This is a repository copy of *Problematizing the concept of the 'borderline' group in performance assessments*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/112354/

Version: Accepted Version

Article:

Homer, MS orcid.org/0000-0002-1161-5938, Pell, G and Fuller, R (2017) Problematizing the concept of the 'borderline' group in performance assessments. Medical Teacher, 39 (5). pp. 469-475. ISSN 0142-159X

https://doi.org/10.1080/0142159X.2017.1296563

© 2017, Taylor & Francis. This is an Accepted Manuscript of an article published by Taylor & Francis in Medical Teacher on 6 March 2017 available online: https://doi.org/10.1080/0142159X.2017.1296563

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Problematizing the concept of the 'borderline' group in performance assessments

Abstract

Introduction

Many standard setting procedures focus on the performance of the 'borderline' group, defined through expert judgements by assessors. In performance assessments such as Objective Structured Clinical Examinations (OSCEs), these judgements usually apply at the station level.

Methods and results

Using largely descriptive approaches, we analyse the assessment profile of OSCE candidates at the end of a five year undergraduate medical degree programme to investigate the consistency of the borderline group across stations. We look specifically at those candidates who are borderline in individual stations, and in the overall assessment. Whilst the borderline group can be clearly defined at the individual station level, our key finding is that the membership of this group varies considerably across stations.

Discussion and conclusion

These findings pose challenges for some standard setting methods, particularly the borderline groups and objective borderline methods. They also suggest that institutions should ensure appropriate conjunctive rules to limit compensation in performance between stations to maximise 'diagnostic accuracy'. In addition, this work highlights a key benefit of sequential testing formats in OSCEs. In comparison with a traditional, single-test format, sequential models allow assessment of 'borderline' candidates across a wider range of content areas with concomitant improvements in pass/fail decision-making.

Introduction

Most standard setting procedures used in Objective Structured Clinical Examinations (OSCEs) focus on the ability and performance of the "borderline" or minimally competent candidate group, defined through expert judgements of assessors which are combined into a generic descriptor of this hypothetical level of performance (Livingston & Zieky 1982; Ben-David 2000; Norcini 2003; Cizek & Bunch 2007, p. 48; Boursicot et al. 2007; McKinley & Norcini 2014). Whilst these descriptors are generally applied at the item (i.e. station) level, the original model (Livingstone and Zieky 1982), defined the borderline group across the whole assessment. However, in the medical education setting of the OSCE, more recent literature (Reznick et al. 1996; McKinley & Norcini 2014) generally suggests that the borderline group is defined at the individual station level. Given this difference in approach, part of the motivation of this paper is to scrutinise in detail the distinction between a definition of borderline across the whole assessment, and one within individual items/stations only.

There are two principal methods of standard setting that invoke the concept of 'borderline' performance; the first is the borderline regression method (Kramer et al. 2003; Pell & Roberts 2006) where one of the global performance grades for each station is 'borderline' (or similar), and the passing score is usually determined using the weighted average total checklist/domain score corresponding to this 'borderline' grade via simple regression of checklist scores on global grades. The second is the borderline groups method (Livingston & Zieky 1982; Cizek & Bunch 2007, p. 112–113), which also requires a 'borderline' grade at the station level, and in this approach the average (usually the median) checklist score within this grade determines the required passing score within a station. More recently, a variation on these methods, the

Objective Borderline method, has been advocated which makes strong assumptions about the checklist score distribution within the borderline group (Shulruf et al. 2013; Shulruf et al. 2015) – essentially that higher global grades are always associated with higher checklist scores within a station.

Prompted by consideration of these different standard setting methods, and the assumptions that underpin them, this work investigates the nature of the borderline group at OSCE station and whole assessment level, and explores the appropriateness of the assumptions behind both of the borderline groups and objective borderline methods.

Previous research on the 'borderline' group

Earlier work has investigated the consistency and congruence of assessor checklist and global grade pass/fail decisions within stations (Pell et al 2015). This showed that for many stations, there is a high degree of disagreement in these decisions and that this appears to be station-specific with no obvious systematic pattern. This earlier work provided a natural stimulus to a wider consideration of the consistency or otherwise of the 'borderline' group in the OSCE. A review of the literature reveals very little specific research on the borderline group and its consistency or otherwise across the assessment. Of particular note, there appear to be no studies that consider the extent to which membership of the borderline group in one station matches that of other stations, or across the assessment as a whole.

