Published article:
Pell, G, Fuller, R, Homer, M and Roberts, T (2010) *How to measure the quality of the OSCE: A review of metrics.* Medical Teacher, 32 (10). 802 - 811 . ISSN 0142-159X
http://dx.doi.org/10.3109/0142159X.2010.507716

How to measure the quality of the OSCE: a review of metrics

Author details

Godfrey Pell
Richard Fuller
Matthew Homer
Trudie Roberts

Medical Education Unit
Leeds Institute of Medical Education
Worsley Building
University of Leeds
Leeds
LS2 9JT

Corresponding author

Godfrey Pell
Principal Statistician
Leeds Institute of Medical Education
Worsley Building
University of Leeds
Leeds
LS2 9JT

Tel: +44 (0) 113 23434378
Fax: +44 (0) 113 23432597
Email: G.Pell@leeds.ac.uk

Notes on contributors

Godfrey Pell is a senior statistician who has a strong background in management. Before joining the University of Leeds he was with the Centre for Higher Education Practice at the Open University. Current research includes standard setting for practical assessment in higher education, and the value of short term interventionist programmes in literacy.

Dr Richard Fuller is a Consultant Geriatrician, as well as being Sub-Dean of the Faculty of Medicine and Health, and Chair of the MBChB at Leeds. His research interests include clinical assessment, in particular monitoring and improving the quality of the OSCE.

Dr Matthew Homer is a Research Fellow at the University of Leeds, working in the both the Schools of Medicine and Education. He works on a range of research projects and provides general statistical support to colleagues. His research interests include the statistical side of assessment, particularly related to OSCEs.

Trudie Roberts is a Consultant Physician, a Professor of Medical Education and is Director of the Leeds Institute of Medical Education. Her research interests include clinical assessment.

Practice points

- It is important to always evaluate the quality of a high stakes assessment, such as an OSCE, through the use of a range of appropriate metrics.

- When judging the quality of an OSCE, it is very important to employ more than one metric to gain an all round view of the assessment quality.

Abstract

With an increasing use of criterion based assessment techniques in both undergraduate and postgraduate healthcare programmes, there is a consequent need to ensure the quality and rigour of these assessments. The obvious question for those responsible for delivering assessment is how is this 'quality' measured, and what mechanisms might there be that allow improvements in assessment quality over time to be demonstrated? Whilst a small base of literature exists, few papers give more than one or two metrics as measures of quality in OSCEs.

In this guide, aimed at assessment practitioners, the authors aim to review the metrics that are available for measuring quality and indicate how a rounded picture of OSCE assessment quality may be constructed by using a variety of such measures, and also to consider which characteristics of the OSCE are appropriately judged by which measure(s). The authors will discuss the quality issues both at the individual station level and across the complete clinical assessment as a whole, using a series of ‘worked examples’ drawn from OSCE data sets from the authors’ institution.

Key words

OSCE, quality, CBA

Introduction

With increasing scrutiny of the techniques used to support high level decision making in academic disciplines, Criterion Based Assessment (CBA) delivers a reliable and structured methodological approach. As a competency-based methodology, CBA allows the delivery of ’high stakes’ summative assessment (e.g. qualifying level or degree level examinations), and the demonstration of high levels of *both* reliability and validity. This assessment methodology is attractive, with a number of key benefits over more ’traditional’ unstructured forms of assessment (e.g. *viva voce*) in that it is absolutist, carefully standardised for all candidates, and assessments are clearly designed and closely linked with performance objectives. These objectives can be clearly mapped against curricular outcomes, and where appropriate, standards laid down by regulatory and licensing bodies, that are available to students and teachers alike. As such, CBA methodology has seen a wide application beyond summative assessments, extending into the delivery of a variety of work-based assessment tools across a range of academic disciplines (Norcini 2007, Postgraduate Medical Education and Training Board 2009). CBA is also now being used in the UK in the recruitment of junior doctors, using a structured interview similar to that used for selecting admissions to higher education programmes (Eva 2004)

