Article:

Homer, M orcid.org/0000-0002-1161-5938, Fuller, R, Hallam, J orcid.org/0000-0002-1044-0515 et al. (1 more author) (2020) Setting defensible standards in small cohort OSCEs: Understanding better when borderline regression can ‘work’. Medical Teacher, 42 (3). pp. 306-315. ISSN 0142-159X

https://doi.org/10.1080/0142159X.2019.1681388

Setting defensible standards in small cohort OSCEs: understanding better when borderline regression can ‘work’

Abstract

Introduction

Borderline regression (BRM) is considered problematic in small cohort OSCEs (e.g. n<50), with institutions often relying on item-centred standard setting approaches which can be resource intensive and lack defensibility in performance tests.

Methods

Through an analysis of post-hoc station- and test-level metrics, we investigate the application of BRM in three different small-cohort OSCE contexts: the exam for international medical graduates wanting to practice in the UK, senior sequential undergraduate exams, and Physician associates exams in a large UK medical school.

Results

We find that BRM provides robust metrics and concomitantly defensible cut scores in the majority of stations (percentage of problematic stations 5%, 14% and 12% respectively across our three contexts). Where problems occur, this is generally due to an insufficiently strong relationship between global grades and checklist scores to be confident in the standard set by BRM in these stations.

Conclusion

This work challenges previous assumptions about the application of BRM in small test cohorts. Where there is sufficient spread of ability, BRM will generally provide defensible standards, assuming careful design of station-level scoring instruments. However, extant station cut-scores are preferred as a substitute where BRM standard setting problems do occur.

Key words

Standard setting; OSCE; small cohorts; borderline regression

Practice Points

- Standard setting is always a challenge, and this is particularly true in small cohort performance tests
- The borderline regression method (BRM) is often thought of as problematic in small cohorts
- This work indicates that across a variety of OSCE small cohort contexts, BRM can produce defensible standards in many stations
- Where problems with BRM do occur this is often due to a weak relationship between global grades and checklist/domain scores, and/or to a lack of spread in scores
- Extant pass marks derived from larger cohort ('main') exams are useful to employ when BRM standards are problematic in small cohorts.

Glossary

Borderline regression method

This is an examinee-centred method of standard setting often used in OSCEs. At the station level, candidates are scored in two ways independently – one score is based on a checklist or set of domain scores, and the other is a global grading of performance (e.g. fail, borderline, pass, good grade). Scores are regressed on grades, and the cut-score in the station is set at the checklist/domain score corresponding to the borderline grade (Pell, 2010). The overall test cut-score is set at the aggregate of the stations cut-scores. One advantage the borderline regression method is that it uses all scores from the assessment (e.g. not just those at the borderline), and that these scores are based on judgment of the actual performance of candidates – compare with item-centred standard setting methods (e.g. Angoff) where item difficulty is judged in advance of the administration of the assessment.

Pell, G., Fuller, R., Homer, M., and Roberts, T., 2010. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Medical Teacher*, 32 (10), 802–811.

Introduction

Standard setting, particularly in high stakes performance assessments, is always challenging (Cusimano 1996, Ben-David 2000, Cizek 2012). The borderline regression method (BRM) is an examinee-centred approach to setting standards (Livingston and Zieky 1982), where candidate performance in stations (or cases) is scored in two different ways: holistically by a global grade, and also with a checklist or domain-based scoring instrument (Kramer *et al.* 2003, Pell *et al.* 2010, McKinley and Norcini 2014). Under BRM, the latter score is regressed on the global grade, and the station-level standard is set *post hoc* using the regression model predicted score corresponding to the borderline grade. The exam-level standard is then based on the aggregate of the station level cut-scores, with the option to employ additional conjunctive level standards such as the use of the standard error of measurement (Hays *et al.* 2008).

Across a broad set of contexts, BRM is now generally acknowledged as providing defensible standards and has become the default approach in many high stakes performance assessment contexts (Boursicot *et al.* 2007, McKinley and Norcini 2014). By using all interactions between assessors and candidates to set the cut-score rather than just the borderline group, BRM brings additional benefits in comparison to other borderline methods. For example, having stations scored in two different ways means that under BRM there are a range of additional station- and exam-level metrics which give detailed insight into assessment quality (Pell *et al.* 2010). These metrics also allow for the impact of interventions aimed at improving assessments to be measured longitudinally (Fuller *et al.* 2013).

The majority of the BRM literature is generated from assessments with relatively ‘large’ cohorts (i.e. n>50) where the candidate group is typically high-performing (i.e. a single year group at medical school, or a cohort of post-graduate candidates). For BRM to function effectively, the range of checklist marks and global grades should be sufficient to provide a comparatively stable estimate of the cut-score, and so relatively large sample sizes help to ensure sufficient spread in candidate ability to support the application of BRM, despite the cohort as a whole being high-performing.

Most of the literature validating BRM has used a range of empirical approaches to estimate the error in the cut-score it produces – either through resampling approaches (Muijtjens *et al.* 2003, Homer *et al.* 2016) or via regression-based formulae (Kramer *et al.* 2003, Wood *et al.* 2006, Hejri *et al.* 2013). Estimating this error is generally considered as an important constituent of validity evidence linked to a particular standard setting approach (American Educational Research Association 2014, p. 108). The evidence suggests that in comparison with other standard setting approaches, BRM has lower error at modest candidate sample sizes (n> 50). The resampling-based work has also attempted to use data from larger cohorts to extrapolate cut-score error for smaller samples (n<50), and indicates that the estimated error in the cut-score becomes quite large at cohort sizes below 50 candidates (Homer *et al.* 2016).

Faced with these standard setting challenges, institutions with small cohorts have generally relied on test-centred approaches such as Angoff-type methods where checklist item or, more commonly, station-level difficulty is judged *a priori* by a group of experts (McKinley and Norcini 2014). These methods can be time-consuming and resource intensive, and may not be particularly reliable in themselves - the difficulty of conceptualising the 'just passing' candidate, and then articulating a passing standard for a series of interdependent (and unobserved) activities within a complex OSCE station can prove difficult in practice (Boulet *et al.* 2003). Most of the evidence exploring the problems with test-centred standard setting methods has been developed in the context of knowledge testing (Clauser *et al.* 2009, Margolis *et al.* 2016). It is difficult to imagine that judging what scores would reflect the minimally competent performance in an OSCE station based merely on knowledge of the station content, rather than how candidates actually perform on the day – the 'reality check' (Livingston and Zieky 1982), would not encounter similar problems. There appears to be contradictory evidence in the literature on this specific point with some agreeing that Angoff-type approached perform less well than BRM (Schoonheim-Klein *et al.* 2009). However, more recent work argues that Angoff-type approaches can work to an extent in OSCEs (Dwyer *et al.* 2016) although interpretation of the findings in this particular paper are complicated by an attempt in it to set two standards, for junior and senior residents, in the same examination. Despite this newer evidence, the resource intensive nature of Angoff, and the doubts about its efficacy when employed in a range of high stakes assessment formats, remain.