We might speculate that this limitation partly reflects the awareness of the issue of case-specificity, where it is well known that success on one station (or case) does not strongly predict performance on other stations (Eva et al. 1998; Norman et al. 2006; Wimmers et al. 2007; Wimmers & Fung 2008; Mattick et al. 2008). However, the extent to which this finding, which has only been shown to apply to the cohort as a whole, also applies to those near the passing standard (i.e. borderline) is to our knowledge left unexplored in the literature. Tweed and Wilkinson (2010) question whether the ‘borderline’ group exists across the entire assessment in the sense that an individual candidates’ true ability will either meet program requirements or not. In their argument, it is therefore the assessment process, with its attendant error, which makes overall pass/fail decisions difficult and leads to the concept of ‘borderline’. Theirs is a useful first critique of the notion of the ‘borderline’ which will be explored later.

Research aims

In this paper we are interested in investigating the following two overarching research questions:

- To what extent is the borderline group of candidates consistent between stations and between examination sites?
- How do borderline judgements compare at the station and overall test level?

We will demonstrate that the answers to these questions have important assessment design implications, certainly in assessment settings common to single medical schools.

Methods

Assessment context

We focus on the fifth and final year of an undergraduate medical programme at the University of Leeds, UK where the more complex behaviours assessed lend themselves to detailed analysis. The OSCE examination is made up of a first part of 13 integrated stations with simulated and real patients. Each station is highly integrated, involving patient assessment, clinical reasoning and decision making, patient management (including clinical investigation and prescribing) and escalation/referral across a range of common and critical patient presentations. The full assessment is in a sequential format (Cookson et al. 2011; Pell et al. 2013) with an additional 12 stations for those not performing adequately over the first part. All the data we present in this paper is from the first part of the sequence that all candidates must sit. There are usually around 250-280 candidates spread across multiple parallel circuits and across examination centres in up to four different hospital sites.

OSCE assessors undergo comprehensive training, with refresher training as appropriate, and on the day of the examination there is a pre-OSCE run through and discussion of the purpose of each station and of what is being assessed. To help minimise unwanted assessor variance in scoring, there is also consideration of how examiners are expected to behave in terms of the ‘script’, and in their interaction with candidates.

Within each circuit, a station is scored by single assessor (so there are a similar number of assessors as students) using a key features checklist (Farmer & Page 2005) together with a global grade on a five point scale (fail, borderline, clear pass, good pass and excellent pass). The borderline regression method (Kramer et al. 2003; Pell & Roberts 2006) is used for standard setting within stations, and the aggregate of the station passing scores is used to set the cut-score for the overall initial passing decision (adjusted upwards by two standard errors of measurement to eliminate, or at the very least minimise, false positives in the first part of the sequence) (Hays et al. 2008). All candidates are randomised into groups (i.e. circuits), with the exception of those who receive extra time as part of reasonable adjustments to assessment based on recognized disabilities.

The results presented here focus on recent ‘final’ year data, for Year 5 2015, with 283 candidates across four examination sites. Stations are drawn from a well-established and regularly reviewed station bank, and in this particular assessment extensive post hoc analysis indicates acceptable station level metrics (Pell et al. 2010; Fuller et al. 2013; Pell et al. 2015; Homer et al. 2016), determined by reasonable values for R-squared, low between-group error variance, and no stations detracting from overall reliability.

Definitions of borderline and analytic approaches

There are three separate definitions of ‘borderline’ that are worth distinguishing at this stage:

1. Station-level borderline based on station borderline grade – this applies to candidates who are awarded a ‘borderline’ global grade in the station.
2. Station-level borderline based on station checklist mark - this is based on the candidates ‘near’ the total checklist pass mark for the station assessment as set by the borderline regression.
3. Test-level borderline – this is based on the candidates ‘near’ the overall cut-score for the first part of the sequential assessment.

To facilitate clearer understanding (and to keep the analysis relatively straightforward), we will use the appropriate deciles to categorise candidates as ‘near’ the borderline in definitions 2 and 3. The overall aim in this work is to analyse the consistency of these different classifications – in order to gain understanding of the extent to which the ‘borderline’ group does or does not form a consistent group of candidates across different elements of the assessment. This analysis also includes the comparison of borderline classifications across the four different sites that the assessment takes place in.

In terms of specific methods, this paper makes use descriptive statistics, cross-tabs and basic inferential statistics (e.g. chi-square tests) – both within stations and across the assessment as a whole – to assess the consistency or otherwise of borderline classifications of candidates.

Results and interpretation

Station level variation in the borderline group

The successive columns of Table 1 show respectively station level metrics for (i) the percentage of candidates who were awarded a global grade 'borderline' by assessors, (ii) the percentage within this group who actually passed based on the checklist score under borderline regression, and (iii) the percentage of candidates failing the station. To illustrate the variation in each metric across stations, we have highlighted the highest (dark shading) and lowest (light shading) values for each column.

