifying excellence in numerically graded OSCEs, improving alignment between examiner global judgements and awarded grades. Threshold choice should be context-dependent, but the Two-Point Regression method provides a robust and adaptable framework.

ARTICLE HISTORY
Received 30 June 2026
Accepted 15 August 2026

KEYWORDS
Standard setting; OSCE; assessment; borderline regression

Introduction

Assessments such as written exams (e.g. MCQ) and Objective Structured Clinical Examinations (OSCEs) are widely used in higher education to determine the knowledge and skill of students. They may differentiate students in a variety of manners such as competent/incompetent, pass/fail as well as giving a numerical score based on performance, which may be scaled to a numerical grade equivalent. Within the UK Higher Education (HE) context it is common for the numerical grade equivalent for passing a Level 7 course to be set at 50% [1,2]. However, the individual 'cut score' per assessment may change relative to the perceived difficulty of an assessment *via* standard setting, which determines the minimum acceptable level or score a candidate must achieve to pass an assessment, making decisions on grades more defensible [3–5]. The individual cut score is then standardised to the required numerical grade equivalent of the institution.

Practice points

- Current borderline regression methods for standard setting an OSCE may over award excellent grades.
- An extended Two-Point Regression method, which adds a second cut score for excellence can control the number of excellent grades received.
- The Two-Point Regression method may improve defensibility of grades awarded across the full scale of marks.

Standard setting helps to ensure fairness and consistency between cohorts of students and is of

CONTACT Andrew M. Lunn Alunn1@Lancashire.ac.uk School of Medicine and Dentistry, The University of Lancashire, Preston, PR1 2HE.

particular use in improving the defensibility of high stakes exams, such as those taken on professional courses. The same process is however rarely followed when setting boundaries for higher grade classifications [3]. This is of particular importance where large numbers of students may receive high scores. Due to the nature of the skills being tested in an OSCE, the average raw percentage achieved is often high, with published examples showing averages as ranging from 50% to as high as around 80% in individual stations [6,7]. This can result in a compressed window between the cut score and excellence boundary, diminishing the opportunity to differentiate between competence and excellence, making differentiation between 'competent' and 'excellent' students difficult and asking questions of the defensibility of numerical grades, particularly those at the top end of the scale. This suggests the possibility of standard setting a second boundary for excellence, dividing the cohort into three groups rather than two to aid in grading. A competent student may be defined as being clearly competent to perform the main task of the station, whereas an excellent student would additionally perform fluently and at a high level of skill, with an appropriate level of confidence.

Various standard setting methods exist, which may be more suited to different assessment types. Widely used methods include Angoff, Ebel and Hofstee, with criterion referenced approaches typically being favoured in the UK [3]. A now well-established methodology for standard setting Objective Structured Clinical Examinations (OSCEs) is the borderline regression model (BLR) [8–10]. The method is similar to, but distinct from the borderline group method and utilizes a linear regression model such that all examinees are included (rather than a sub-set), reducing potential bias, particularly in smaller cohorts [6,9,10]. Following the typical implementation of the BLR model, examiners will score students' performance in a station, based on a mark scheme and then additionally give an overall judgement on students' ability in a global rating. The global rating scales vary, but will include at least one fail, borderline and pass category. Following the OSCE, student performance in a station is plotted as global score on the X-axis and mark scheme score on the Y-axis. A linear regression is fitted to the resulting graph, the formula for that line is then resolved for the borderline category, giving the cut score for that station, any student who achieves it is deemed to have passed that station. The sum of the station cut scores determines the overall cut score for passing an OSCE [11]. This may be used alone or together with a conjunctive condition such as a minimum number of stations passed [9,11,12]. For a BLR method to be used a minimum number of students must take the assessment, variably quoted as 50–70 students, to ensure a sufficient range of mark scheme and global judgement scores have been used [13].

Where a standard set cut score is used it may be necessary to transform a raw mark (the basic numerical score achieved by a student) into a grade. Grades may be categorical, such as pass/fail, or 'scaled' to a numerical value in line with institutional regulations. In this process the standard set cut score is transformed to match the passing grade as set out in local academic regulations, with all other scores being transformed similarly following a distribution [14]. This is achieved with accepted formulas, formally set out by McLachlan et al. in 2023 [14]. It is critical to emphasise that, whilst the numerical value changes, the relative rank of students and binary pass/fail outcomes are unaffected. This process can however impact on the award of higher classification boundaries such as 1st class and honours, which typically have fixed thresholds and therefore do not undergo the same scrutiny of standard as the pass mark when standard setting. In the UK HE setting, a final numerical grade of $\geq$70% is often viewed as the threshold for a 1st class degree (or honours for medicine) [1].