Investigating the challenge of appropriate, defensible standard setting in small cohorts

In this paper we investigate the use of BRM in a range of different small cohort contexts in order to develop a more contextualised evidence-base regarding the conditions under which BRM in small cohorts with trained clinical assessors might provide defensible standards. We are particularly interested in better understanding the issues (and their prevalence) that arise when using this approach to standard setting in small cohorts. As our main metric of station-level quality, we use R-squared, the measure of the strength of linear association between the global grade and the checklist-score (Pell *et al.* 2010). Formally, this quantifies the proportion of shared variance between the two scores, with high values (e.g. 0.8) indicating a strong association and providing some evidence of validity in the scoring and standard setting under BRM, whereas low values (e.g. less than 0.4 or 0.5) potentially signify problems in the station ((Pell *et al.* 2010). We also employ visual inspection of scatter graphs to assess the degree of spread in scores – we give more details of our approach in methods section.

This paper focuses on standard setting at station level, and does not explore wider examination issues related to the use of conjunctive standards at the test level (e.g. minimum stations passed or standard errors of measurement) (Cizek and Bunch 2007, chap. 2, Hays *et al.* 2008). The use of these conjunctive standards in small cohort performance testing merits separate investigation.

Assessment contexts with small cohorts

We investigate the use of BRM in three quite different high stakes assessment contexts:

1. The OSCE for international medical graduates seeking professional registration to practise medicine in the UK

This examination is administered by the General Medical Council (GMC) in the UK, and is part of a sequence of knowledge and performance testing referred to as PLAB – Professional and Linguistic Assessment Board test (General Medical Council 2019). The OSCE component (PLAB2) is designed to cover all aspects of clinical practice a UK-trained doctor might expect see during their first day of their second year of medical practice following graduation from medical school and completion of the first Foundation Year of postgraduate training. The examination consists of 18 stations which are each scored by clinically trained assessors via a holistic judgement of the performance in a four point global grade (0=unsatisfactory, 1=borderline, 2=satisfactory, 3=good). Candidates are also scored in three separate domains (*Data gathering, technical and assessment skills*, *Clinical management skills*, and *Interpersonal skills*). Each domain is scored on a 4-point scale and these are aggregated to a total station score out of 12.

Each PLAB2 administration consists of a morning and afternoon circuit, usually with the same assessors in each station, with all assessment outcomes for the 30-35 candidates combined for the BRM standard setting for the day – in other words, each administration is treated independently of any other in terms of standard setting, and only data from the day in question is used to calculate cut-scores under BRM. In order to sit PLAB2, candidates have to pass the PLAB1 applied knowledge test. There are of the order of 100 administrations of PLAB2 per year, and the station level-data used in this paper consists of 198 test administrations over September 2016 to October 2018. The large volume of PLAB2 administrations essentially necessitates standard setting to be examinee-centred (i.e. *post hoc*), and BRM has been used since 2016.

PLAB2 stations in this study were drawn from a bank of 264 stations in total, and an appropriate blueprinting process was carried out for each of the 198 administrations to select each set of 18 stations in the exam. As a consequence, the frequency of use of any individual station varies in the data, ranging from 1 to 48 with a median of 11 over the period. Station-level data consists of a range of station and test-level metrics (Pell *et al.* 2010) such as the cut-score, R-Squared, reliability coefficient-station-deleted (reliability of the overall exam with station removed), and station 'facility' (i.e. station pass rate for each administration). In addition, scatter graphs of global grades versus total domain scores for each station in each administration are also available for visual inspection.

2. The second part of a sequential OSCE for undergraduate medical students in a UK medical school

As part of more innovative approaches to assessment, the development of adaptive approaches to testing has seen the introduction of sequential testing models, where assessment is delivered in two parts. An initial screening OSCE for all candidates, with a further sequence for weaker candidates provides both an adaptive test format, and overall enhanced decision-making (Pell *et al.* 2013, Homer *et al.* 2018). This part of the study draws on work from Year 4 and Year 5 (qualifying) OSCEs, where the full cohort of approximately 300 students take the initial screening sequence, with 20-50 students being recalled to sit the second sequence OSCE. Pass/fail decisions for this smaller cohort of candidates are made based on performance across ***both*** sequences (26 and 25 stations in total in Years 4 and 5 respectively).

Stations are scored by clinical assessors using a key features checklist (Farmer and Page 2005) and a global grade on a five point scale – 0=fail, 1=borderline, 2=pass, 3=good pass, 4=excellent pass. We consider candidate-level data from six sequence 2 administrations (2017-2019 inclusive). Stations selected for use in sequence 2 are those that have been used in a previous sequence 1 administration, and so have pre-existing passing scores generated from the satisfactory use of BRM in the full cohort (i.e. with sufficiently good metrics such as R-squared).

Standard setting is undertaken by BRM or substitution with previous pass marks if there is concern, for any reason, with the BRM standards in the new administration. Part of our research is to quantify how often this substitution might prove necessary. For the first sequence, post-hoc analysis provides a sophisticated range of test- and station level metrics, including measures of assessor disparity (Pell *et al.* 2010, 2015). However, for the second sequence, such analysis is necessarily limited, given the atypical nature of this relatively small sub-group of candidates. The focus in this sequence 2 analysis is on scatter graphs of grades versus checklist scores and associated metrics (e.g. R squared).

3. OSCEs for Physician associates in a UK medical school

A physician associate (PA, 'physician assistant' in the US) is a relatively new healthcare professional in the UK. The training programme typically consists of a graduate entry programme which provides a two year university Masters-level qualification, following which new PAs practice as part of a team alongside fully qualified doctors (Health Education England 2015). The University of Leeds PA programme is assessed through a range of knowledge, performance and professional assessments, including end of year high stakes OSCEs. A cohort of typically 20-30 students are assessed using a 'traditional single test OSCE', where stations are scored by clinical assessors via a five-point global grade, and a key features checklist (as in context 2).

In this study we use PA candidate-level data from four administrations (Y1 and Y2 in 2018 and 2019). Standards are usually set using a modified Angoff method at the station level (McKinley and Norcini 2014), and part of our research explores the extent to which BRM can be used as a replacement or, at least, as the default approach to standard setting in these exams. Consequent to the application of Angoff standard setting, a limited range of post hoc analysis quality data is available (e.g. station level facility, R-squared, scatter graph inspection).