TABLE 1 HERE

Table 1 shows that any apparent borderline group lacks stability in the sense that the size and nature of this group varies greatly across stations. We take each data column in turn and describe what the corresponding metric is telling us about the nature of the 'borderline' group.

Firstly, the relative size of the borderline group (as defined by assessors in their global grades) varies significantly across stations - between 9% and 26% of the cohort (second column of Table 1; chi-square=57.1, df=12, p<0.001, phi=0.12). Clearly, this would be impossible if there were a single unified ‘borderline’ group across the stations.

Secondly, the proportion of the borderline group in each station (i.e. those with borderline grades in the station) who actually pass (based on checklist marks) is again variable – between 35% and 56% (third column of Table 1; chi-square=62.6, df=12, p<0.001, phi=0.13). If assessor judgments of borderline behaviour were consistent across a set of well-designed stations (i.e. those with acceptable station level quality metrics (Pell et al. 2010; Pell et al. 2015)), we would expect the percentages of candidates passing and failing within this group to be consistent across these.

Thirdly, based on the percentage of candidates who fail in a station (final column of Table 1), the rank position of the borderline group can vary by up to a complete decile (comparing the 10% and the 22% failing in the third column of Table 1; chi-square=34.7, df=12, p=0.001, phi=0.10). Note that under methods such as borderline regression, the pass mark is designed to provide a consistent standard across stations, thereby automatically taking into account variation in station difficulty. For a consistent group of borderline candidates across the stations, we would expect an approximately uniform failure rate.

Some of these findings can be explained by ‘measurement error’, for example, as a result of variation in assessor judgments exacerbating apparent differences across stations as evidenced in Table 1. We would argue that if the relatively large effect sizes shown were due to measurement error alone, then the station level metrics for many of these stations would be clearly unsatisfactory – but as previously highlighted, this is not the case. Hence, we can be confident that the differences evidenced in Table 1 represent real and important effects, and are not just as a result of measurement error.

Congruence of borderline within stations and across the exam

For a single representative station (station 5 in Table 1), with otherwise good quality metrics (e.g. reasonable R-squared, good inter-grade discrimination (Pell et al. 2010; Pell et al. 2015;)), Table 2 cross-tabulates global grades and performance deciles based on checklist scores. This analysis indicates that the borderline grade encompasses a wide range of checklist performance levels – from the 4th to the 10th deciles. It is notable that despite acceptable R-squared values for this station (R-squared=0.58) this does not preclude a relatively wide range of checklist marks for particular global grades. Previous work has highlighted that apparently reasonable metrics can hide a range of potential problems (Pell et al. 2010; Fuller et al. 2013).

TABLE 2 HERE

One would expect some level of spread in checklist marks for any grade, including the ‘borderline’ grade. However, the evidence in Table 2 suggests that simplistic notions of ‘borderline’ performance are problematic, even within a station, and at the very least it might be better to think of ‘borderline’ performance as multi-faceted.

We have also compared borderline decisions in this station with those in each of the other 12 stations in the assessment. To do this we use a set of twelve 2×2 cross-tabs – borderline/not-borderline in station 5 against borderline/not-borderline in each other station. We find the average level of disagreement in these classifications (i.e. the proportion in the off-diagonal) is 27%. Again this is evidence that however the borderline group is defined, its constitution is not consistent across stations.

In Table 3, we compare performance decile within the station 5 with decile across the assessment as a whole (based on total checklist score across the 13 stations).

TABLE 3 HERE

If the classifications in the station were well-aligned with those in the assessment as a whole we would see all cases in the on-diagonal (top-left to bottom right in bold), but in fact only 17% of cases (48 out of 283) are situated here. Exploring borderline performance at the station level (e.g. deciles 8-10), Table 3 shows that only 45 (shaded) out of the 83 cases in total (bottom three rows; 54%) are in the same decile range by total score. This indicates that the borderline group is not consistent when comparing station and overall performance.

Further analysis indicates that the total number of station-level borderline grades a student receives across the assessment as a whole varies between 0 and 7 (out of a maximum 13) – with these latter students all in the lowest performance decile (i.e. the 10th) overall. When looking at weaker performing students overall, those in the 9th performance decile based on aggregate checklist scores have between 1 and 6 borderline grades across stations. For those in the 10th decile, the corresponding figures are 2 and 7. These analyses again indicate that the borderline group is not made up of a single group of students.

