Article:
Pell, G, Homer, MS and Fuller, R (2015) Investigating disparity between global grades and checklist scores in OSCEs. Medical Teacher. ISSN 0142-159X

https://doi.org/10.3109/0142159X.2015.1009425

Title: Investigating disparity between global grades and checklist scores in OSCEs

Short title: Disparity in global and checklist OSCE scores

Names of authors

Godfrey Pell, Matt Homer, Richard Fuller

Names of institution

Leeds Institute of Medical Education, University of Leeds

Corresponding author

Godfrey Pell, LIME, School of Medicine, University of Leeds, Leeds LS2 9JT
Telephone: +44 (0) 113 3434378, Fax: +44 (0) 113 3434375, Email: g.pell@leeds.ac.uk

Abstract

Background

When measuring assessment quality, increasing focus is placed on the value of station level metrics in the detection and remediation of problems in the assessment.

Aims

This paper investigates how disparity between checklist scores and global grades in an OSCE can provide powerful new insights at the station level whenever such disparities occur, and develops metrics to indicate when this is a problem.

Method

This retrospective study uses OSCE data from multiple examinations to investigate the extent to which these new measurement of disparity complements existing station level metrics.

Results

In stations where existing metrics are poor, the new metrics provide greater understanding of the underlying sources of error. Equally importantly, stations of apparently satisfactory 'quality' based on traditional metrics are shown to sometimes have problems of their own – with a tendency for checklist score 'performance' to be judged stronger than would be expected from the global grades awarded.

Conclusions

There is an ongoing tension in OSCE assessment between global holistic judgements and the necessarily more reductionist, but arguably more objective, checklist scores. This paper develops methods to quantify the disparity between these judgements, and illustrates how such analyses can inform ongoing improvement in station quality.

Practice points

- OSCEs are well-established in many institutions and this implies that the investigation of quality issues requires a focussed approach, and the impact of particular interventions should be quantifiable.
- Substantial misalignment between checklist marks and global grades indicates quality issues at the station level.
- Existing metrics cannot always identify such problems: this paper provides a method for quantifying this misalignment that can be modified to suite local conditions
- Within the borderline group (i.e. those with 'borderline' grades), a high proportion passing the station can indicate a lack of understanding of the station's objectives on the part of assessors (similarly with a low proportion).

Notes on contributors

GODFREY PELL, BEng, MSc, FRSS, C.Stat, C.Sci, is the senior statistician at Leeds Institute of Medical Education, who has a strong background in management. His current research focuses on quality within the OSCE, including theoretical and practical applications. He acts as an assessment consultant to a number of medical schools.

MATT HOMER, BSc, MSc, PhD, CStat is a senior researcher and teacher at the University of Leeds, working in both the Schools of Medicine and Education. His research generally has a quantitative methodological focus, and within medical education relates to evaluating and improving assessment quality, standard setting and psychometrics.

RICHARD FULLER, MA, MBChB, FRCP, is a consultant physician and Director of the Leeds MBChB undergraduate degree programme at Leeds Institute of Medical Education. His research interests focus on monitoring and improving the quality of assessment at programme levels, with particular focus on performance assessment.

Introduction

Measuring assessment quality is an essential component of high stakes assessment, encompassing a range of institutional activities that includes the selection of appropriate test formats, blueprinting content, item design, extensive post-hoc analysis of candidate performance data and the use of this data to justify assessment decisions and further improve assessment (Hays, Gupta, & Veitch, 2008; Pell, Fuller, Homer, & Roberts, 2010). The use of Objective Structured Clinical Examinations (OSCEs) as a key form of performance assessment has gained widespread academic support as a fair, robust test format that has seen considerable research-informed development (Newble, 2004; Pell et al., 2010). OSCEs have clear strengths, especially when appropriate standard setting methodologies are employed, allowing careful specification of content, standardisation and an opportunity to undertake extensive measurement and post hoc analysis to determine assessment quality. Measurements of reliability are routinely used as an element of determining assessment quality (Streiner & Norman, 2003, Chapter 8), with an increasing focus on the value of station level metrics in the detection and remediation of a range of problems with OSCE formats (Fuller, Homer, & Pell, 2013; Pell et al., 2010).