There are conflicting views on the grading of OSCEs. They may be graded as pass/fail to determine competence alone; or students may receive numerical grades as a scaled percentage where more than simple competence may be wished to be determined. As a competency based exam, measuring pass/fail competency as an outcome is fundamental to the assessment, and examiner variability is widely noted [15,16]. However, it remains commonplace that OSCEs are graded numerically and contribute to overall degree performance. Given that OSCEs examine the skills that graduates will practice, it is compelling that their final degree performance is linked to them. Regardless of requiring a binary pass/fail or numerical grade equivalent, standard setting methodologies such as BLR may be used. However, when numerically grading OSCEs (rather than a binary pass/fail) only using standard setting to set a cut score for competence poses issues in the higher score classifications.

Borderline regression has to date only been used to divide students' performance in OSCE examination into two categories (pass/fail). It is accepted that once the regression model has been produced using all examinees, the 'borderline' category determines where the cut score should go, however, defining where the borderline point is presents more challenges [17]. The training and practices of examiners has a major influence on where the cut score will end up and relies on consistent use of all scoring categories to give a reliable regression analysis [10,18]. Many standard setting methodologies can however be used to divide students into more than two groups. For many methods such as Angoff and Ebel, this would need to be set by a panel [3,4,19,20]. It is feasible however to modify the use of the borderline regression technique to set multiple boundaries using the same linear regression produced to aid in the differentiation between both pass/fail and pass/excellent categories.

Rationale

It is with the motivation of providing more defensible grading and differentiation between 'competent' and 'excellent' students that a Two-Point Regression (TPR) model was developed, building on the borderline regression model (BLR). In addition to the standard BLR cut score a second threshold for excellence will be set. The two cut scores would be used to divide students into three groups: fail, pass, excellent, then used as pinpoints when scaling to university academic regulations and determining numerical grades.

Research questions

The following research questions were set.

Primary

1. Does the Two-Point Regression (TPR) model produce a defensible second cut score, for excellence?

Secondary

1. What is a reasonable level to set the second cut score at? (Context dependant)
2. Does the second cut score for excellence reliably distinguish between students based on their performance?
3. Does the second cut score give an acceptable proportion of top grades?

Methods

Study design

This study retrospectively applied a modified two-point borderline regression standard setting model to 10 OSCEs undertaken over a period of four years by medical students at a UK university. This methodology was chosen to maximise the sample size as the proposed methodology can be applied to existing BLR data retrospectively

We propose a method where, following an OSCE, student scores for each global judgement are plotted and a linear regression fitted for each station. The equation for which, when solved for the borderline point would give the pass mark, as is standard for BLR. The same equation would then be solved for the excellence threshold, producing a second cut score for excellence. The pass mark would then be used to determine who had passed the station and the sum of the pass marks determining the overall OSCE cut score as normal. The sum of the second cut scores (for excellence) would then determine the overall OSCE cut score for receiving an 'excellent' grade, which together with previously published scaling formulas would be used to scale students' raw marks to institutional standard boundaries (e.g. 50% pass and 70% excellence in a UK context). Variations would be required depending on the specific running of OSCEs and academic regulations (such as cut-off grade for excellence) at individual institutions. We present the method that would work at this institution, highlighting where variation would be required. The formulas used are easily modified for different contexts, where excellence may be considered at 90% for example.

The boundary for passing was calculated using a typical borderline regression, which in this 4-point model (0–3) is set at 1 where 0 = fail, 1 = borderline, 2 = competent and 3 = exemplary. The boundary for excellence has not previously been defined for standard setting. There is however evidence that it is not a simple linear relationship, precluding a simple equal split of the global scale [17]. Therefore, a boundary based on examiner global rating where a reasonable proportion of students have performed significantly better than 'competent' was sought. To gain insight into a reasonable boundary in this context a number of options were calculated setting the boundary for excellence on this 4-point global judgement scale (0–3) at eX= 2, 2.25, 2.5 and 3 these figures represented realistic points between 'competent' and full marks where excellence may be defined. They also include a lower bound of competence (eX = 2) and what is already defined as 'exemplary' in the global scale, which could be viewed as synonymous with excellence (eX = 3). All examiners were taught to discriminate between candidates and calibrate their scoring during 3-hour face-to-face interactive training sessions before becoming an OSCE examiner. They are also required to attend refresher training at least every 3 years. This is reinforced for every OSCE by a 45-minute calibration session with examiners marking the same station on the day of the OSCE. This enables tailoring of generic guidance to the specific station and year group being assessed.

For global judgements, the following word pictures were used:

- **Exemplary.** This candidate is clearly competent to perform the main task of the station and is performing fluently and at a high level of skill, with an appropriate level of confidence.

- **Competent**. This candidate is competent to perform the main task of the station.
- **Borderline**. You have genuine uncertainty about this performance. On balance, this candidate is probably just able to perform the main task of the station.
- **Fail**. This candidate is not yet able to perform the main task of the station.