Considering now overall pass/fail decisions, the failure rate in this OSCE is approximately 5%. At the station level by contrast, the failure rate is somewhat higher (10-20% - see Table 1, final column), whilst up to 26% of the cohort are defined as ‘borderline’ in the global grade (Table 1, first column). So for the overall exam, the borderline group is clearly within the 10th decile, whereas at the station level is usually the 9th (or perhaps even 8th) decile for that station. Again, we evidence little consistency in these classifications.

Variation in use of borderline grades by Examination Centre/Site

For another station in the same Y5 assessment in 2015 (station 12 in Table 1), Figure 1 shows that there is significant disagreement between assessors on different sites in terms of the checklist performance which leads to a borderline grade under the borderline regression method – see the vertical variation in where each regression line cuts the x-value corresponding to the borderline grade (x=1). These within-site passing scores are 13, 15, 17 and 18 checklist marks, whereas the overall passing score across these sites is 16 when all data are combined.

FIGURE 1 HERE

Recalling that candidates are randomly distributed to examination centres, we conclude that the relationship between checklist scores and global grades shows important differences across sites, suggesting in particular that the understanding of the borderline candidate is different across them. As with earlier evidence, any conception of ‘borderline’ appears far more complex and situated than one might first imagine. This is an area of active research and we hypothesize that, despite common assessor training, different local environmental, clinical and workplace factors (e.g. patient safety alerts, critical incidents, hospital specific prescribing guidance) all interplay to generate some of this difference (Govaerts 2016).

Discussion

Implications for assessment design

When comparing across stations, there are four key findings worth particular emphasis. Firstly, membership of the borderline group is not consistent between stations, so a candidate who is ‘borderline’ in one station may be a good pass in another. Secondly, the size of the ‘borderline’ group within stations can vary substantially. Thirdly, the percentage of candidates failing a particular station can position the borderline group between (say) the 8th and 10th decile in terms checklist scores. Finally, the proportion of candidates within the borderline group who pass or fail a station (based on checklist score) is highly variable.

Taken together, these results show that across the whole assessment it is not possible to define a single ‘borderline’ group – in other words, case-specificity, which in known to exist at the cohort level (Eva et al. 1998; Norman et al. 2006; Mattick et al. 2008), is also important with regard to the key group of candidates in the pass/fail critical region. This specificity makes it difficult to generalise about the borderline candidate’s performance from assessed areas of the curriculum that appear in the OSCE to non-assessed areas, further problematizing the description of the hypothetical borderline candidate in standard setting methods. We argue that this finding provides important evidence in favour of a sequential testing model of assessment (Emons et al. 2007; Tweed & Wilkinson 2010; Cookson et al. 2011; Pell et al. 2013; Wainer & Feinberg 2015), where weaker candidates are given a supplementary test covering a wider range of task/constructs than those included in the initial (‘screening’) test. The key ‘borderline’ group is not consistent in their performance, and hence we need to sample more widely from the universe of potential OSCE stations to improve diagnostic accuracy in this group, ensuring better decision making for them – and this is precisely what a sequential model of assessment provides. Taking this argument a little further, the high degree of case-specificity witnessed in this work also underlines the importance of bespoke remediation programmes for those students who fail and have to repeat the year under a sequential assessment regime (Pell et al. 2013).

Implications for standard setting

An important corollary of these analyses is that the variability across stations in this metric raises concerns over methods which set the pass mark assuming a given distribution of marks within the 'borderline' group. Such methods include the well-known and established borderline groups' method (Livingston & Zieky 1982; Kramer et al. 2003; Wood et al. 2006; Cizek & Bunch 2007, p. 112–113)[1] which, by taking an average checklist score amongst those 'borderline', essentially assumes that an equal number of candidates in the 'borderline' group will pass as will fail in a station. In other words, this method assumes that the central location of the scoring distribution is the most appropriate estimate for the cut-score for the station. We have shown that there is considerable variation in the proportion passing/failing (Table 1, column 3), which implies that the average (usually median) checklist score of the borderline group is not necessarily a good estimate of the appropriate 'borderline' cut-score in the station. In fact, the median percentage of the borderline group who pass across the set of 13 stations in the 2015 assessment is 44% (i.e. less than half). This implies that the borderline group aggregate pass mark across the assessment would tend to be estimated too highly in this case, as the borderline group method assumes this figure to be actually 50% in each of the stations.