In 'traditional' test formats, OSCEs have two assessment outcomes within each station, a checklist score and a global grade (other formats of the OSCE have seen a move away from checklists, for example, in the USA's main licencing exam, the history taking checklist has been eliminated due to concerns regarding its poor discrimination). The alignment between the checklist/marking scheme score and overall global grade within a station is an important characteristic, and one would expect that in a high quality station (i.e. one that is working well as part of a reliable and valid assessment) this alignment should be strong. However, it is possible that there are candidates who have a strong checklist score, but whose global grades are poor. Conversely, there may well be candidates for whom checklist marks are low but for whom the global grade is good. The degree to which these two performance

measures are aligned, or more importantly misaligned, in this critical borderline area remains under-researched. A number of studies have looked at checklist discrepancies and/or rating discrepancies (Boulet, McKinley, Whelan, & Hambleton, 2003), but to our knowledge none has investigated discrepancies *between* checklist scores and global ratings in a station and what this might mean. A poor correlation between the two outcomes would indicate problems in a station. As well as poor station design and associated support materials, these problems might also include inadequate assessor training, poor behaviour by individual or groups of assessors etc. We show empirically in the results section that an acceptable correlation can allow misalignment issues concerning pass/fail decisions to exist but go unrecognised. Any degree of misalignment increases the likelihood of failing 'competent' students or passing 'incompetent' students at the station level. This is the key issue investigated in this paper.

The ability to undertake detailed post hoc analysis to explore the extent of any misalignment provides additional insight into error variance at station level and challenges assumptions about the true nature of standardisation in OSCEs (Newble, 2004). At the same time, research in these areas is complemented by a growing body of literature that seeks to understand the complex area of assessor decision making in performance testing (Govaerts, 2011; Kogan, Conforti, Bernabeo, Iobst, & Holmboe, 2011; Sadler, 2009; Yorke, 2011). Employing constructivist views of assessment, this literature reveals that the factors affecting assessor decision-making can be highly individualised, contextualised and influenced by characteristics such as assessor experience and seniority (Kogan et al., 2011; Pell & Roberts, 2006). This can be summarised as a complex interaction of test format design issues, construct, assessor behaviours and candidate performances within the OSCE environment, sometimes described as a 'black box' of variance (Gingerich, Regehr, & Eva, 2011; Kogan et al., 2011). Where misalignment occurs, our work seeks additional understanding of this 'black box' with regard to this error variance.

There is a growing dissatisfaction with 'traditional' (i.e. reductionist) checklist marking schedules both in healthcare and wider education (Sadler, 2009), with an accompanying growth in the use of global/domain based marking schema, supported by work that indicates that global grades are more reliable than checklist scores (Cohen, Colliver, Robbs, & Swartz, 1996; Regehr, MacRae, Reznick, & Szalay, 1998). It is important to note that some of the misalignment between scores and grades in a station can reflect poor checklist design, and that these two performance 'scores' may measure quite different traits. If candidate performance is measured on global grades alone, understanding of error variance is likely to be diminished, with an inability to undertake more detailed triangulation with checklist outcomes. This approach poses the real possibility that assessments take place without an ability to investigate concerns about the nature of variance in marks, implying that error in assessor judgements may be more likely to go unrecognised. Ideally, a structure of OSCE design and analysis should be implemented that generates a range of station level metrics (derived from appropriate marking schedules) that then allows for comprehensive data triangulation. Such an approach facilitates a more in-depth evaluation of the quality of the assessment, allowing exploration of the complex dynamics of performance assessments.

In exploring the concept of misalignment, this paper addresses these issues through a better understanding of the relationship at the station level between global grades and checklist scores, with a detailed investigation of the degree of misalignment between checklist-based pass/fail decisions and assessor global grades. This work also examines the 'directionality' of this misalignment, namely whether assessors judge more students to have passed or failed, based on global grades, than are realised by the checklist score which, under borderline methods of standard setting, determine pass/fail decisions at the station level. Investigating the level of disagreement allows the development of additional metrics to measure and explore this misalignment, including directionality, further. Additionally, we investigate the proportions of students who pass and fail each station having been assigned a global borderline grade. One might reasonably expect that these proportions are

approximately equal in well-designed stations (in essence this is the assumption behind the borderline group method of standard setting (Ben-David, 2000). Hence, when this is not the case, it will be shown that this provides additional evidence with regard to quality issues in such stations.

In summary, this paper adds to the range of post hoc quality station-level metrics which can be generated and, together with existing post hoc metrics, provides a greater understanding of error variance in stations. This in turn facilitates an enhanced focus in quality improvement at the station level. We emphasise that these metrics should be treated as diagnostic, and institutions should bear in mind that a station with poorly performing metrics can affect pass/fail decisions for the entire assessment (Fuller, Homer, & Pell, 2011).