Methods

Common methods across all three candidate contexts

Using a range of station level data as detailed above, we investigate station- and test-level metrics, and examine the relationship between global grades and total key feature/domain scores in stations (Pell *et al.* 2010). For stations with 'low' R-Squared value (e.g. below 0.4) (Pell *et al.* 2010), we also assess the extent to which grades and scores show sufficient variation within each station. Our overall approach is to assess whether there is evidence that BRM is 'working' at the station level – based on a sufficiently satisfactory positive relationship between checklist/domain scores and global grades, and an adequate spread of grades/scores within each station.

Methods specific to PLAB2 exams

For the PLAB2 data, in addition to the common analyses outlined above, we also use simple descriptive and correlational methods at the station level (n=3645) to probe the relationships between cut-scores, facility (pass rate) and R-squared values, and to measure the extent to which cut scores vary for the same station across multiple administrations.

Methods specific to sequential exams

For the sequential context, we also compare the BRM-set standards with those generated from main cohorts in a previous administration in order to assess the consistency of the standards across these quite different cohorts of students. In contrast to the other two contexts considered in this paper, the students sitting the sequence two examinations are by definition an 'extreme sub-group' in the sense that they are not representative of the full cohort, having failed to perform sufficiently strongly in the first sequence to 'pass' based on this alone. This has implications for the appropriate interpretation of the usual range of metrics used for assuring station and exam quality, akin to a traditional 'resit OSCE' (Pell *et al.* 2010). Typical measures of reliability are not usually appropriate as scores are likely to have a limited range which lowers correlation between scores (Bland and Altman 2011), and station failure rates will be expected to be high compared to when used in the full cohort.

Methods specific to the PA exams

For the PA exam we also compare station-level and overall BRM standards with those from the modified Angoff approach that is currently employed to provide the actual standard. This comparative work allows for the possibility of providing additional validity evidence for the BRM-set standards. Note that for the other two contexts, PLAB2 and Sequential, Angoff judgements are not available so this additional analysis is not possible. We also calculate the standard error of the overall pass mark in each exam using a resampling approach (Homer *et al.* 2016).

Results

We take each context in turn and summarise the key analyses we have carried out in each to assess the evidence that BRM is providing defensible standards. We first give the overall reliability of the exams, and then move on to the BRM-specific analysis.

1. PLAB2 exams

The reliability of these 198 18-station examinations, as measured by Cronbach's alpha, is generally good with the 5th, 50th and 95th percentiles of the distribution of 198 alpha values being 0.64, 0.79, and 0.87 respectively.

The strength of the relationship between global grades and domain scores (R-squared)

For our main analysis, we first look at the distribution of R-squared values across the 198 administrations. Across the 3,564 stations in the analysis, the mean value of R-squared is 0.75 (standard deviation 0.12; 5th, 50th and 95th percentiles=0.51, 0.77, 0.89 respectively). This data indicates that in the vast majority of these stations the strength of the relationship between global grades and domain scores is very good – in turn suggesting that BRM is generally providing defensible standards for this examination (Pell *et al.* 2010).

Where the value of R-squared is relatively low, this implies that the domain scores are not discriminating strongly between different global grades (Pell *et al.* 2010). The most extreme example in the whole data set (R-squared=0.10) is shown in Figure 1 (a station where a patient presents with a urinary problem):

FIGURE 1 HERE

This station level plot of a single administration of this station highlights the degree of correlation between a global grade and domain 'score'. Each 'dot' on the plot represents an individual assessor-candidate observation, with bigger 'dots' reflecting a number of identical observation/scores. In Figure 1, there is a relative lack of discrimination in scores – most candidates are scoring quite highly in both global grades and domain scores, and there is a lack of spread in both of these measures.

For such poor values of this important metric, the key question is what the impact is on the defensibility of the standard set? In general across the full PLAB2 dataset, we find there is a weak negative correlation between R-squared values in stations and corresponding station-level cut-scores (r=-0.13, n=3564, p<0.001). This suggests that low values of R-squared are typically associated with slightly higher cut-scores, potentially leading to higher failure rates.

The urinary station (single administration shown in Figure 1) has been administered 22 times over the period concerned. Interestingly, the median R-squared across these administrations is 0.70 which very strongly suggests that the low R-squared is not typical for this station, and therefore is likely to be either an individual assessor issue, and/or a problem with lack of spread in the scores in the station in this particular administration, rather than an underlying problem with the station design. The cut-scores across these 22 administrations show some variation, with a standard deviation of 0.86 domain marks (equivalent to 7.2% of the scale) – and the cut-score with the lowest

R-squared value (as shown in Figure 1) produces the highest cut-score across all administrations of this station. However, we emphasise that the prevalence of stations with poor R-squared values across the data set as a whole is low (e.g. 5%).

Station level pass rates

The pass rate at the station level has a median value of 74% across the 3564 station administrations. This indicates that typically a significant proportion of candidates are scoring relatively poorly in stations, and that there is a reasonably wide range of abilities within most cohorts – in other words, a significant number of candidates with each cohort are receiving low scores in addition to a number of candidates doing well. This range of candidate performances within stations is a key requirement for the successful application of BRM in these contexts, an issues we will return to in the Discussion.

Variation in the standard within stations

The variation in cut-scores for each station across the full dataset has a (median) standard deviation of 5.6% which is suggests that generally the cut-scores set by BRM are broadly stable across administrations.

2. Sequence 2 exams

It is not appropriate to calculate reliability figures for Sequence 2 in isolation as this sub-group, is by definition, weaker than the full cohort. Instead, we use decision theory to estimate overall reliability based on Sequence 1 scores alone. (Pell *et al.* 2013). In this context, omega-total (Revelle and Zinbarg 2008) for the full sequence is of the order of 0.80 (or greater) for each of the six examinations studied.

Across the six sequential examinations from 2017-2019, Table 1 gives the number of stations where BRM metrics and spread of marks were judged sufficiently satisfactory for the corresponding cut scores to be used with confidence in the standard setting for the second part of the sequence.

Where this was not the case, the previous cut-score for the station, derived from administration in a full cohort and with satisfactory BRM metrics, was used.

TABLE 1 HERE

The analysis summarised in Table 1 suggest that BRM provides a feasible, and defensible approach to standard setting in these exams for a large percentage of stations (86%). A good example of where borderline regression is not doing so is shown in Figure 2, a Knee examination station (Year 5, 2018, 22 candidates). The problem here is lack of spread in the global grades (only two of five possible grades employed), which leads to a low R-squared (=0.14) and subsequently to concern about the accuracy of the BRM pass mark for this administration.