Another standard setting method is the more recently advocated objective borderline method (Shulruf et al. 2013; Shulruf et al. 2015), which assumes a strong dependency between global grades and checklist scores in a station. It assumes a monotonic relationship – that those with pass grades always score at least as highly on the checklist than those with borderline grades. The results from this work clearly show that this assumption does not commonly hold in real OSCE assessment outcome data (Table 2 shows a single station, but several other stations show a similar pattern), and hence this casts doubt on the validity of cut-scores calculated by this method, certainly when applied in the context of our own OSCE data.

A final comment in respect of implications for standard setting is that high variability of the borderline group within an entire assessment makes Angoff-type judgements (Cizek & Bunch 2007, chap. 6; McKinley & Norcini 2014) more conceptually challenging for assessors. Evidence from this paper shows that a consistent borderline group does not exist across the assessment, which implies that making judgements about borderline level performance for individual items/station level is likely to be problematic. Our arguments here complement other work that indicates problems with Angoff-type standard setting methods in knowledge tests (Clauser et al. 2009; Margolis & Clauser 2014; Margolis et al. 2016).

In respect of the borderline regression method, (Kramer et al. 2003; Pell & Roberts 2006) all these issues are far less acute since each cut-score calculation uses all encounters between candidates and assessors, not just those in or around the 'borderline', however defined. This is a very powerful advantage of this method of standard setting in OSCEs over other 'borderline' or methods.

Another important consequence of this work is that compensation across stations should be limited in these high stakes settings (Cizek & Bunch 2007, p. 20–22). Our findings show that there is high case-specificity for borderline candidates, and allowing unlimited compensation of checklist marks across stations would not necessarily guarantee the desired level of allround competence that is justifiably required by licensing bodies and all stakeholders. These findings therefore support the need for conjunctive standards, to limit excessive compensation, in addition to candidates achieving the overall cut score (Ben-David 2000; Cizek 2012, p. 36–38). These might include requiring a minimum proportion of station passes across the exam – whilst individual station pass/fail decisions have relatively low reliability, across the exam as a whole such a standard has satisfactory reliability. An alternative to this is to have minimal requirements for individual domains, although in this approach there is the risk of lowering reliability across a relatively small numbers of stations (McKinley & Norcini 2014). If such additional conjunctive requirements are not included, it is very likely that candidates weak in several important areas could pass based on sufficiently strong compensatory performance elsewhere in the assessment.

The examiner centre/site differences in the relationship between checklist marks and global grades we witness (Figure 1) suggest that assessor decision-making is not a purely objective process, but can depend on the culture, environment and team that they are working in. In other words, it is context specific as Goaverts (2016) and others have recently discussed (Sadler 2009; Yorke 2011a; Yorke 2011b; St-Onge et al. 2015; Chahine et al. 2015). This could have important implications for the equivalence of standards in national assessments carried out at different geographical locations.

Issues of assessment quality

One final point merits reiteration – namely that a single metric alone (e.g. a measure of reliability) is never sufficient for measuring the ‘quality’ of an assessment when making a high stakes pass/fail decision (Pell et al. 2010; Pell et al. 2015). We emphasise that the work we have presented here is based on ‘good’ quality OSCEs with reasonable station level metrics. Investigating the patterns and relationships between assessment outcomes (e.g.

global grades and checklist scores) is always a revealing, and sometimes a worrying, process but is one that always gives useful and important insights into the assessment.

Study limitations

This is a study within a single UK medical school with a particular programme of assessment in place, including reliance on a specific method of standard-setting (i.e. borderline regression). Although this paper investigates a single cohort for a single year-group, other unpublished work provides confidence that the findings are, however, representative of other cohorts and other years within our institution. A key approach to developing this research further would be a multi-school project to explore wider applicability beyond one institution, and the authors would welcome additional research collaboration to explore these issues.

Conclusion

In conclusion, we argue that examining bodies and institutions must take into account the high level of variability of borderline candidates when designing their assessment systems and passing rubrics. Using simple classical methods, our study shows that the borderline group is a complex and varying concept across elements of the assessment, and that this has important assessment policy implications. Developing our understanding of what is really happening at the borderline level in performance assessments should be a key focus in improving diagnostic accuracy, and in better understanding assessor decision making - ultimately leading to safer and more valid decisions with regard to student progression. Ongoing work aims to develop this understanding further using latent variable methods (Templin & Jiao 2012; Boscardin 2012; Homer et al. 2015) to classify candidates into groups based on their performance profile across an OSCE. It is hoped that such approaches will provide additional empirical evidence concerning any apparent ‘borderline’ group across the assessment as a whole.