For scoring domains, the following word pictures were used as guidance:

- **Excellent for year of study**. Covers the areas in this domain very effectively, fluently and with appropriate confidence for the year of study.
- **Satisfactory for year of study**. Covers the main areas in this domain effectively, but with a little less confidence or fluency. Any deficiencies are very minor and do not prevent the task being achieved.
- **Borderline for year of study**. Patchy or hesitant performance in this domain. Demonstrates inadequacies in more than one area in this domain. No serious defects.
- **Unsatisfactory for year of study**. Demonstrates inadequacies in several important aspects of this domain, some of which are serious or detrimental to patient care.

Step-by-step process followed

- cX = where on the global scale the boundary for the cut score is set
- eX = where on the global scale the boundary for an excellent score is set
  1. An OSCE examination is performed
  2. Examiners document candidate scores and global judgements on performance
     a. In this context domain-based marking was used with 5–7 domains per station
  3. For each station candidate's global judgements (x) are plotted against their scores (y)
     a. Global judgement Fail = 0, Borderline = 1, Competent = 2 and Exemplary = 3
  4. A linear regression is fitted and the equation determined
  5. The equation is solved for y, where cX = 1, for the borderline point. This determines the cut score for passing a station. (The border line point for passing may be different depending on how the global judgements were taken).
  6. The equation is then solved at a point where students would be deemed to have per formed excellently on the station. Within this 4-point global judgement we have solved this equation at eX = 2, 2.25, 2.5 and 3 to explore the most suitable point
     a. The excellence point will differ depending on the context
  7. The cumulative cut (cX) and excellence (eX) scores across all stations determine the over all OSCE cut scores
  8. If students' overall mark is:
     a. Below the cut score, they **fail the OSCE**
     b. Above the cut score but below the score for excellence they **pass the OSCE (subject to conjunctive conditions)**
     c. Above the score for excellence, they **pass the OSCE with excellence (subject to conjunctive conditions)**
        i. An example conjunctive condition may be that the student must also pass a minimum number of stations e.g. pass 9 out of 12 stations in addition to achieving the overall cut score
  9. The raw scores are then scaled to numerical grade equivalents [14].
     a. Cut scores may map differently according to local guidance, for example the numerical grade equivalent for excellence may be 70%, 75% or even greater. This is easily incorporated into the formulas used.

Data analysis

The process outlined above was conducted for ten medical OSCEs. For each OSCE a standard 1-point BLR was conducted and also four variations of the 2-point method, setting the global judgement required for excellence to different points (eX = 2, 2.25, 2.5 and 3).

For each method summary statistics and the following were determined: percentage required to achieve an 'excellent' grade, number of students passing and failing (not taking into account conjunctive conditions), number and percentage of students receiving an 'excellent' grade, mean scaled grade for all students and mean global score received by 'excellent' students.

Chi^2 analysis was performed on the number of students receiving an 'excellent' grade. Mean scaled grade and mean global score for 'excellent' students was analysed using ANOVA with post-hoc Tukey test if the data was parametric and Kruskal-Wallis test with post-hoc Dunn's test if non-parametric.

All data curation and analysis was carried out in IBM SPSS Version 31.0.0.0 and Excel for Microsoft Office 365.

Ethics

The study was conducted in accordance with the Helsinki Declaration of 1975, (revised 2008). Ethics approval was obtained from the University of

Lancashire (REF HEALTH 01346, 5/3/26). All data was treated following the Data Protection Act (2018) requirements and/or GDPR 2016.

Results

To test this method, it has been implemented using data from ten OSCEs taken by medical students in their clinical years. All OSCEs were conducted on a minimum of 99 students, with an average of 153 students per OSCE. A total of 145 stations were included with a minimum of 12 stations and maximum of 16 stations in any one OSCE. Cohort year and demographic information have been removed to preserve anonymity.

A summary of results is shown in Table 1, these are graphically represented in Figures 1, 2 and 3. The number of students passing and failing the assessments (not considering conjunctive conditions) remained the same for all methods. The number of students receiving an 'excellent' grade ($\geq$70% after scaling) reduced when using the TPR method and further as the global judgement required for excellence increased.

The average raw percentage thresholds for excellence for the 1-point BLR, and TPR method set at eX= 2, 2.25, 2.5 and 3 were 70.1%, 71.8%, 77.2%, 82.3 and 92.7% respectively. This was commensurate with the mean global score received by students who achieved an 'excellent' grade, which was 2.13, 2.15, 2.28, 2.44 and 2.80 respectively.

The percentage of students receiving an 'excellent' grade and the mean overall scaled grade reduced when using the TPR method and further as the global judgement required for excellence increased.

When the set global score for excellence was plotted against the resulting average global score of 'excellent' students, a strong linear relationship is seen R^2=0.9939, Figure 3.

Data for eX = 3 was excluded from analytical statistical tests as there was only a single student in one OSCE that achieved an excellent grade when the global score for excellence was set to eX = 3.

Chi^2 analysis of the number of students receiving an 'excellent' grade revealed that there was no significant difference between the 1-point BLR and the TPR when the global score for excellence was set to eX= 2 (satisfactory) ($p = 0.552$). It further showed that there was a statistically significant difference between the 1-point BLR and the TPR when eX was set to 2.25 and 2.5 ($p < 0.001$) Table 2.