FIGURE 2 HERE

For this station, the BRM cut-score from a previous use of this station in a full cohort with good metrics is preferred.

Comparison with main cohort standards

We next compare the BRM small cohort overall standard (for the 10 or 12 sequence 2 stations in Year 4 and Year 5 respectively) with that for the standard generated from previous full cohort data. In four out of the six administrations under consideration, there appears to be a tendency for BRM to produce a slightly higher standard in small cohorts (of the order of 5%). This issue is clearly worthy of additional research.

3. PA exams

Overall test level reliability as calculated by Cronbach's alpha for each of the four PA examinations from 2018 is at least 0.80 in each case.

Table 2 details the number of stations in each examination where BRM metrics were judged satisfactory based on visual inspection of scatter graphs and values of R-squared. We see that in the majority of stations (88%) the data suggest that BRM standards are defensible.

TABLE 2 HERE

Standard error of the cut-score

The standard errors of the overall BRM pass mark for the PA exams is estimated using resampling methods, and are of the order of 1% across each of the four examinations. These values are considered acceptable, in that they are lower than extrapolated values from main exams found in the literature (≈1.4%) (Muijtjens *et al.* 2003, Homer *et al.* 2016).

Comparison with Angoff judgements

As part of established practice, the PA OSCE team produces an Angoff-type judgement of each station in terms of the expected proportion of minimally competent PA candidates who would pass the station. There is obvious interest in comparing this approach with a BRM derived standard. Figure 3 gives a scatter graph for the 2018 Year 1 examination comparing BRM standards (horizontally) with those from the Angoff (vertically) – both calculated as the percentage of the total station score. The blue (dashed) line is the line of best fit ($r=0.68$, $n=16$, $p=0.004$), the orange (bold) line is $y=x$ (i.e. if cut-scores for each standard setting method were the same in each station they would all be on this line):

FIGURE 3 HERE

Figure 3 shows that for more challenging stations (bottom left corner of graph), Angoff tends to give a higher cut-score compared to BRM (dots above the bold line) for the 2018 Year 1 PA exam, and for easier stations (top right) it tends to give a lower cut-score (dots below bold line). These differences result in Angoff giving a higher overall cut-score - 69% vs. 66% for BRM, and this corresponds to one additional failure for the cohort of approximately 20 candidates under an Angoff set cut-score across all stations.

At the station level, there are also more individual station failures in the 2018 Year 1 PA exam under Angoff (78 vs. 52). Finally, BRM gives more variation in passing scores - BRM has a broader (horizontal) range of 39% in cut-scores across stations compared to a more constricted Angoff (vertical) range of 19% - with similar findings in the other PA data. A reasonable interpretation of these analyses suggests that BRM is producing a more realistic range of cut-scores, whereas Angoff scores show a more restricted range.

Discussion

Across a large set stations from three different and diverse small cohort OSCE contexts, BRM has been shown to function effectively in the vast majority of stations (over 86% in each context). Where BRM produces satisfactory station level metrics, we argue this contributes important evidence towards the wider validity argument in the justification of the use of high stakes test outcomes (Kane 2013).

Existing assumptions about the effectiveness of BRM as a standard setting method have tended to be informed by high stakes data from larger cohorts of candidates from a positively skewed population (Pell *et al.* 2010, McKinley and Norcini 2014) – namely, where the vast majority of students occupy a 'competent-excellent' range of ability . In each of our three contexts, we have provided evidence that BRM can deliver defensible standards in the majority of stations. Where this is not the case, this is usually because of a poor relationship between global grades and checklist/domain scores, that brings into question the BRM standard (Pell *et al.* 2010). This problem is often brought about by a lack of sufficient spread in candidate scores (e.g. Figure 1), and for small cohorts, the risk of a 'restricted range' of scoring is obviously more likely compared to larger cohorts. Despite this hypothesis, the prevalence of this was not particularly high across any of the three contexts, perhaps reflecting the interplay of good OSCE station (and scoring) design and the ability of (and support for) assessors global judgments about observed performance of candidates.

One might hypothesise that BRM standard setting should 'work' better in PLAB2 compared to the other contexts, since this exam has a relatively high failure rate (typically 26% at the station level) which indicates that scores are more variable within the cohort – a characteristic that, from a technical point of view, makes BRM more likely to function effectively (Draper and Smith 1998, chap. 3). The failure rates in the sequential context are generally lower (median failure rate in station of the order of 18%), and for the PA examinations are lower still (≈12%). Our analysis does indeed partially confirm this hypothesis, with only a small proportion (≈5%) of PLAB2 stations being problematic for the application of BRM, whereas the prevalence of issues with BRM are a little higher in the other two contexts (14% for sequence 2, and 12% for PA). The issue of spread of marks/grades (and candidate ability) and the impact on standards (and error associated with this) under BRM is an area that requires more research, but our work might indicate that BRM remains unsuited to some small cohort assessment formats testing where there is a limited spread of candidate ability (e.g. highly specialised postgraduate examinations).

Assessor stringency and assessment design

One important difference between our three contexts and larger cohort exams is the lack of parallel circuits in the former (Harden *et al.* 2015, chap. 6). In other words, there is (usually) a single assessor corresponding to each station in our three contexts, whereas in a large undergraduate medical school, for example, there might be of the order of 20 or more parallel circuits, so that each station is assessed by a large number of individual assessors. In one sense, having a single assessor is more likely to bring consistency to the marking of the station since all candidates will be observed by the same set of assessors across the exam. On the other hand, in small cohorts single assessors are confounded with stations and there is no easy way to directly compare assessor stringency at the station level (Pell *et al.* 2010, Yeates *et al.* 2018). The presence of a single assessor per station (rather than multiple across circuits) may also contribute to a lack of calibration compared to a group of peers all examining the same station. Pertinent to this issue, recent work by Crossley and colleagues (2019) describes a complex balance of 'guarded curiosity' (of comparative judgments by peers), affective bias and 'moderated conservatism' where assessors balance openness to change alongside loyalty to personal judgments.