Practice points

- The borderline group, however defined, is inconsistent in membership (i.e. shows high levels of case-specificity) when comparing across stations and across the entire assessment.
- For candidates in the overall pass/fail critical region, more information is required to infer competence in comparison to better performing candidates. Hence, a sequential testing model of assessment is likely to provide greater ‘diagnostic accuracy’, since such candidates are tested on a wider range of the curriculum than under traditional models of assessment.
- Different proportions of the borderline group pass and fail when comparing across stations. Hence, standard setting methods that assume a pre-defined distribution for this group could be prone to inaccuracy.

Notes on contributors

Matt Homer BSc, MSc, PhD, CStat, is an Associate Professor, working in both the Schools of Medicine and Education. His medical education research focuses on psychometrics and assessment quality, particularly related to OSCEs and knowledge tests.

Godfrey Pell, BEng, MSc, C.Stat, C.Sci, is 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.

Richard Fuller, MA, MBChB, FRCP, FAcadMed is a consultant physician, Professor of Medical Education and Director of the undergraduate degree programme at Leeds Institute of Medical Education. His research interests focus on the ‘personalisation’ of assessment, to support individual learner journeys, through 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.

Declarations of interest

The authors report no declarations of interest.

References

Ben-David MF. 2000. AMEE Guide No. 18: Standard setting in student assessment. Med Teach. 22:120–130.

Boscardin CK. 2012. Profiling students for remediation using latent class analysis. Adv Health Sci Educ Theory Pract. 17:55–63.

Boursicot KAM, Roberts TE, Pell G. 2007. Using borderline methods to compare passing standards for OSCEs at graduation across three medical schools. Med Educ. 41:1024–1031.

Chahine S, Holmes B, Kowalewski Z. 2015. In the minds of OSCE examiners: uncovering hidden assumptions. Adv Health Sci Educ. 21:609–625.

Cizek GJ, editor. 2012. Setting Performance Standards: Foundations, Methods, and Innovations. 2 edition. New York: Routledge.

Cizek GJ, Bunch MB. 2007. Standard setting a guide to establishing and evaluating performance standards on tests [Internet]. Thousand Oaks, Calif.: Sage Publications; [cited 2013 Aug 21]. Available from: http://SRMO.sagepub.com/view/standard-setting/SAGE.xml

Clauser BE, Mee J, Baldwin SG, Margolis MJ, Dillon GF. 2009. Judges’ Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. J Educ Meas. 46:390–407.

Cookson J, Crossley J, Fagan G, McKendree J, Mohsen A. 2011. A final clinical examination using a sequential design to improve cost-effectiveness. Med Educ. 45:741–747.

Emons WHM, Sijtsma K, Meijer RR. 2007. On the consistency of individual classification using short scales. Psychol Methods. 12:105–120.

Eva KW, Neville AJ, Norman GR. 1998. Exploring the etiology of content specificity: factors influencing analogic transfer and problem solving. Acad Med J Assoc Am Med Coll. 73:S1-5.

Farmer EA, Page G. 2005. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 39:1188–1194.

Fuller R, Homer M, Pell G. 2013. Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. Med Teach. 35:515–517.

Govaerts MJB. 2016. Competence in assessment: beyond cognition. Med Educ. 50:502–504.

Hays R, Gupta TS, Veitch J. 2008. The practical value of the standard error of measurement in borderline pass/fail decisions. Med Educ. 42:810–815.

Homer M, Pell G, Fuller R. 2015. Identifying “borderline” students in performance assessments – what does the empirical evidence say?

Homer M, Pell G, Fuller R, Patterson J. 2016. Quantifying error in OSCE standard setting for varying cohort sizes: A resampling approach to measuring assessment quality. Med Teach. 38:181–188.

Kramer A, Muijtjens A, Jansen K, Düsman H, Tan L, Van Der Vleuten C. 2003. Comparison of a rational and an empirical standard setting procedure for an OSCE. Med Educ. 37:132–139.

Livingston SA, Zieky MJ. 1982. Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. [Internet]. [cited 2015 Mar 19]. Available from: http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED227113

Margolis MJ, Clauser BE. 2014. The Impact of Examinee Performance Information on Judges’ Cut Scores in Modified Angoff Standard-Setting Exercises. Educ Meas Issues Pract. 33:15–22.

Margolis MJ, Mee J, Clauser BE, Winward M, Clauser JC. 2016. Effect of Content Knowledge on Angoff-Style Standard Setting Judgments. Educ Meas Issues Pract [Internet]. [cited 2016 Feb 15]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/emip.12104/abstract

Mattick K, Dennis I, Bradley P, Bligh J. 2008. Content specificity: is it the full story? Statistical modelling of a clinical skills examination. Med Educ. 42:589–599.