ANOVA analysis with post-hoc Tukey testing of the mean scaled grade of all students after scaling revealed that there was no significant difference between the 1-point BLR and the TPR when eX was set to 2 (satisfactory) ($p = 0.552$). It further showed that there was a statistically significant difference between the 1-point BLR and the TPR when eX was set to 2.25 and 2.5 ($p < 0.001$). This also revealed a statistically significant difference in score received when eX is set to 2.25 and 2.5 ($p = 0.031$) Table 2.

ANOVA analysis with post-hoc Tukey testing of the mean global score for students receiving an 'excellent' grade revealed that there was no significant difference between the 1-point BLR and the TPR when eX was set to 2 (satisfactory) ($p = 0.84$). It further showed that there was a statistically significant difference between the 1-point BLR and the TPR when eX was set to 2.25 and 2.5 ($p < 0.001$). This also revealed a statistically significant difference in score received when eX is set to 2.25 and 2.5 ($p < 0.001$) Table 2.

Discussion

After being tested with 10 relevant OSCEs the Two-Point Regression (TPR) method has been shown to have potential as a robust, consistent method by which to set a second threshold to standard set which students receive a nominally 'excellent' grade after scaling. This method applies the same principles of setting the cut score to define a second threshold for excellence, potentially providing a defensible method of assigning grades at the upper of the scale.

The results in this study suggest that a typical 1-point BLR method rewards students who are only deemed 'competent' on global rating, with 'excellent' grades as shown by the average global rating of 'excellent' students when using the standard 1-point BLR. Average global judgements were reached by a minimum of 12 examiners and students receiving a grade $\geq$70% were found to be 2.13, which is only just over what is deemed 'competent.' This suggests that average 'competent' students are receiving 'excellent' grades. This assertion of 'competent' students receiving 'excellent' grades is further confirmed when using the TPR method and setting the excellence threshold to 'competent' (eX = 2). This returned a similar average global rating of 2.15 for students who received an 'excellent' grade of $\geq$70%, which was not significantly different to the 1-point BLR (2.13) ($p = 0.84$). In contrast to this, when the excellence threshold was increased to eX= 2.25 and eX= 2.5 using the TPR method, the average global scores for 'excellent' students were appreciably and significantly increased to 2.28 and 2.43 respectively ($p < 0.001$ in these cases). The awarding of 'excellent' grades of $\geq$70% to students with global rating of 2.13, suggesting that they are in the view of the examiners, simply

Table 1. Results for each individual OSCE using a standard 1-point borderline regression and the proposed Two-Point regression solving the line of best fit for four options eX = 2, 2.25, 2.5 and 3.

		Cronbach's alpha	Cut score (%)	Number of students	Number fails on score alone	Number passes	Threshold for excellence (%)	Number excellent students	% of students receiving excellent	Mean scaled grade (%)	Mean global score for excellent students (/3)
OSCE 1	1-BLR	0.79	51	119	9	110	70	42	35.29	66	2.10
	eX = 2						72	38	31.93	65	2.10
	eX = 2.25						77	18	15.13	62	2.16
	eX = 2.5						82	2	1.68	59	2.57
	eX = 3						92	0	0	55	N/A
OSCE 2	1-BLR	0.60	53	114	0	114	71	76	66.67	69	2.11
	eX = 2						73	45	39.47	68	2.13
	eX = 2.25						78	20	17.54	64	2.20
	eX = 2.5						83	3	2.63	62	2.33
	eX = 3						93	0	0	59	N/A
OSCE 3	1-BLR	0.75	51	113	1	112	70	50	44.25	69	2.16
	eX = 2						72	45	39.82	67	2.18
	eX = 2.25						78	24	21.24	64	2.28
	eX = 2.5						83	7	6.19	61	2.40
	eX = 3						94	0	0	58	N/A
OSCE 4	1-BLR	0.77	52	184	3	181	71	94	51.09	73	2.15
	eX = 2						73	82	44.57	71	2.17
	eX = 2.25						78	37	20.11	67	2.32
	eX = 2.5						84	14	7.61	64	2.45
	eX = 3						94	0	0	60	N/A
OSCE 5	1-BLR	0.72	49	172	1	171	69	70	40.70	71	2.19
	eX = 2						70	68	39.53	70	2.19
	eX = 2.25						76	35	20.35	66	2.30
	eX = 2.5						81	13	7.56	64	2.46
	eX = 3						92	0	0	60	N/A
OSCE 6	1-BLR	0.73	51	144	1	143	69	86	59.72	70	2.12
	eX = 2						72	64	44.44	69	2.18
	eX = 2.25						77	29	20.14	65	2.34
	eX = 2.5						82	12	8.33	63	2.46
	eX = 3						93	1	0.69	59	2.8
OSCE 7	1-BLR	0.72	51	204	7	197	70	95	46.57	67	2.14
	eX = 2						72	79	38.73	66	2.16
	eX = 2.25						77	32	15.69	63	2.32
	eX = 2.5						82	6	2.94	61	2.50
	eX = 3						92	0	0	58	N/A
OSCE 8	1-BLR	0.64	51	199	4	195	70	65	32.66	66	2.14
	eX = 2						70	65	32.66	66	2.13
	eX = 2.25						75	22	11.06	63	2.31
	eX = 2.5						80	8	4.02	61	2.44
	eX = 3						89	0	0	58	N/A
OSCE 9	1-BLR	0.77	51	183	10	173	70	88	48.09	67	2.10
	eX = 2						72	79	43.17	66	2.12
	eX = 2.25						78	29	15.85	63	2.24
	eX = 2.5						83	9	4.92	61	2.40
	eX = 3						94	0	0	58	N/A