Methods

Initial exemplification and exploration

We use OSCE data from multiple cohorts and different undergraduate year groups in a single UK Medical School to investigate the analysis of agreement between global 'decisions' and checklist decisions. The construct of our OSCEs, and the level and extent of post hoc analysis of station level metrics has been described in detail previously (Pell et al., 2010). Our OSCE format uses *year-specific* global grade descriptors (indexed as clear fail, borderline and three passing grades) alongside a specific marking schema that develops from a traditional checklist format in our junior OSCEs (third year) to a sophisticated 'key features' format in the final, qualifying OSCE (fifth year). There is a comprehensive training program for assessors which comprises group workshops, on-line refresher courses, and on-the-day pre-examination station run-throughs involving assessors from parallel circuits.

Within each station, a pass mark is calculated from all the grades/marks within the station using the Borderline Regression Method (Kramer et al., 2003; Pell & Roberts, 2006). This method of standard setting generates a range of new metrics which allows us to inspect in detail the degree of alignment of grade and checklist. To exemplify the development of our method, we analyse recent OSCE station level data from a cohort of 234 fourth year medical students (on a five year program undergraduate program). Table 1 shows candidate performance described by global grade and by the decision based on checklist performance under the borderline regression method in an *individual* station from this OSCE.

		Global grade					
		Clear fail	**Borderline**	**Clear pass**	**Very good pass**	**Excellent pass**	**Total**
Checklist decision based on set standard	**Pass**	0	26	105	46	15	**192**
	Fail	7	22	9	0	0	**38**
Total		**7**	**48**	**114**	**46**	**15**	**230**

Table 1: Pass/fail decisions for a single station versus global grades awarded

The first useful statistical measures are the overall failure rate at the station (38/230=16.5%) and the percentage in the 'Borderline' grade (48/234=20.9%), the latter of which is arguably high since in one in five encounters the assessors are unable to make a clear pass/fail global decision. The methods we develop will enable us to examine the congruence between the two assessor judgements: global grades and checklist marks, particularly within borderline categories.

In a typical station, one would usually expect more borderline students to pass than fail since these students are located towards the left tail of the scoring distribution, but to the right of the failing students (compare the 26 to the 22 in the Borderline column of Table 1). In this station (Table 1) the proportion of obviously 'misclassified' students of either type (i.e. global 'clear pass' but low checklist score, and vice versa) is shown in the shaded cells and is arguably quite small (9/234=3.9%). The method will be further developed and then used to explore the extent to which misclassification affects OSCE stations more generally, thereby allowing discussion of the inferences that can be drawn from such analysis.

Treatment of 'Borderline' grades

Ideally, to aid analysis we would like to allocate the 48 students who were awarded a 'Borderline' grade in this station (Table 1) to either a pass or fail overall 'grade'. We have investigated a number of ways of doing this (not reported in detail here): for example, equal apportioning of borderline students to passing and failing, apportioning these based on frequencies of neighbourhood grades, as well as carrying out an analysis based on excluding the borderline students. These analyses indicate that the most effective approach is to classify all 'Borderline' students as 'Not clear pass', which then allows us to collapse Table 1 to a simplified format in Table 2:

		Global 'decision'		
		At least 'Clear pass'	**Not 'Clear pass'**	**Total**
Checklist decision based on set standard	**Pass**	166	26 (b)	**192**
	Fail	9 (c)	29	**38**
Total		**175**	**55**	**230**

Table 2: Pass/Not pass classifications for a single station

We can now make more sense of the degree of misclassification in this station, as represented by the two shaded cells in the off-diagonal of Table 2. We have 15.2% (=35/230) in the off-diagonal, and ratio of 26:9 between these two cells (i.e. approximately 3:1). Nine candidates (cell 'c') has been awarded a global 'clear pass' by an assessor, but failed to achieve sufficient checklist marks, whereas 26 students achieve satisfactory checklist marks but are awarded poor global grades by assessors (cell 'b'). These misclassifications can also be seen in the scatter graph Figure 1 (bottom right and top left sections respectively).