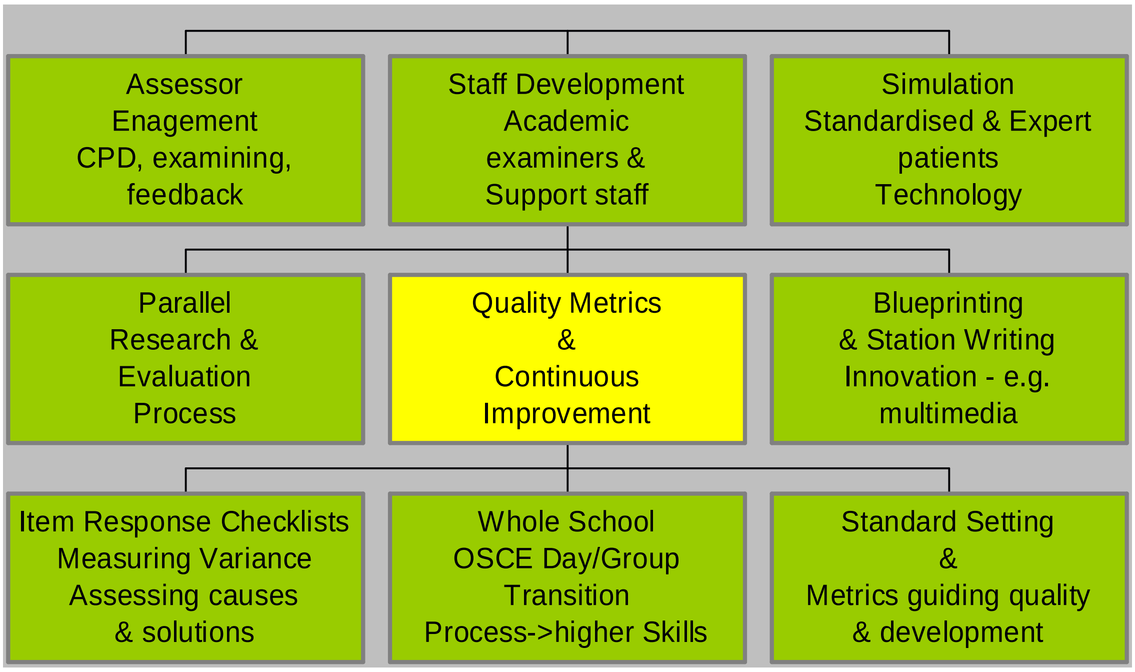
The Objective Structured Clinical Examination (OSCE) uses CBA principles within a complex process that begins with 'blueprinting' course content against pre-defined objectives (Newble 2004). The aim here is to ensure both that the 'correct' standard is assessed, and that the content of the OSCE is objectively mapped to curricular outcomes. Performance is scored, at the station level, using an item checklist, detailing individual (sequences of) behaviours, and by a global grade, reliant on a less deterministic overall assessment by examiners.[1] (Cohen 1997; Regehr 1998)

Central to the delivery of any successful CBA is the assurance of sufficient quality and robust standard setting, supported by a range of metrics that allow thoughtful consideration of the *performance of the assessment as a whole*, rather than just a narrow focus on candidate outcomes (Roberts 2006). 'Assessing the assessment' is vital, as the delivery of OSCEs are complex and resource intensive, usually involving large numbers of examiners, candidates, simulators and patients, and often taking place across parallel sites. This complexity means CBA may be subject to difficulties with standardisation, and is heavily reliant on assessor behaviour, even given the controlling mechanism of item checklists. No single metric is sufficient in itself to meaningfully judge the quality of the assessment process, just as no single assessment is sufficient in judging, for example, the clinical competence of an undergraduate student. Understanding and utilising metrics effectively are therefore central to CBA, both in measuring quality and in directing resources to appropriate further research and development of the assessment (Wass 2001). This guide will examine the metrics available, using final year OSCE results from recent years as exemplars of how exactly these metrics can be employed to measure the quality of the assessment. It is designed to support assessment quality at "local" medical school level, as applied to individual OSCE constructs and data generation. However, it should be made clear that looking at the OSCE metrics is only part of the overall process of reviewing OSCE quality – see Figure 1 for an overview of the total process.

---

[1] The global grades in our OSCEs are recorded numerically as 0=Clear fail, 1=Borderline, 2=Clear pass, 3=Very good pass, 4=Excellent pass.

Figure 1: Quality assessment is a complex process



For national assessment boards, the OSCEs are designed centrally to a common standard. However, at the local level with the assessment designed within specific medical schools, some variation, for example in station maxima, is possible depending upon the importance and complexity of the station to those setting the exam. These absolute differences between stations will adversely affect the reliability metric making the 0.9 value, often quoted, unobtainable. It is possible to standardise the OSCE data and thereby obtain a higher reliability metric but this would not be a true representation of the assessment as set with respect to the objectives of the assessing body. This guide is aimed primarily at those involved with clinical assessment at the local level within individual medical schools, where, although the assessment may take place across multiple sights, it is a single administration. Those involved with national clinical assessments are likely to have a different perspective.

Methods of standard setting

The method of standard setting will determine the metrics available for use in assessing quality. For example, with the borderline methods currently in favour, only the regression method will give some indication of the relationship between global grade and checklist score and also the level of discrimination between weaker and stronger students.