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

Norcini JJ. 2003. Setting standards on educational tests. Med Educ. 37:464–469.

Norman G, Bordage G, Page G, Keane D. 2006. How specific is case specificity? Med Educ. 40:618–623.

Pell G, Fuller R, Homer M, Roberts T. 2010. How to measure the quality of the OSCE: A review of metrics - AMEE guide no. 49. Med Teach. 32:802–811.

Pell G, Fuller R, Homer M, Roberts T. 2013. Advancing the objective structured clinical examination: sequential testing in theory and practice. Med Educ. 47:569–577.

Pell G, Homer M, Fuller R. 2015. Investigating disparity between global grades and checklist scores in OSCEs. Med Teach. 37:1106–1113.

Pell G, Roberts TE. 2006. Setting Standards for Student Assessment. Int J Res Method Educ. 29:91–103.

Reznick RK, Blackmore D, Dauphinée WD, Rothman AI, Smee S. 1996. Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. Acad Med J Assoc Am Med Coll. 71:S19-21.

Sadler DR. 2009. Indeterminacy in the use of preset criteria for assessment and grading. Assess Eval High Educ. 34:159–179.

Shulruf B, Poole P, Jones P, Wilkinson T. 2015. The Objective Borderline Method: a probabilistic method for standard setting. Assess Eval High Educ. 40:420–438.

Shulruf B, Turner R, Poole P, Wilkinson T. 2013. The Objective Borderline method (OBM): a probability-based model for setting up an objective pass/fail cut-off score in medical programme assessments. Adv Health Sci Educ. 18:231–244.

St-Onge C, Chamberland M, Lévesque A, Varpio L. 2015. Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. Adv Health Sci Educ. 21:627–642.

Templin J, Jiao H. 2012. Applying model-based approaches to identifying Performance Categories. In: Cizek, Gregory J, editor. Setting Perform Stand Found Methods Innov. 2 edition. New York: Routledge; p. 379–398.

Tweed M, Wilkinson TJ. 2010. Should there be borderline candidates or should there be a zone of uncertainty around assessment decisions? Med Teach. 32:869.

Wainer H, Feinberg R. 2015. For want of a nail: Why unnecessarily long tests may be impeding the progress of Western civilisation. Significance. 12:16–21.

Wimmers PF, Fung C-C. 2008. The impact of case specificity and generalisable skills on clinical performance: a correlated traits-correlated methods approach. Med Educ. 42:580–588.

Wimmers PF, Splinter TAW, Hancock GR, Schmidt HG. 2007. Clinical Competence: General Ability or Case-specific? Adv Health Sci Educ. 12:299–314.

Wood TJ, Humphrey-Murto SM, Norman GR. 2006. Standard Setting in a Small Scale OSCE: A Comparison of the Modified Borderline-Group Method and the Borderline Regression Method. Adv Health Sci Educ. 11:115–122.

Yorke M. 2011a. Assessing the complexity of professional achievement. In: Jackson N, editor. Learning to be professional through a higher education. London: Sceptre.

Yorke M. 2011b. Summative assessment: dealing with the “measurement fallacy.” Stud High Educ. 36:251–273.