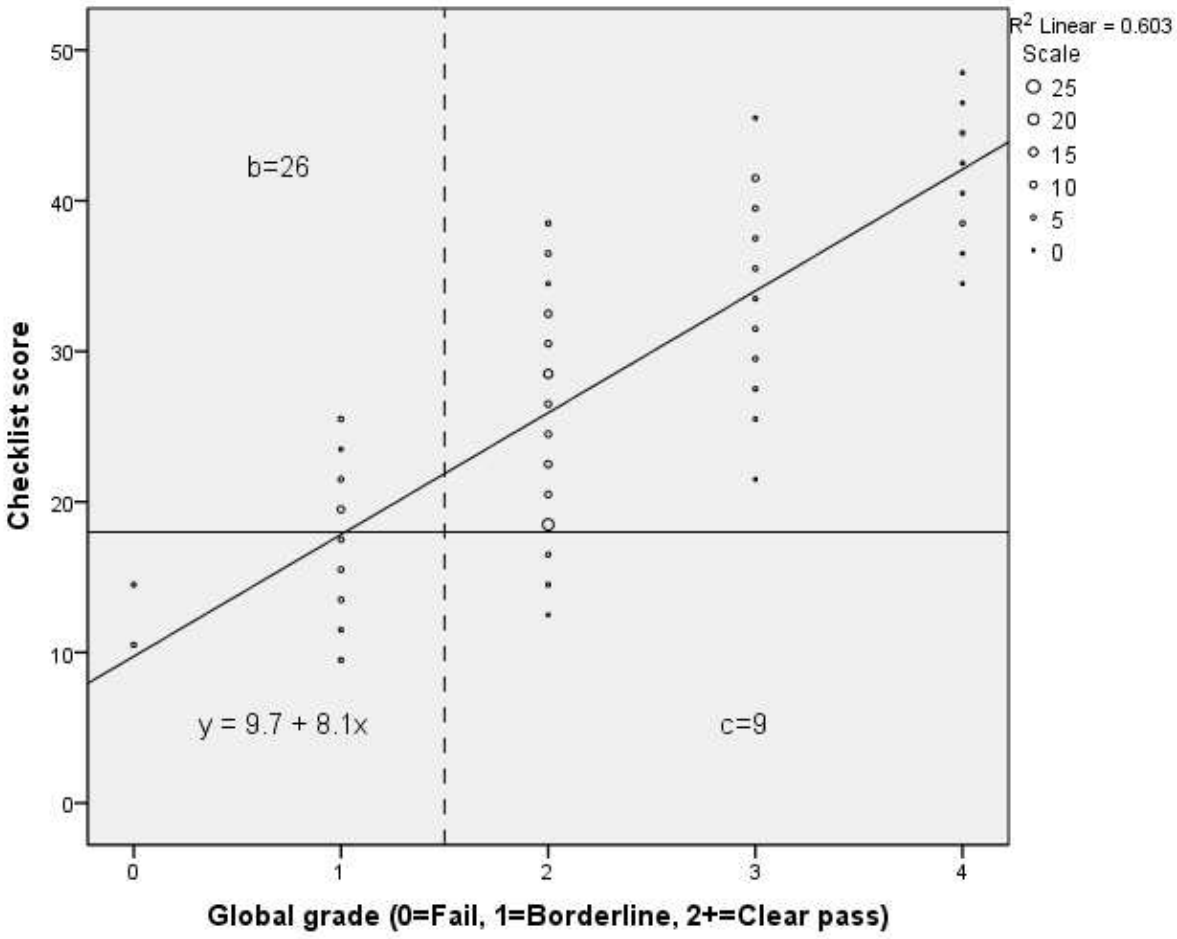

Figure 1: Scattergraph of checklist marks versus global grades

The latter asymmetry (c=9, b=26) indicates that assessors are unsure about awarding passing grades to students whose checklist performance is actually good enough to pass the station overall. This phenomenon often reflects a complex dynamic of variance factors – where issues about candidate performance (e.g. safe practice or professional behaviour) are not appropriately represented in the checklist, but are instead revealed through assessors' overall decisions as reflected in the global grade. We will examine different examples of misclassification later in the results section.

Formulating measures of misclassification

One of the key areas of research in this study is to explore the possibility of developing useful metrics to quantify the degree of misalignment that is present in Table 2, as a step towards highlighting stations that require further investigation.

We wish to test whether or not there is a 'significant' difference between two dichotomous 'pass/fail' measures on the same subjects (see Table 2). The McNemar test (Field, 2013, p. 555) is designed to assess precisely whether such a difference in paired nominal data is statistically significant. This test focuses on the shaded elements in the off-diagonal of Table 2 using the following formula:

$$\chi^2 = \frac{(\boldsymbol{b}-\boldsymbol{c})^2}{(\boldsymbol{b}+\boldsymbol{c})}$$

For the station illustrated in Table 2, this gives a McNemar value of 8.3. Under the null hypothesis of no difference between the two measures, this follows a chi-square distribution with one degree of freedom, so that the critical value is 3.84 for a standard 5% test of no difference between the two measures, suggesting a significant level of misclassification.