(*continued*)

Table 1. Continued.

		Cronbach's alpha	Cut score (%)	Number of students	Number fails on score alone	Number passes	Threshold for excellence (%)	Number excellent students	% of students receiving excellent	Mean scaled grade (%)	Mean global score for excellent students (/3)
OSCE 10	1-BLR	0.67	51	99	1	98	71	45	45.45	69	2.10
	eX = 2						72	39	39.39	68	2.13
	eX = 2.25						78	15	15.15	64	2.30
	eX = 2.5						83	4	4.04	61	2.42
	eX = 3						94	0	0	59	N/A

		Number of students	Number of fails	Number passes	Threshold for excellence (%)	Number excellent students	% of students receiving excellent	Mean scaled grade (%)	Mean Global score for excellent students (/3)
1 point BLR	Average	153.1	3.7	149.4	70.1	71.1	47.0	68.7	2.13
	SD				0.7		10.3	2.3	0.03
	Min	99.0	0.0	98.0	69.0	42.0	32.7	66.0	2.10
	Max	204.0	10.0	197.0	71.0	95.0	66.7	73.0	2.19
eX = 2	Average	153.1	3.7	149.4	71.8	60.4	39.4	67.6	2.15
	SD				1.0		4.3	2.0	0.03
	Min	99.0	0.0	98.0	70.0	38.0	31.9	65.0	2.10
	Max	204.0	10.0	197.0	73.0	82.0	44.6	71.0	2.19
eX = 2.25	Average	153.1	3.7	149.4	77.2	26.1	17.2	64.1	2.28
	SD				1.0		3.2	1.5	0.06
	Min	99.0	0.0	98.0	75.0	15.0	11.1	62.0	2.16
	Max	204.0	10.0	197.0	78.0	37.0	21.2	67.0	2.34
eX = 2.5	Average	153.1	3.7	149.4	82.3	7.8	5.0	61.7	2.44
	SD				1.2		2.3	1.6	0.06
	Min	99.0	0.0	98.0	80.0	2.0	1.7	59.0	2.33
	Max	204.0	10.0	197.0	84.0	14.0	8.3	64.0	2.57
eX = 3	Average	153.1	3.7	149.4	92.7	0.1	0.069	58.4	2.80
	SD				1.6		0.9	1.4	N/A
	Min	99.0	0.0	98.0	89.0	0.0	0.0	55.0	N/A
	Max	204.0	10.0	197.0	94.0	1.0	0.69	60.0	N/A

Followed by average results for each standard setting methodology, min = minimum score achieved, max = maximum score achieved.

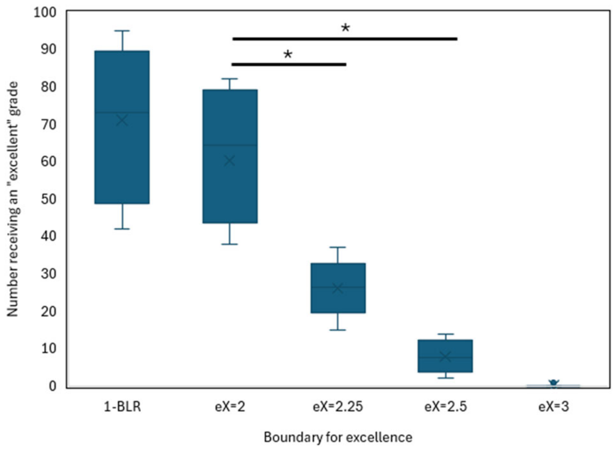

Figure 1. Number of 'excellent' students across all 10 OSCEs when using each standard setting methodology. Where 'excellent' is a grade of $\geq$70% after scaling. $^*p < 0.001$.