Table 1: Comparison of the borderline methods of standard setting

Borderline Groups (BLG)	Borderline Regression (BLR)
• Easy to compute • Only 3 global ratings required (fail, borderline, pass) • Only uses borderline data – a proportion of assessor/candidate interactions • Needs sufficient candidates in the borderline group (20+) • Produces limit quality assurance metrics	• Not so easy to compute • Usually requires 5 global ratings (fail, borderline, pass, credit, distinction) • Uses all assessor/candidate interactions • Does not need any borderline grade students • Produces a good variety of quality assurance metrics
• Also **Contrasting Groups**: similar characteristics to BLG	

The authors favour the borderline regression method because it uses all the assessment interactions between assessors and candidates, and also generates a wide range of metrics. One of the criticisms sometimes levelled at the borderline regression method is its possible sensitivity to outliers. These outliers occur in three main groups:

- Students who perform very badly and obtain a near zero checklist score.
- Students who achieve a creditable checklist score but who fails to impress the assessor overall.
- The assessor who gives the wrong overall grade.

These issues will be discussed in more detail at the appropriate points throughout the guide.

Assessing Assessment – General Overview

Table 2 details a 'standard' report of metrics from a typical OSCE (20 stations over two days, total testing time ~ three hours, spread over four examination centres). The borderline regression method was used for standard setting[2] (Pell & Roberts 2006). Typically such an OSCE will generate roughly 60,000 data items (i.e. individual student-level checklist marks), which form a valuable resource for allowing quality measurement and improvement. As a result of utilising such data, we have seen our own OSCEs deliver progressively more innovation, whilst simultaneously maintaining or improving the levels of reliability.

A selection of these overall summary metrics will be used in this guide to illustrate the use of psychometric data 'in action', and to outline approaches to identifying and managing unsatisfactory station-level assessment performance.

Table 2 Final year OSCE Metrics[3]

Station	Cronbach's alpha if item deleted	R^2	Inter-grade discrimination	Number of failures	Between-group variation (%)
1	0.745	0.4	4.21	53	31.1
2	0.742	0.5	5.23	24	30.1
3	0.738	0.5	5.14	39	33.0
4	0.742	0.5	4.38	39	28.0
5	0.732	0.5	4.14	29	20.5
6	0.750	0.4	4.74	43	40.3
7	0.739	0.5	4.51	36	19.5
8	0.749	0.4	3.45	39	33.8
9	0.744	0.5	4.06	30	36.0
10	0.747	0.5	3.91	26	29.9
11	0.744	0.5	4.68	37	37.6
12	0.744	0.5	2.80	23	32.3
13	0.746	0.6	3.99	30	22.0
14	0.746	0.6	5.27	54	27.3
15	0.739	0.5	3.49	44	25.9
16	0.737	0.5	3.46	41	34.3
17	0.753	0.5	3.58	49	46.5
18	0.745	0.5	2.42	15	25.4
19	0.749	0.4	3.22	52	39.5
20	0.754	0.5	4.50	37	34.1

[2] However, under any of the borderline methods of standard setting, where a global grade is awarded in addition to the checklist score, all of these metrics would be useful in measuring the quality of the assessments. For other types of standard setting, where such a global grade does not form part of the standard setting procedure (e.g. Ebel and Angoff, see Cusimano 1996) inter-grade discrimination and coefficient of determination (R^2) will not apply.

[3] The authors would like to point out that this is not the most recently available OSCE data, but is from an earlier year and was chosen specifically to demonstrate some of the quality issues that have arisen in OSCE assessments.

Number of candidates=241

Metric 1 - Cronbach's Alpha[4]

This is a measure of internal consistency (commonly, though not entirely accurately, thought of as 'reliability'), whereby in a good assessment the better students should do relatively well across the board (i.e. on the checklist scores at each station).[5] The (overall) value for alpha that is usually regarded as acceptable in this type of high stakes assessments, where standardised and real patients are used, and the individual station metrics are not standardised, is 0.7 or above. Where station metrics are standardised a higher alpha would be expected. Alpha for this set of stations was 0.754, and it can be seen (from the second column of Table 1) that no station detracted from the overall 'reliability', although stations 17 and 20 contributed little in this regard.

Since alpha tends to increase with the number of items in the assessment, the resulting *'alpha if item deleted'* scores should all be lower than the overall alpha score if the item/station has performed well. Where this is not the case this may be caused by the any of the following reasons:-

- The item is measuring a different construct to the rest of the set of items.
- The item is poorly designed.
- There are teaching issues – either the topic being tested has not been well taught, or has been taught to a different standard across different groups of candidates.
- The assessors are not assessing to a common standard.