Figure

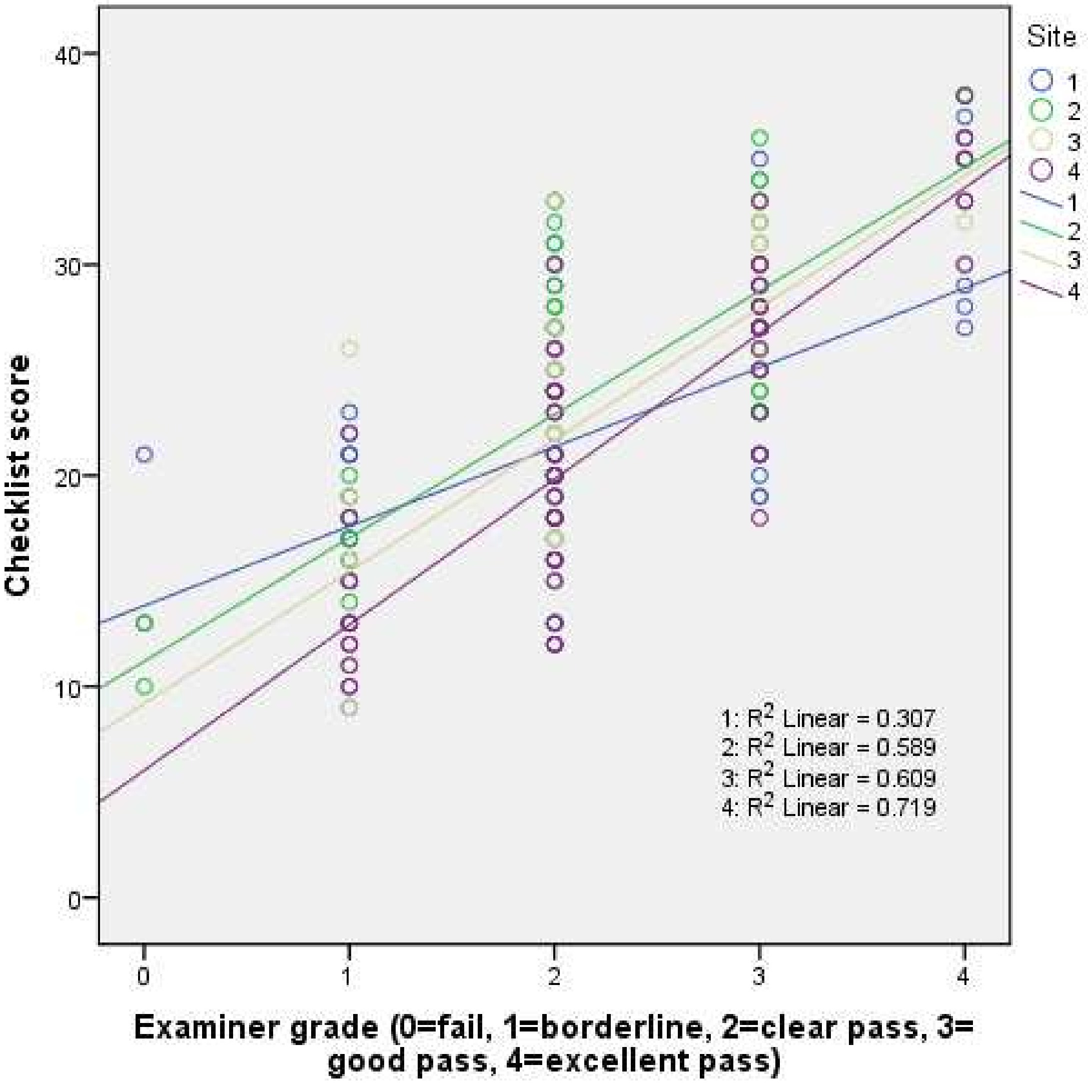

Figure 1: Variation in passing scores by site (station 12)

Tables

Table 1

Station	Percentage with borderline global grade within each station	Percentage of borderline (grade) group who pass each station	Percentage of all candidates failing the station
1	20	39	19
2	19	43	16
3	26	41	20
4	16	52	18
5	16	38	21
6	14	56	13
7	13	42	17
8	9	56	12
9	11	50	10
10	22	49	21
11	14	49	14
12	14	44	20
13	22	35	22

Table 1: Station-level metrics related to borderline candidates

Table 2

		Station global grade					**Total number of candidates**
		Fail	**Borderline**	**Clear pass**	**Good pass**	**Excellent pass**	
Station decile based on checklist score	**1**	0	0	2	9	15	**26**
	2	0	0	5	18	5	**28**
	3	0	0	16	17	2	**35**
	4	0	2	14	9	1	**26**
	5	0	2	15	7	0	**24**
	6	0	2	34	6	0	**42**
	7	0	3	16	0	0	**19**
	8	0	8	13	3	0	**24**
	9	2	15	15	2	0	**34**
	10	9	13	3	0	0	**25**
Total number of candidates		**11**	**45**	**133**	**71**	**23**	**283**

Table 2: Within-station checklist decile versus global grade (station 5).

Table 3

		Overall decile based on total OSCE score										Total
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
Station decile based on checklist score	**1**	7	4	2	6	2	2	0	2	1	0	**26**
	2	7	4	4	5	5	1	1	0	0	1	**28**
	3	4	3	8	7	7	3	2	1	0	0	**35**
	4	2	2	1	2	2	4	4	5	3	1	**26**
	5	0	4	4	3	2	4	2	3	0	2	**24**
	6	6	4	3	1	3	4	6	4	5	6	**42**
	7	2	2	1	0	3	2	3	4	1	1	**19**
	8	1	1	2	4	1	3	3	1	6	2	**24**
	9	0	3	2	0	1	5	4	5	8	6	**34**
	10	0	0	1	1	3	1	2	4	4	9	**25**
Total		**29**	**27**	**28**	**29**	**29**	**29**	**27**	**29**	**28**	**28**	**283**

Table 3: Within-station checklist decile versus overall decile (station 5).