We would obviously expect a strong association between the global and checklist decisions, and whilst the McNemar test is sensitive to *absolute* differences between b and c, it is not sensitive to the actual *proportion* of students amongst the whole cohort present in the off-diagonal, thus not capturing the effect size of the missclassification. To overcome this limitation, we propose a modification that additionally takes account of the off-diagonal proportion in the cohort. For convenience purposes, we have labelled this the 'Pell-McNemar' (PM) misclassification measure:

$$\text{PM} = \frac{(b-c)^2}{(b+c)} \times \frac{(b+c)}{MN} = \frac{(b-c)^2}{MN}$$

$$\text{i.e. PM} = \frac{(b-c)^2}{MN} \qquad \text{(formula 1)}$$

Here N is the total number of students in the assessment, and M is a judgement of the proportion misclassified in the cohort that is considered reasonable. Given the multiple factors that contribute to assessor variance, we have modelled a range of values for M, showing that a misclassification rate of 10% (i.e. M=0.1) is a useful threshold value for similar assessment programmes with a stable assessor population. Newer programmes, or those which have yet to acquire a stable pool of trained and experienced assessors, might model an adjusted range of this value, presumably higher than 10%. Whilst the PM statistic should not be used for formal significance testing, our empirical work indicates that as a diagnostic tool a PM value of ≥ 4 highlights a need for further investigation of underlying reasons for misclassification in the station.

The values of the McNemar and PM statistics for the data in Table 2 are 8.3 and 12.6 respectively. So on both measures we conclude that there is a 'significant' misalignment between the checklist and global grade pass/fail 'decisions' and hence that further investigation of the performance of this station should be considered; for example, reviewing the congruence between checklist and assessor support material.

Applying this approach, but for the borderline group only, we also calculate a McNemar and a PM measure as per the formulae above but with the equivalent of b and c values from the borderline group only (and in the on-diagonals zeroes). So, b is the number of borderline students passing, and c is the number failing. High values of these measures suggest that in awarding a global grade, the assessors might use additional candidate behaviours which are

not captured in the checklist. This can work to either the student's advantage or disadvantage in their global grade and will affect the station pass mark

Across a selection of stations taken from a number of OSCE assessments, these measures will now be explored in more detail in the Results section. Our detailed analysis of OSCE assessments over the last four years indicates that in approximately 15-20% of stations there is either substantial disparity between global grades and checklist scores, or substantial asymmetry in pass/fail outcomes in the borderline group, or both.

Results – Application in Practice

Having developed the model, we now apply it to OSCE data from two successive final (fifth year undergraduate) year qualifying examinations to consider our new metrics alongside some well-known pre-existing metrics (Pell et al., 2010). Our final year stations use a 'key features' approach to checklist marking, forming a hybrid between a traditional checklist and global/domain grades alone. We use this data to illustrate how an analysis of 'disparity' provides additional insight into OSCE quality, complementing existing forms of post-hoc analysis.

From a composite selection of stations from two OSCEs assessments, Table 3 illustrates the array of station level metrics now available to assist interpretation as part of measuring OSCE quality. We have used Station 6 in this table to highlight the issues of misclassification (Tables 1 and 2) in the earlier methods section. Table 3 combines a 'standard set of metrics' (columns 2-5) as previously described in the literature (Pell et al., 2010), whilst the remaining columns report the new metrics that investigate levels of misclassification. We have selected a number of stations from across both final level OSCE

examinations to demonstrate how the new misclassification metrics provide a greater level of interpretation.

For each station, we have also focused on the misclassification in the borderline group in detail (last three columns of Table 3).

Station	Pre-existing quality metrics				Pass/fail grid		Disparity measures	Disparity tests	Borderline group only		
	R Square	Inter-grade Discrimination	Number of Failures (cohort size ~250-300)	% of Between-group variation	Checklist pass/Global clear pass	Checklist pass/Global not clear pass	% in the off-diagonal (i.e. mis-diagnosed)	McNemar	% of Total cohort	Number passing [1]	McNemar
					Checklist fail/Global clear pass	Checklist fail/Global not clear pass	Asymmetry in the off-diagonal	PM	% Passing	Number failing [2]	PM
1	0.53	6.0	36	27	213	23	12.9	3.5	16.5	23	0.0
					12	24	1.9	4.5	51.1	22	0.0
2	0.52	4.3	32	6	230	10	6.3	0.5	12.5	10	5.8
					7	25	1.4	0.3	29.4	24	7.2
3	0.42	2.6	41	20	223	8	12.1	8.8	7.0	5	4.3
					25	16	0.3	10.6	26.3	14	3.0
4	0.67	5.3	42	18	207	23	10.7	10.0	19.5	23	0.9
					6	36	3.8	10.6	44.4	30	1.8
5	0.51	4.8	42	24	179	9	11.3	2.5	13.9	8	8.0
					17	25	0.5	2.8	25.0	24	11.1
6	0.60	8.1	38	21	166	26	15.2	8.3	20.9	26	0.3
					9	29	2.9	12.6	54.2	22	0.7
7	0.60	5.6	29	17	175	26	14.8	9.5	17.0	25	3.1
					8	21	3.3	14.1	64.1	14	5.3
8	0.61	4.3	54	4	151	25	12.6	15.2	25.2	22	3.3
					4	50	6.3	19.2	37.9	36	8.5