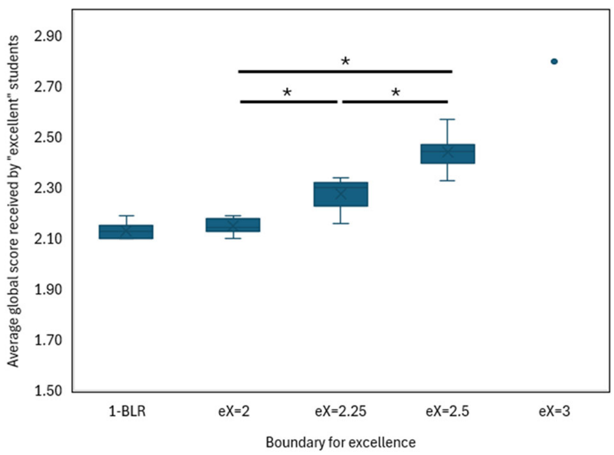

Figure 2. Average global score achieved by 'excellent' students across all 10 OSCEs when using each standard setting methodology. Where 'excellent' is a grade of $\geq$70% after scaling. $^*p < 0.001$.

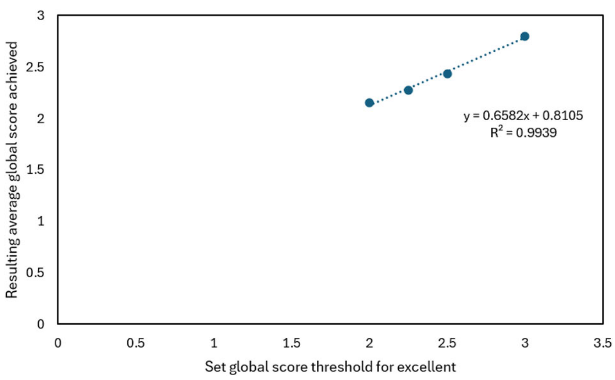

Figure 3. Set global score for excellence plotted against the resulting average global score of 'excellent' students following that method. Where 'excellent' is a grade of $\geq$70% after scaling.

'competent' rather than 'excellent' is arguably not of a suitable standard.

Defining what 'excellent' is poses theoretical and practical challenges similar to those experienced when defining competence. The highest point of the four-point global scale used here is named 'Exemplary,' which is arguably synonymous with excellent. It is potentially reasonable then to simply define this as the point of excellence and set the excellence threshold at eX = 3. As seen in Table 1, this leads to an unreasonably low number of students receiving an excellent grade, with just 1 out of 1,531 candidates from the 10 OSCEs receiving an excellent grade, Suggesting it is an inappropriately high boundary. This further adds to the argument that examiners do not treat a grading scale in a linear fashion, and that the correct setting of the boundary and training of examiners is critically important. The question of defining competence (and thus excellence) is time consuming and context dependant, with translation of such a nebulous idea to a simple number or category conceptually tricky. However, research in the area suggests that expert examiners making such judgements remains preferable to a fixed cut score or threshold [21,22]. This study has therefore focussed on finding a reasonable threshold that is significantly above competence, across multiple OSCEs based on sector norms. Applying the principles of defining minimal competence to excellence would suggest that standard setting with expert judgement may be more defensible than not addressing the boundary at all. What these results show is that the inclusion of a second threshold does make a real and significant difference to the awarding of 'excellent' grades. It can also consistently modify the level of performance required to achieve 'excellence' making available a method to ensure the standard of candidates receiving an 'excellent' grade can be controlled in the same way as for those passing.

The average global score results also show the internal consistency of the TPR method, as the average scores received by 'excellent' students are consistently comparable to the threshold set when determining the second cut score for excellence. This is further supported when a linear regression between the average global score for 'excellent' students and the threshold for global score for excellence is conducted, revealing a strong correlation, R^2=0.9939, Figure 3. This figure is however only fitted to four data points so should be treated with caution, but nonetheless suggests internal consistency.

The proposed TPR method not only changes the average global score of the students receiving an 'excellent' grade, but it predominantly does this by modifying the grades of all students who pass (failing students are unaffected). This leads to changes in the cohort average grade and the proportion of students receiving an 'excellent' grade. The difference in cohort average grade received between the 1-point BLR and TPR where $x = 2$ is minimal, 68.7% and 67.6% respectively ($p = 0.84$). This further suggests that the typical 1-point BLR only requires a student to perform competently to receive an 'excellent' grade. The average

Table 2. Analytical statistics, numbers quoted for each test are P-values.