Exploring this theme further, the impact of assessor scoring stringency (i.e. 'hawks and doves') (Yeates and Sebok-Syer 2017) is likely to be greater on BRM standards in small cohorts. With 'generous' markers, the scores will tend to be in the top right corner of the scatter diagram (see Figure 1) and this means that there is considerable uncertainty in the 'correct' cut-score when extrapolating back towards the borderline grade via the regression line. This issue is perhaps less acute with stricter markers as then the scores will be near the borderline grade and the degree of extrapolation, and hence the 'error', is therefore likely to be smaller in comparison. A related, but perhaps, counter-intuitive point is that less stringent assessors might actually raise the cut-score under BRM – the administration of the urinary station shown in Figure 1 has the highest cut-score out of the 22 administration of the same station in the dataset, and yet the assessor is giving relatively high scores and is likely to be on the dovish end of the assessor stringency scale. Clearly, these issues are complex and would benefit from further investigation, perhaps using statistical simulation methods (Currie and Cleland 2016, Homer *et al.* 2016), which could aid further thinking with regard to the selection and training of assessors in small cohort exams.

A final comment concerning design issues relates to the nature of the rating scale for the global grade. More research is needed to investigate whether, for example, four or five point scales are more appropriate depending on the context – and perhaps in small cohorts the evidence might favour a shorter scale. One thing we recommend is that the scale broadly reflects the ability profile of the candidate pool – thereby making it more likely that all grades are actually used by assessors. Faculty should conceptualise the range of performances that are seen, either side of just 'safe' to produce the rating scale, and this process can be reviewed regularly. In all three of our contexts the scale is asymmetric, with multiple passing grades and a single fail grade. To our knowledge, there is no published work comparing the efficacy of differently constructed global rating scales under BRM.

Comparisons with other standard setting approaches

As a general principle, we prefer standards set using data from the actual examination, rather than that derived from data from previous administrations – elements of station design, current medical practice, and standards of assessor training are constantly developing. It is only in the current administration that all these factors can be taken into account fully in the pattern of scores/grades awarded, and the standard thereby set. However, analysis of six sequential test administrations has shown small differences in the BRM derived standards of Sequence 2 stations when derived directly from the small cohort examined versus those from that same stations derived from larger cohorts (e.g. when blueprinted as part of Sequence 1). This is clearly an important issue and merits further investigation to better understand whether these differences are the result of substantive, systematic changes in assessor behaviour between small and large cohorts. In the sequence 2 examination, the students are by definition a weaker sub-group, and to an extent the stakes may be perceived by assessors to be higher than they are in the sequence 1 examination. Might this knowledge influence the marking that assessors provide?

In the PA exam, we have seen that assessors tend to shy away from extremes when providing their Angoff scores (and that this is far less of a problem with BRM) – this carries an echo of the well-known issue in workplace-based assessment where there is a reluctance for assessors to use the full scale (Crossley and Jolly 2012). There may also be social reasons for this presumed reluctance in an *a priori* standard setting meeting (Fitzpatrick 1989) but the current study has no data to speak to this issue.

Conclusions

Challenging established assumptions, and using relatively simple methods (e.g. visual inspection of scatter graphs and calculation of R-squared values), this study has shown that the use of BRM in context of the small cohorts can be generally successful. We have avoided in this paper more technical approaches to judging robustness of regression-based approaches (e.g. robust regression, Bayesian methods, or more sophisticated modelling approaches) but these might well merit further application in the future (Wilcox 2005, chap. 10, Tavakol *et al.* 2018). There are also philosophical issues that are worth of further consideration, but beyond the scope of this paper, regarding what exactly we mean by the standard for a station that is used regularly – for example, is it the standard on the day in question, or should we use all data from previous administrations to derive the standard?

However, from a practical point of view, extant cut-scores, preferably based on previous satisfactory station performance (Pell *et al.* 2010), should ideally be available for all stations in small cohort exams so that when problems with BRM do occur (e.g. a lack of spread of scores) these cut-scores can substitute without the need for the removal of the stations. We recognise that poor metrics might well indicate that there is a problem in the station (e.g. a design issue, or a lack of understanding of expected level of performance, or an assessor training issue). The principle of parsimony and the need to maintain the blueprint would suggest not removing the station from the exam unless scores are clearly erroneous. However, such stations should be flagged for post-examination review to better understand the causes of the problems.

One could argue that the need for 'backup' cut-scores might also be true to an extent for larger cohorts for the rare occasions when BRM proves problematic based on *post hoc* analysis – unless one is happy with removing poorly performing stations which itself threatens the quality of blueprinting process and ultimately the validity of the assessment (Downing and Haladyna 2004).

As with all development and use of high quality assessment tools, we comment finally that care always needs to be taken in the overall design of station level scoring instruments (i.e. global rating scales, and key features checklists or domain scoring scales) based on clear articulation of the purpose(s) of the assessment and the inferences to be drawn from its outcomes (Kane 2013). Whilst the effective use of BRM should include consideration of adequate cohort size, it is also contingent on a range of other factors including the degree of examinee heterogeneity, good station design and satisfactory assessor training and behaviour.

Acknowledgements

We thank the GMC for providing access to the anonymised data used as part of this study. We also thank our friend and colleague, the late John Patterson, for prompting the initial research on small cohorts in 2012 which eventually led to this work.

Declaration of interest

All authors have no declaration of interest to make.

Ethics

The University of Leeds and the GMC gave permission for the anonymized data to be used for research in this paper. The co-chairs of the University of Leeds School of Medicine ethics committee confirmed to the authors that formal ethics approval for this study was not required as it involved the use of routinely collected student assessment data which were fully anonymized prior to analysis.

References

American Educational Research Association, 2014. *Standards for Educational and Psychological Testing*. Washington, D.C: American Educational Research Association.

Ben-David, M.F., 2000. AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, 22 (2), 120–130.

Bland, J.M. and Altman, D.G., 2011. Correlation in restricted ranges of data. *BMJ*, 342 (mar11 1), d556–d556.

Boulet, J.R., De Champlain, A.F., and McKinley, D.W., 2003. Setting defensible performance standards on OSCEs and standardized patient examinations. *Medical Teacher*, 25 (3), 245–249.

Boursicot, K.A.M., Roberts, T.E., and Pell, G., 2007. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. *Medical Education*, 41 (11), 1024–1031.

Cizek, G.J., ed., 2012. *Setting Performance Standards: Foundations, Methods, and Innovations*. 2 edition. New York: Routledge.

Cizek, G.J. and Bunch, M.B., 2007. *Standard setting a guide to establishing and evaluating performance standards on tests*. Thousand Oaks, Calif.: Sage Publications.

Clauser, B.E., Mee, J., Baldwin, S.G., Margolis, M.J., and Dillon, G.F., 2009. Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*, 46 (4), 390–407.

Crossley, J. and Jolly, B., 2012. Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical Education*, 46 (1), 28–37.

Crossley, J.G.M., Groves, J., Croke, D., and Brennan, P.A., 2019. Examiner training: A study of examiners making sense of norm-referenced feedback. *Medical Teacher*, 0 (0), 1–8.