Table 3: Full set of OSCE metrics for selected stations from two OSCE assessments

[1] The students counted in this cell are a subset of the students Checklist pass/Global not clear pass (i.e. of those in the top right cell of the Pass/fail grid)
[2] The students counted in this cell are a subset of the students Checklist fail/Global not clear pass (i.e. of those in the bottom right cell of the Pass/fail grid)

Stations with established problems based on existing station level metrics

As part of the validation of the new metrics, we examine their application where existing station level metrics already highlight concerns about quality. In Table 3, station 3 shows a poor r-squared value with an accompanying low value for the slope of the regression line (inter-grade discrimination), already suggesting that the station is not discriminating well between students based on ability. From this analysis, we would anticipate there to be a wide range of checklist marks for each global grade. The pass/fail grid reveals a high level of asymmetry in the off-diagonal with three times as many candidates (25:8) achieving a global pass grade from assessor but poor checklist marks compared to those not achieving a global pass whilst having good checklist marks.

Both the McNemar test and the PM measure indicate significant levels of misalignment in the two assessor judgements for this station. The borderline group is small (7%) and, alongside the evidence of misalignment, this indicates that whilst many assessors are making decisive global passing decisions in this station, these do not always concur with the checklist marks awarded. As a result of the small *proportion* of students judged to be borderline, the asymmetry within the borderline group is not judged important by the PM measure, although using the McNemar measure this would be significant. This contrasts with metrics in stations 7 and 8 (Table 3), which both have large numbers in the borderline group.

In summary, the additional use of the misalignment measures shows that in this station, assessors are almost certainly measuring different traits/attributes in their global grades compared to those specified in the checklist. This could be an example of the hawks and doves phenomenon, but occurring in just one of the two station measures (i.e. checklist scores or global grades). More detailed support material might be required as assessors

appear to be using differential assessment criteria when awarding the global grade that are not reflected in the checklist, perhaps based on personal experience.

Stations with no 'apparent' problems

We now examine stations where the 'standard' set of metrics would not highlight underlying quality issues in respect of assessor decision making and judgements. This analysis reveals a central theme: candidates as a whole achieve comparatively better performance (determined by the checklist score) than would be expected by assessors' prediction (determined by the global grade).

Station 1 provides a useful illustration of this theme. The station has 'acceptable' standard metrics based around a R square value >0.5, good inter-grade discrimination and a low number of failures. However, the pass/fail grid reveals that there are 12 candidates who failed based on key features checklist score, but who were awarded a clear pass global grade, and 23 candidates in the converse position. This misalignment exceeds our threshold value (PM=4.5 compared to the suggested cut-off of 4). Examining the borderline group in detail, we see that 16.5% of the cohort (i.e. 1 in 6) are in this group, and further that there are approximately equal numbers who actually pass or fail in the borderline group. This suggests that assessors are generally having difficulty in making global decisions. Note, this issue is not apparent in the pre-existing station level metrics. The relatively high proportion of the borderline group suggests a more widespread problem for assessors, which would merit investigation of the efficacy of existing assessor support material for this station.

Station 2 shows a similar pattern to station 1 in respect of misclassification, but with a smaller proportion of students (17 out of 272), meaning that both McNemar and PM

measures are not significant. However, a closer inspection shows that 12.5% of the cohort has been recorded as borderline by assessors. The directionality of the classification in this group shows that more students end up failing (24) than passing (10). This analysis suggests either (i) assessor dissatisfaction with the item checklist and a subsequent reluctance to award failing grades to those who ultimately do fail, or (ii) a tendency of assessors to over-rate candidate performance in the global grade, perhaps based on other observed aspects of a candidate's performance.

Stations with relatively high numbers of 'Borderline' students

As a final part of the work examining the impact of the misalignment measure, we review stations where the proportion of borderline grades awarded is relatively 'high'. Station 8, which focusses on medico-legal responsibilities after the death of a patient, has the highest proportion of borderline grades (25.2% of the whole cohort) amongst the stations listed in Table 3. Review of traditional station level metrics (columns 2 to 5) shows an acceptably performing station, but with a high number of student failures.

There is a high level of asymmetry in the direction of assessor (global) predictions, primarily focused on assessor overall concerns with students who are achieving adequate checklist marks (25 vs. 4). However, when we examine the borderline group only, the majority (36 out of 58) fail the station (with 22 passing) – suggesting that assessors are reluctant to award failing grades to failing students. This would point to assessor uncertainty with regard to the construct the station is intending to measure despite an established programme of assessor training.