Number excellent vs not excellent students	Chi^2	eX = 2	eX = 2.25	eX = 2.5
	1-point BLR	0.552	$<$0.001	$<$0.001
Mean scaled grade	**ANOVA + Tukey**			
	Tukey	eX = 2	eX = 2.25	eX = 2.5
	1-point BLR	0.552	$<$0.001	$<$0.001
	eX = 2		$<$0.001	$<$0.001
	eX = 2.25			0.031
	ANOVA + Tukey			
Mean global score for excellent students	**Tukey**	eX = 2	eX = 2.25	eX = 2.5
	1-point BLR	0.84	$<$0.001	$<$0.001
	eX = 2		$<$0.001	$<$0.001
	eX = 2.25			$<$0.001

grade received when x is set to 2.25 reduces to 64.1%, which is significantly different to the 1-point BLR and TRP eX = 2 ($p < 0.001$); and when eX is set to 2.5 the average grade reduces to 61.7%, which is significantly different to the 1-point BLR, TRP eX = 2 and TRP eX = 2.25 ($p < 0.001$). The average grade reduces even further to 58.4% when eX is set to 3. Furthermore, the average percentage of students receiving an 'excellent' grade when using the standard 1-point BLR is 47% (min = 32.7%, max = 66.7%). This percentage is moderately reduced to 39% when using the TPR and setting the threshold to eX = 2 (competent). When eX is set to 2.25 and 2.5 using the TPR, the percentage of 'excellent' students is reduced much further to 17.2% and 5.0% respectively (Table 1). These reductions are far greater and statistically significant ($p < 0.001$ in both cases). Setting eX to 3, reduces the average percent of students receiving an excellent grade to a negligibly small 0.069%.

Setting the threshold for excellence to both 2.25 and 2.5 has been shown to produce sets of 'excellent' students who performed significantly better than 'competent.' What is less clear however is what is an acceptable average percentage of students to be receiving an 'excellent' grade and thus what a reasonable point for the second threshold is. Whilst we do not wish to define a set percentage of students to receive an 'excellent' grade, it is of interest to compare this to national averages. Office for Students (OfS) data on overall degree classification showed a peak in 1st class degrees ($\geq$70%) in 2020-21 at 37.7%, a figure which they deemed too high. This figure is now falling again and the average over the latest published data from 2010 to 2023 is 26.4% [23]. Whilst this is data for an overall degree rather than a single assessment, the figure of 26.4% represents a good estimate for a reasonable maximum proportion of students receiving an 'excellent' grade. This would suggest that a threshold of eX = 2.25 would be reasonable, as it gave an average of 17.2% (max 21.2%, min 11.1%) of students an 'excellent' grade, the threshold would however be context dependant.

This post hoc method allows for manipulation of the percent of students receiving an 'excellent' grade and is open for misuse. It would therefore be important (as with many standard setting methodologies) to set the excellence threshold before the OSCE (e.g. eX = 2.25 or eX = 2.5 etc). In this context an average global score and therefore excellence threshold of eX = 2.25 seems reasonable. It is also statistically significantly above 'competent' but still allows for a reasonable number of students to be rewarded with an 'excellent' grade (17%), which as previously discussed is similar to the UK average of grades of $\geq$70% achieved in degree courses [23]. This is achieved consistently across all OSCEs analysed. In this example we were scaling to an 'excellent' grade being set at 70% as per local university regulations, the formulas used are however easily modified to different academic regulations.

It is important that any proposed method, such as the TPR method here is suitable for use in real world settings, not just idealised ones. As such, a two-point standard set must be achievable using other OSCE relevant methods. As mentioned in the introduction, various widely used standard setting methodologies such as Angoff and Ebel, can already be used to set multiple boundaries, utilising a panel to determine the absolute standard required [3,4,19,20]. The TPR method would be compatible with these methods, however practically the implementation of the method would require long term planning such that the standard setting process of any necessary complementary methods is established simultaneously to set the cut score and boundary for excellence.

Limitations

This study only tested the proposed Two-Point Regression method on medical students in one university. Whilst there will be a range of student abilities in the cohorts tested, this limits generalisability. Furthermore, only one OSCE mark scheme was included, using domain-based marking and a 4-point global judgement scale. To address these limitations future work should focus on repeating the process in different contexts, including different universities, professions and marking schemes. The numbers of passing students in this study does not take into account conjunctive conditions, however the implementation of the TPR method would not affect common conjunctive methods such as minimum number of stations passed or the addition of a standard error of Measurement to the cut score.

Use of the TPR method would be subject to the same requirements of a standard 1-point BLR, requiring a large student cohort, with a good representation of students in each global category.

Conclusions

When using numerical grading in OSCEs a rigorous application of standard setting across the full range of student abilities is required to ensure defensible awarding of exceptional grades. This study suggests that use of a typical 1-point BLR can lead to students who are considered as only 'competent' by global rating from multiple examiners to receive 'excellent' grades. The results suggest that this issue can be addressed with the use of the extended BLR model, where a second threshold for excellence is determined in a Two-Point Regression (TPR). This allows the minimum standard of excellence displayed by a student to be set in the same way it is for minimum competence with a robust, simple, modification to the existing widely used BLR. The TPR method may represent a psychometrically coherent extension of the traditional BLR method, determining more defensible grading of higher scoring students in an OSCE with numerical grading.

Author contributions

CRediT: **Andrew M. Lunn**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – original draft, Writing – review & editing; **Christopher J. Harrison**: Conceptualization, Methodology, Writing – original draft, Writing – review & editing; **John C. McLachlan**: Conceptualization, Writing – original draft, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The author(s) reported there is no funding associated with the work featured in this article.