Currie, D.G.P. and Cleland, P.J., 2016. Sequential objective structured clinical examinations: Number of stations. *Medical Teacher*, 38 (8), 857–858.

Cusimano, M.D., 1996. Standard setting in medical education. *Academic Medicine: Journal of the Association of American Medical Colleges*, 71 (10 Suppl), S112-120.

Downing, S.M. and Haladyna, T.M., 2004. Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education*, 38 (3), 327–333.

Draper, N.R. and Smith, H., 1998. *Applied Regression Analysis, 3rd Edition*. 3rd edition. New York: Wiley-Blackwell.

Dwyer, T., Wright, S., Kulasegaram, K.M., Theodoropoulos, J., Chahal, J., Wasserstein, D., Ringsted, C., Hodges, B., and Ogilvie-Harris, D., 2016. How to set the bar in competency-based medical

education: standard setting after an Objective Structured Clinical Examination (OSCE). *BMC Medical Education*, 16 (1), 1.
Farmer, E.A. and Page, G., 2005. A practical guide to assessing clinical decision-making skills using the key features approach. *Medical Education*, 39 (12), 1188–1194.
Fitzpatrick, A.R., 1989. Social Influences in Standard Setting: The Effects of Social Interaction on Group Judgments. *Review of Educational Research*, 59 (3), 315–328.
Fuller, R., Homer, M., and Pell, G., 2013. Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. *Medical Teacher*, 35 (6), 515–517.
General Medical Council, 2019. What is the PLAB 2 exam? [online]. *General Medical Council*. Available from: https://www.gmc-uk.org/registration-and-licensing/join-the-register/plab/plab-2-guide/what-is-the-plab-2-exam [Accessed 7 Jan 2019].
Harden, R., Lilley, P., and Patricio, M., 2015. *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment., 1e*. 1 edition. Edinburgh ; New York: Churchill Livingstone.
Hays, R., Gupta, T.S., and Veitch, J., 2008. The practical value of the standard error of measurement in borderline pass/fail decisions. *Medical Education*, 42 (8), 810–815.
Health Education England, 2015. Physician associate [online]. *Health Careers*. Available from: https://www.healthcareers.nhs.uk/explore-roles/medical-associate-professions/roles-medical-associate-professions/physician-associate [Accessed 7 Jan 2019].
Hejri, S.M., Jalili, M., Muijtjens, A.M.M., and Van Der Vleuten, C.P.M., 2013. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, 18 (10), 887–891.
Homer, M., Fuller, R., and Pell, G., 2018. The benefits of sequential testing: Improved diagnostic accuracy and better outcomes for failing students. *Medical Teacher*, 40 (3), 275–284.
Homer, M., Pell, G., Fuller, R., and Patterson, J., 2016. Quantifying error in OSCE standard setting for varying cohort sizes: A resampling approach to measuring assessment quality. *Medical Teacher*, 38 (2), 181–188.
Kane, M.T., 2013. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50 (1), 1–73.
Kramer, A., Muijtjens, A., Jansen, K., Düsman, H., Tan, L., and van der Vleuten, C., 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. *Medical Education*, 37 (2), 132–139.
Livingston, S.A. and Zieky, M.J., 1982. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.
Margolis, M.J., Mee, J., Clauser, B.E., Winward, M., and Clauser, J.C., 2016. Effect of Content Knowledge on Angoff-Style Standard Setting Judgments. *Educational Measurement: Issues and Practice*, 35 (1), 29–37.
McKinley, D.W. and Norcini, J.J., 2014. How to set standards on performance-based examinations: AMEE Guide No. 85. *Medical Teacher*, 36 (2), 97–110.
Muijtjens, A.M.M., Kramer, A.W.M., Kaufman, D.M., and Van der Vleuten, C.P.M., 2003. Using Resampling to Estimate the Precision of an Empirical Standard-Setting Method. *Applied Measurement in Education*, 16 (3), 245–256.
Pell, G., Fuller, R., Homer, M., and Roberts, T., 2010. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. *Medical Teacher*, 32 (10), 802–811.
Pell, G., Fuller, R., Homer, M., and Roberts, T., 2013. Advancing the objective structured clinical examination: sequential testing in theory and practice. *Medical Education*, 47 (6), 569–577.
Pell, G., Homer, M., and Fuller, R., 2015. Investigating disparity between global grades and checklist scores in OSCEs. *Medical Teacher*, 37 (12), 1106–1113.
Revelle, W. and Zinbarg, R.E., 2008. Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, 74 (1), 145.

Schoonheim-Klein, M., Muijtjens, A., Habets, L., Manogue, M., van der Vleuten, C., and van der Velden, U., 2009. Who will pass the dental OSCE? Comparison of the Angoff and the borderline regression standard setting methods. *European Journal of Dental Education*, 13 (3), 162–171.
Tavakol, M., Pinner, G., and Doody, G.A., 2018. The Bayesian borderline regression method: Identifying pass marks for small cohorts. *Medical Teacher*, 0 (0), 1–1.
Wilcox, R.R., 2005. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press.
Wood, T.J., Humphrey-Murto, S.M., and Norman, G.R., 2006. Standard setting in a small scale OSCE: a comparison of the Modified Borderline-Group Method and the Borderline Regression Method. *Advances in Health Sciences Education: Theory and Practice*, 11 (2), 115–122.
Yeates, P., Cope, N., Hawarden, A., Bradshaw, H., McCray, G., and Homer, M., 2018. Developing a video-based method to compare and adjust examiner effects in fully nested OSCEs. *Medical Education*, 0 (0).
Yeates, P. and Sebok-Syer, S.S., 2017. Hawks, Doves and Rasch decisions: Understanding the influence of different cycles of an OSCE on students' scores using Many Facet Rasch Modeling. *Medical Teacher*, 39 (1), 92–99.

Author information

Matt Homer, BSc, MSc, PhD, PGCE, CStat, is an Associate Professor in the Schools of Education and Medicine at the University of Leeds. Within medical education, he has a research interest in assessment design, standard setting methodologies and psychometrics analysis. He also advises the UK General Medical Council on a range of assessment issues.

Richard Fuller MA, MBChB, FRCP, is a Consultant Geriatrician/Stroke Physician and Vice-Dean of the School of Medicine at the University of Liverpool. His current research focuses on the application of intelligent assessment design in campus and workplace-based assessment formats, assessor behaviours, mobile technology delivered assessment and the impact of sequential testing methodologies.