Discussion

The continued drive towards authenticity and integration within clinical performance assessment has been accompanied by an increased awareness of the complex dynamic between station construct, candidate performance and assessor decision making, particularly within OSCE settings. At the heart of this dynamic exists a need for institutions to demonstrate the overall quality of the OSCE, and an established series of measurements are employed to calculate overall test reliability, station level quality metrics and increasingly, error variance estimates (Pell et al., 2010). This requirement has prompted further investigation into the multiple contributions to error and variance within the OSCE, allowing both recognition and improvement of structural, systematic and construct related factors (Fuller et al., 2013). Despite this, there remains a large degree of 'noise' in performance testing, recently conceptualised within workplace assessment settings as a 'black box' (Kogan et al., 2011).

Researchers have begun to unpack this 'noise', and work within healthcare assessment has focused on assessor behaviours and decision making in the complex, changing nature of the OSCE station (Govaerts, 2011). Work from other professional disciplines has highlighted a wider tension between the balance of global grades and checklists/marking rubrics, revealing active 'transgressions' as assessors trust of holistic, global judgements overrides their use of checklist criteria(Marshall, 2000). Other work reveals the challenges of using global grades and descriptors alone, as assessors seek to make sense of complex constructs such as 'safety' or 'professionalism' within descriptors. Such constructs are often represented by single words, and despite assessor training, multiple re-interpretations lead to variation in judgements – with some researchers conceptualising this as *anticipated* variance, rather than just simply error Govaerts, van der Vleuten, Schuwirth, & Muijtjens, 2007). This complex dynamic has been conceptualised through a series of observations as 'indeterminancy', challenging the theoretical background to the accepted use of checklists, rubrics and grading schemes (Sadler, 2009).

As OSCE developers, how can we 'bridge' the gap between our ability to measure variance factors and understand these contributory factors in the context of our own assessments and environments? Whilst an extensive literature exists in respect of the use of global grades and checklists within OSCEs (Cunnington, Neville, & Norman, 1996; Humphrey-Murto & MacFadyen, 2002; Wan et al., 2011), little has been done to explore the nature of the alignment between the two. The premise of our work is that such measures could help us understand the extent and nature of misalignment, particularly at the critical 'borderline' of pass/fail, and then better understand some of the 'quality' challenges facing us at the station level. We encourage other institutions to model the PM calculations using their own data to determine the most suitable values of the parameters M and N (formula 1) to meet local conditions.

This study has revealed the extent of misalignment between assessors' checklist decisions and their 'predictions' (i.e. the global grades) across a range of different academic cohorts and levels of assessment in a large scale OSCE. Within stations with 'adverse' standard station level metrics (Fuller et al., 2013; Pell et al., 2010), the misalignment measures complement these well, highlighting where assessors are dissatisfied with station and checklist constructs. As stated earlier in the introduction, the misalignment could be the result of a number of problems (including but not limited to, for example, assessor training, support materials, 'rogue' assessors and so on). Of more importance is the deeper insight into stations which might have been judged as 'acceptable' based on pre-existing metrics. The unsatisfactory characteristic of many of these stations lies in assessor judgement of the borderline group. Interpreting the misalignment measure in this context reveals different directionalities – with assessors showing difficulty in awarding fail grades and a tendency to over rate student performance in the borderline group. Such findings resonate with assessors awarding 'bestowed credit' – rewarding or penalising other candidate activities that are not featured within grading and checklist systems, and an activity that has been identified as a threat to the fidelity of performance assessments (Sadler, 2010).

We estimate that the incidence of substantial asymmetry of pass/fail outcomes within the borderline group occurs in approximately 10% of stations. In other words, there are incidences where the large majority of candidates in the borderline group pass the station (or, conversely, fail the station). Hence, the mean mark for this borderline group, giving the cut-score as per the borderline group method, is lower (or higher) than that under the borderline regression method. We would argue from a quality perspective that this is further evidence in favour of BRM, since under the borderline group method these issues would remain unknown.

A key limitation of our work is that it is undertaken in a single institution's assessment system, albeit with data spread over different stages of the course and over a number of cohorts. Although some of the approaches described might initially appear complex, we feel the methods used in this paper are quite easily replicated and make use of routinely available assessment data. We would be happy to correspond with those interested in replicating this work in a multi-centre study. This data sharing across institutions would allow further investigation of the generalisablity of the findings presented in this paper.