Notes on contributors

Andrew M. Lunn, School of Medicine and Dentistry, The University of Lancashire, Preston, PR1 2HE.

Christopher J. Harrison, School of Medicine and Dentistry, The University of Lancashire, Preston, PR1 2HE.

John C. McLachlan, School of Medicine and Dentistry, The University of Lancashire, Preston, PR1 2HE.

ORCID

Andrew M. Lunn http://orcid.org/0000-0003-2884-2755
Christopher J. Harrison http://orcid.org/0000-0002-6008-4483
John C. McLachlan http://orcid.org/0000-0001-5493-2645

References

1. Assessment USCfQ. Degree classification: transparent, consistent and fair academic standards UK. Standing Committee for Quality Assessment; 2018.
2. Universities U. Understanding degree algorithms. London Universities UK; 2017.
3. Cizek GJ. Setting performance standards: foundations, methods, and innovations. Routledge; 2012.
4. Muijtjens AM, Schuwirth LW, Cohen-Schotanus J, et al. Benchmarking by cross-institutional comparison of student achievement in a progress test. Med Educ. 2008;42(1):82–88. doi:10.1111/j.1365-2923.2007.02896.x
5. Boulet JR, De Champlain AF, McKinley DW. Setting defensible performance standards on OSCEs and standardized patient examinations. Med Teach. 2003;25(3):245–249. doi:10.1080/0142159031000100274
6. Smee S, Coetzee K, Bartman I, et al. OSCE standard setting: three borderline group methods. Med Sci Educ. 2022;32(6):1439–1445. doi:10.1007/s40670-022-01667-x
7. Khan U. Standard setting OSCE: a comparison of arbitrary and Hofstee methods in a low stake OSCE. TAPS. 2024;9(3):15–21. doi:10.29060/TAPS.2024-9-3/OA3129
8. Hejri SM, Jalili M, Muijtjens AM, et al. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. J Res Med Sci. 2013;18(10):887–891.
9. Kramer A, Muijtjens A, Jansen K, et al. Comparison of a rational and an empirical standard setting procedure for an OSCE. Med Educ. 2003;37(2):132–139. doi:10.1046/j.1365-2923.2003.01429.x
10. 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(2):115–122. doi:10.1007/s10459-005-7853-1
11. McKinley DW, Norcini JJ. How to set standards on performance-based examinations: AMEE Guide No. 85. Med Teach. 2014;36(2):97–110. doi:10.3109/0142159X.2013.853119
12. Homer M, Russell J. Conjunctive standards in OSCEs: the why and the how of number of stations passed criteria. Med Teach. 2021;43(4):448–455. doi:10.1080/0142159X.2020.1856353
13. Homer M, Fuller R, Hallam J, et al. Setting defensible standards in small cohort OSCEs: understanding better when borderline regression can 'work. Med Teach. 2020;42(3):306–315. doi:10.1080/0142159X.2019.1681388

14. McLachlan JC, Davies R, Sawdon M. Converting raw scores to numerical grade equivalents: A data modelling study. Res Sq. 2023; Version 1.
15. Homer M. Pass/fail decisions and standards: the impact of differential examiner stringency on OSCE outcomes. Adv Health Sci Educ Theory Pract. 2022; 27(2):457–473. doi:10.1007/s10459-022-10096-9
16. McManus IC, 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 Med Educ. 2006; 6(1):42. doi:10.1186/1472-6920-6-42
17. McGown PJ, Brown CA, Sebastian A, et al. Is the assumption of equal distances between global assessment categories used in borderline regression valid? BMC Med Educ. 2022;22(1):708. doi:10.1186/s12909-022-03753-5
18. Park SY, Lee S-H, Kim M-J, et al. Comparing the cut score for the borderline group method and borderline regression method with norm-referenced standard setting in an objective structured clinical examination in medical school in Korea. J Educ Eval Health Prof. 2021; 18:25. doi:10.3352/jeehp.2021.18.25
19. Mitzel HC, Lewis DM, Patz RJ, et al. The bookmark procedure: psychological perspectives. Setting performance standards. Routledge; 2013. p. 263–296.
20. Burr SA, Whittle J, Fairclough LC, et al. Modifying Hofstee standard setting for assessments that vary in difficulty, and to determine boundaries for different levels of achievement. BMC Med Educ. 2016;16(1):34. doi:10.1186/s12909-016-0555-y
21. Searle J. Defining competency–the role of standard setting. Med Educ. 2000;34(5):363–366. doi:10.1046/j.1365-2923.2000.00690.x
22. Boursicot K. Setting standards in a professional higher education course: defining the concept of the minimally competent student in performance-based assessment at the level of graduation from medical school. Higher Educ Quart. 2006;60(1):74–90. doi:10.1111/j.1468-2273.2006.00308.x
23. Students of Analysis of degree classifications over time. Changes in graduate attainment from 2010-11 to 2022-23. 2024.