Jennifer Hallam, BSc, PG Dip, MSc, PhD, is an Educational psychometrician in the School of Medicine, University of Leeds. Her current interests include the strategic development of assessment and feedback strategies, specifically for performance based assessments. She also has several national medical education roles which include being on the Board of Directors for the Association for the Study of Medical Education (ASME).

Godfrey Pell, BEng, MSc, CStat, is a principal research fellow emeritus at Leeds Institute of Medical Education, who has a strong background in management. His research focuses on quality within the OSCE, including theoretical and practical applications. He acts as an assessment consultant to a number of medical schools.

Tables

Table 1

Year	Year group	Cohort size for sequence 2	Number of S2 stations for which BRM metrics satisfactory	Issue with BRM metrics
2017	4	41	8 out of 10	Two stations had relatively low R-squared (<0.4)
2017	5	33	11 out of 12	For one station, the spread in global grades was limited which raised questions over the robustness of the BRM cut-score
2018	4	55	10 out of 10	BRM cut-score used in all stations
2018	5	22	7 out of 12	Five stations had relatively low R-squared (<0.4)
2019	4	51	10 out of 10	BRM cut-score used in all stations
2019	5	15	11 out of 12	Whilst the metrics were generally good, it was decided that all previous pass marks should be used given the particularly small cohort size (n=15). BRM was especially problematic in one station in that the intercept (i.e. predicted mark for a failing student) was negative.
Total			**57 out of 66 (86%)**	

Table 1: Summary of stations where BRM metrics indicated appropriate for setting the standard in six sequence 2 examinations

Table 2

Year	Year group	Cohort size	Number of stations for which BRM metrics satisfactory	Issue with BRM metrics
2018	1	25	15 out of 16	One station with low R-squared (<0.4)
2018	2	23	14 out of 16	Two stations with low R-squared (<0.4)
2019	1	25	13 out of 16	Three stations with low R-squared (<0.4)
2019	2	26	14 out of 16	Two stations with low R-squared (<0.4)
Total			**56 out of 64 (88%)**	

Table 2: Summary of stations where BRM was appropriate for setting the standard in four PA examinations