Finally, an on-going problem is that of the apparent 'cost' of such detailed, psychometric analyses within OSCE settings. However, previous work examining longitudinal improvements to the OSCE has highlighted the benefit of such in-depth approaches, particularly when headline whole-exam metrics are acceptable (Fuller et al., 2013). The application of these in-depth post hoc approaches reduces the likelihood of the confounding of multiple station-level problems when measuring the effect of change at the station level. We feel that this paper, whilst furthering the range of numerical/technical post hoc analysis of OSCE data, helps reframe the value of the psychometrics within the 'post-psychometric era' (Hodges, 2013) and provides a practical approach to exploring the impact of the complexities of 'how' assessors assess.

References

Ben-David, M. F. (2000). AMEE Guide No. 18: Standard setting in student assessment. *Medical Teacher*, *22*(2), 120–130. doi:10.1080/01421590078526

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). Quality Assurance Methods for Performance-Based Assessments. *Advances in Health Sciences Education*, *8*(1), 27–47. doi:10.1023/A:1022639521218

Cohen, D. S., Colliver, J. A., Robbs, R. S., & Swartz, M. H. (1996). A large-scale study of the reliabilities of checklist scores and ratings of interpersonal and communication skills evaluated on a standardized-patient examination. *Advances in Health Sciences Education*, *1*(3), 209–213. doi:10.1007/BF00162917

Cunnington, J. P. W., Neville, A. J., & Norman, G. R. (1996). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. *Advances in Health Sciences Education*, *1*(3), 227–233. doi:10.1007/BF00162920

Field, A. P. (2013). *Discovering statistics using IBM SPSS statistics: and sex and drugs and rock "n" roll*. London: Sage Publications.

Fuller, R., Homer, M., & Pell, G. (2011). *What a difference an examiner makes! __Detection, impact and resolution of "rogue" examiner behaviour in high stakes OSCE assessments*. Presented at the AMEE, Vienna.

Fuller, R., Homer, M., & Pell, G. (2013). Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. *Medical Teacher*, *35*(6), 515–517. doi:10.3109/0142159X.2013.775415

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-Based Assessments as Social Judgments: Rethinking the Etiology of Rater Errors: *Academic Medicine*, *86*, S1–S7. doi:10.1097/ACM.0b013e31822a6cf8

Govaerts, M. J. B. (2011). *Climbing the Pyramid: Towards Understanding Performance Assessment*. Maastricht University.

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in-training assessment. *Advances in Health Sciences Education: Theory and Practice*, *12*(2), 239–260. doi:10.1007/s10459-006-9043-1

Hays, R., Gupta, T. S., & Veitch, J. (2008). The practical value of the standard error of measurement in borderline pass/fail decisions. *Medical Education*, *42*(8), 810–815. doi:10.1111/j.1365-2923.2008.03103.x

Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher*, *35*(7), 564–568. doi:10.3109/0142159X.2013.789134

Humphrey-Murto, S., & MacFadyen, J. C. (2002). Standard setting: A comparison of case-author and modified borderline-group methods in a small-scale OSCE. *Academic Medicine*, *77*(7), 729–732. doi:10.1097/00001888-200207000-00019

Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W., & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical Education*, *45*(10), 1048–1060. doi:10.1111/j.1365-2923.2011.04025.x

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. *Medical Education*, *37*(2), 132–139. doi:10.1046/j.1365-2923.2003.01429.x

Marshall, B. (2000). *English teachers: the unofficial guide : researching the philosophies of English teachers*. London; New York: Routledge.

Newble, D. (2004). Techniques for measuring clinical competence: objective structured clinical examinations. *Medical Education*, *38*(2), 199–203.

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. *Medical Teacher*, *32*(10), 802–811. doi:10.3109/0142159X.2010.507716

Pell, G., & Roberts, T. E. (2006). Setting Standards for Student Assessment. *International Journal of Research & Method in Education*, *29*(1), 91–103.

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic Medicine: Journal of the Association of American Medical Colleges*, *73*(9), 993–997.

Sadler, D. R. (2009). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, *34*(2), 159–179. doi:10.1080/02602930801956059

Sadler, D. R. (2010). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education*, *35*(6), 727–743. doi:10.1080/02602930902977756

Streiner, D., & Norman, G. (2003). *Health Measurement Scales: A practical guide to their development and use* (3rd ed.). OUP Oxford.

Wan, M., Canalese, R., Lam, L., Petersen, R., Quinlivan, J., & Frost, G. (2011). Comparison of criterion-based checklist scoring and global rating scales for the Objective Structured Clinical Examination (OSCE) in pre- clinical year medical students. *Medical Education*, *45*(S3). doi:10.1111/j.1365-2923.2011.04089.x

Yorke, M. (2011). Assessing the complexity of professional achievement. In *Learning to be professional through a higher education*. London: Sceptre.