Figures

Figure 1

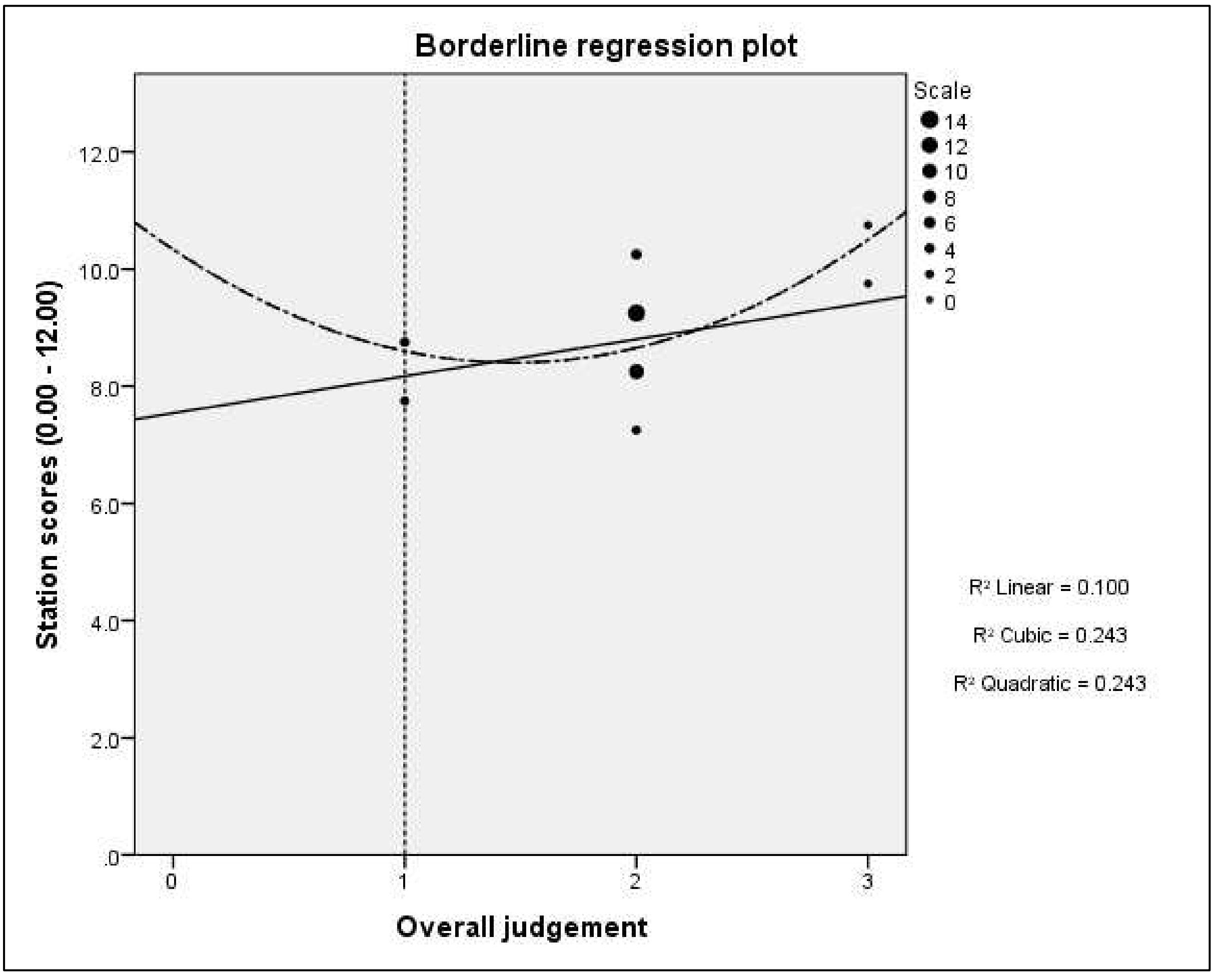


Figure 2

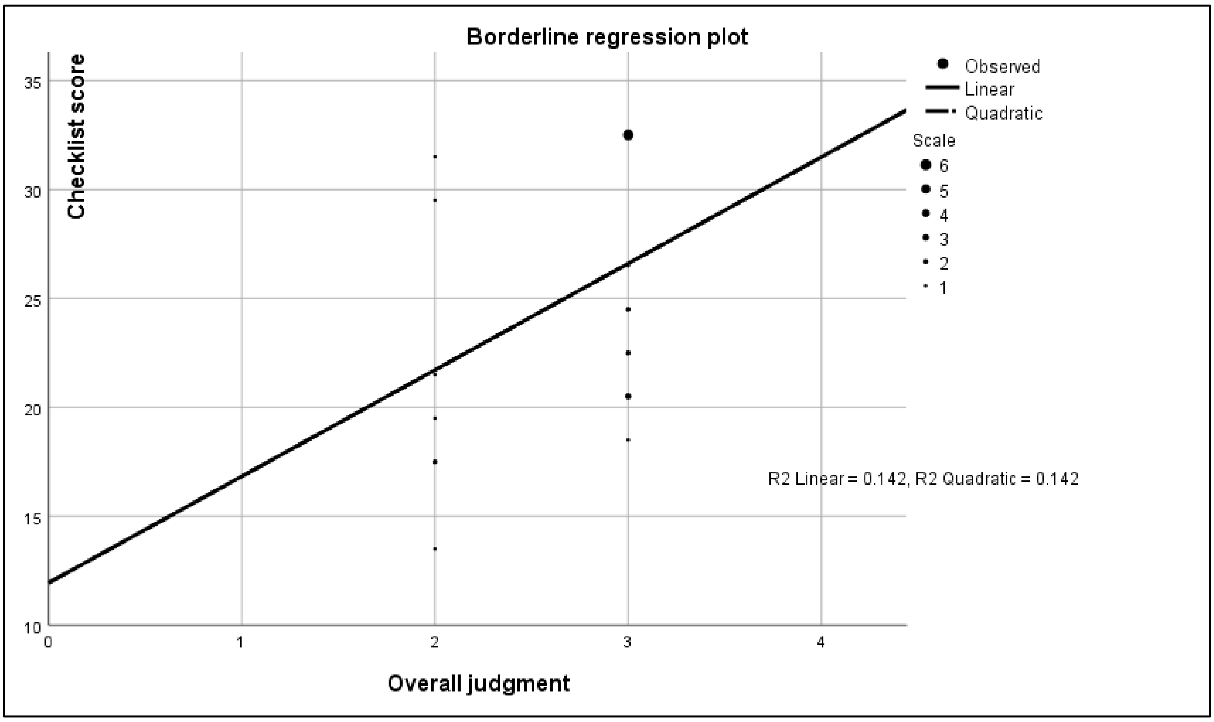
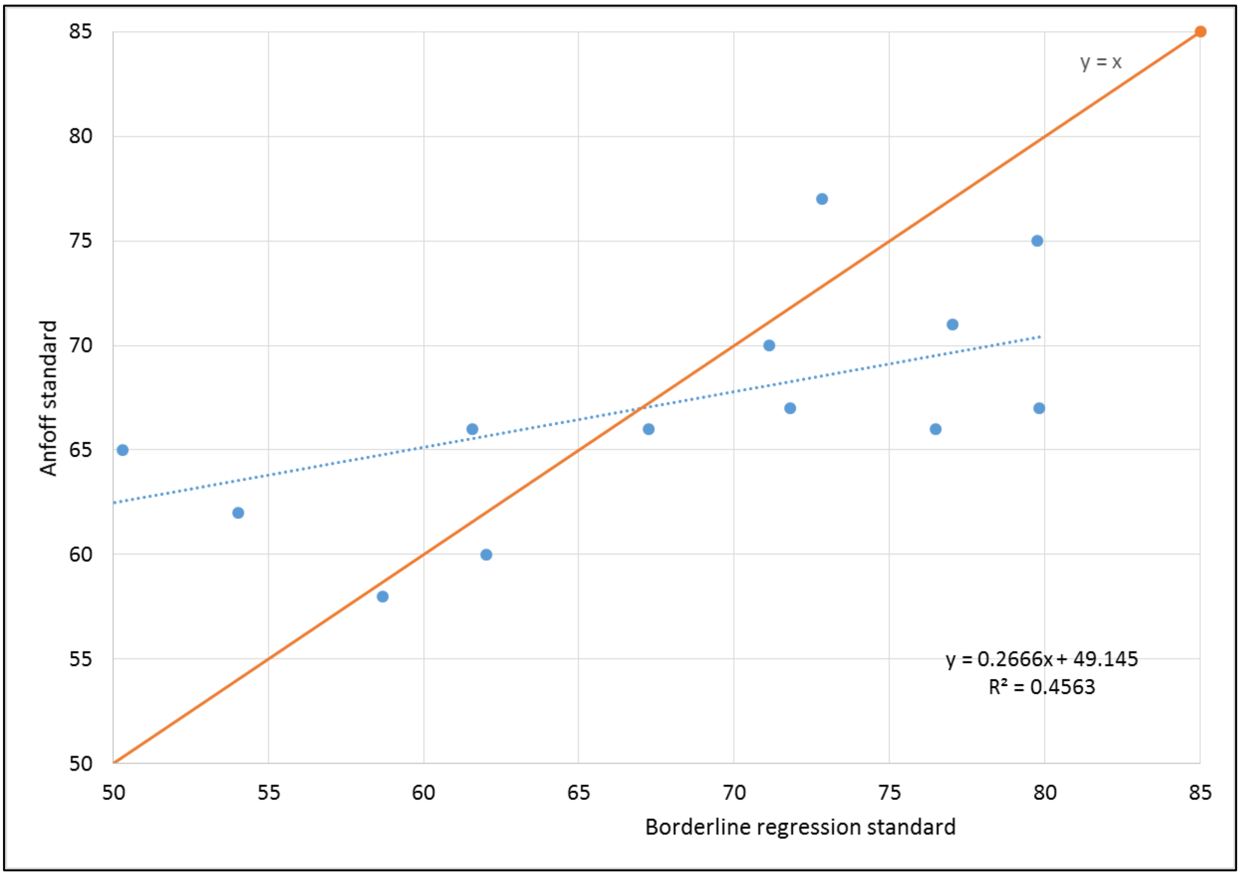

Borderline regression plot
Checklist score
35
30
25
20
15
10
0
1
2
3
4
Overall judgment
Observed
Linear
Quadratic
Scale
6
5
4
3
2
1
R2 Linear = 0.142, R2 Quadratic = 0.142

Figure 3



85
80
75
70
65
60
55
50
Anfoff standard
y = x
y = 0.2666x + 49.145
R² = 0.4563
50
55
60
65
70
75
80
85
Borderline regression standard

Figure Captions

Figure 1: Global grades against domain scores for station with low R-squared (0=unsatisfactory, 1=borderline, 2=satisfactory, 3=good)

Figure 2: Global grades against checklist scores for station with lack of spread, particularly in global grades (0=fail, 1=borderline, 2=pass, 3=good pass, 4=excellent pass)

Figure 3: BRM standards for the PA Year 1 exam in 2018 versus Angoff