In such circumstances, quality improvement should be undertaken by revisiting the performance of the station, and reviewing checklist and station design, or examining quality of teaching in the curriculum.

However, one cannot rely on alpha alone as a measure of the quality of an assessment. As we have indicated, if the number of items increases, so will alpha, and therefore a scale can be made to look more homogenous than it really is merely by being of sufficient length in terms of the number of items it contains. This means that if two scales measuring *distinct* constructs

---

[4] Where the data are dichotomous this is equivalent to the Kuder-Richardson KR20 coefficient.

[5] Strictly speaking, two forms of alpha can be calculated, a non-standardised and standardised version, both available in SPSS, with the former the default. We use the non-standardised one in this guide, a measure of mean intercorrelation weighted by variances, and which yields the same value as the G-coefficient for a simple model of items crossed with candidates.

are combined, to form a single long scale, this can result in a misleadingly high alpha. Furthermore, a set of items can have a high alpha and still be multidimensional. This happens when there are separate clusters of items (i.e. measuring separate dimensions) which intercorrelate highly, even though the clusters themselves do not correlate with each other particularly highly.

It is also possible for alpha to be too high (e.g. >0.9), possibly indicating redundancy in the assessment, whilst low alpha scores can sometimes be attributed to large differences in station mean scores rather than being the result of poorly designed stations.

We should point out that in the authors' medical school, and in many similar institutions throughout the UK, over 1000 assessors are required for the OSCE assessment season. Consequently, recruiting sufficient assessors of acceptable quality is a perennial issue, so it is not possible to implement double-marking arrangements that would then make the employment of G-theory worthwhile in terms of more accurately quantifying differences in assessors. Such types of analysis are more complex than those covered in this guide, and often require the use of additional, less user-friendly, software.

The hawks and dove effect, either within an individual station, or aggregated to significant site effects, may have the effect of inflating the alpha value. However, it is highly likely that this effect will lead to unsatisfactory metrics in the areas of coefficient of determination, between-group within-station error variance, and, possibly, in fixed effect site differences. Our philosophy is that one metric alone, including alpha, is always insufficient in judging quality, and that in the case of an OSCE with a high alpha but other poor metrics, this would not indicate a high quality assessment.

As an alternative measure to 'alpha if item is deleted', it is possible to use the correlation between station score and 'total score less station score'. This will give a more extended scale, but the datum value (i.e. correlation) between contributing to reliability and detracting from it is to some extent dependent on the assessment design and is therefore more difficult to interpret.

Metric 2 - Coefficient of Determination R^2

The R^2 coefficient allows us to determine the degree of (linear) correlation between the checklist score and the overall global rating at each station, with the expectation that higher overall global ratings should generally correspond with higher checklist scores.[6] The square root of the coefficient of determination is the simple Pearsonian correlation coefficient. SPSS and other statistical software packages also gives the adjusted value of R^2 which takes into account the sample size and the number of predictors in the model (one in this case); ideally this value should be close to the unadjusted value.[7]

A good correlation ($R^2 > 0.5$) will indicate a reasonable relationship between checklist scores and global grades, but care is needed to ensure that overly detailed global descriptors are not simply translated automatically by assessors into a corresponding checklist score, thereby artificially inflating R^2. In Table 2, station 14 (a practical and medico-legal skills station) has a good R^2 value of 0.697, implying that 69.7% of variation in the students' global ratings are accounted for by variation in their check list scores. In contrast, station 19 is less satisfactory with an R^2 value of 0.404. This was a new station focussing on patient safety and the management of a needlestick injury. To understand why R^2 was low, it is helpful to examine the relationship graphically (for example, using SPSS Curve estimation) to investigate the precise nature of the association between checklist and global grade - see Figure 2.[8] In this figure, assessor global grades are shown on the *x*-axis and the total item checklist score is plotted on the *y*-axis. Clustered checklist scores are indicated by the size of the black circle, as shown in the key. SPSS can calculate the R^2 coefficient for polynomials of different degree, and thereby provide additional information on the degree of linearity in the relationship.

In station 19 we can see that there are two main problems - a wide spread of marks for each global grade, and a very wide spread of marks for which the fail grade (0 on the x-axis) has been awarded. This indicates that some students have acquired many of the marks from the item checklist, but their overall performance has raised concerns in the assessor leading to a global fail grade.

In the introduction we raised the issue of the effects of outliers on the regression method. Examples of poor checklist scores but with reasonable grades can be observed in Figure 3. Not illustrated here, a frequent occurrence is students that score almost zero on the checklist score. This has the effect of reducing the value of the

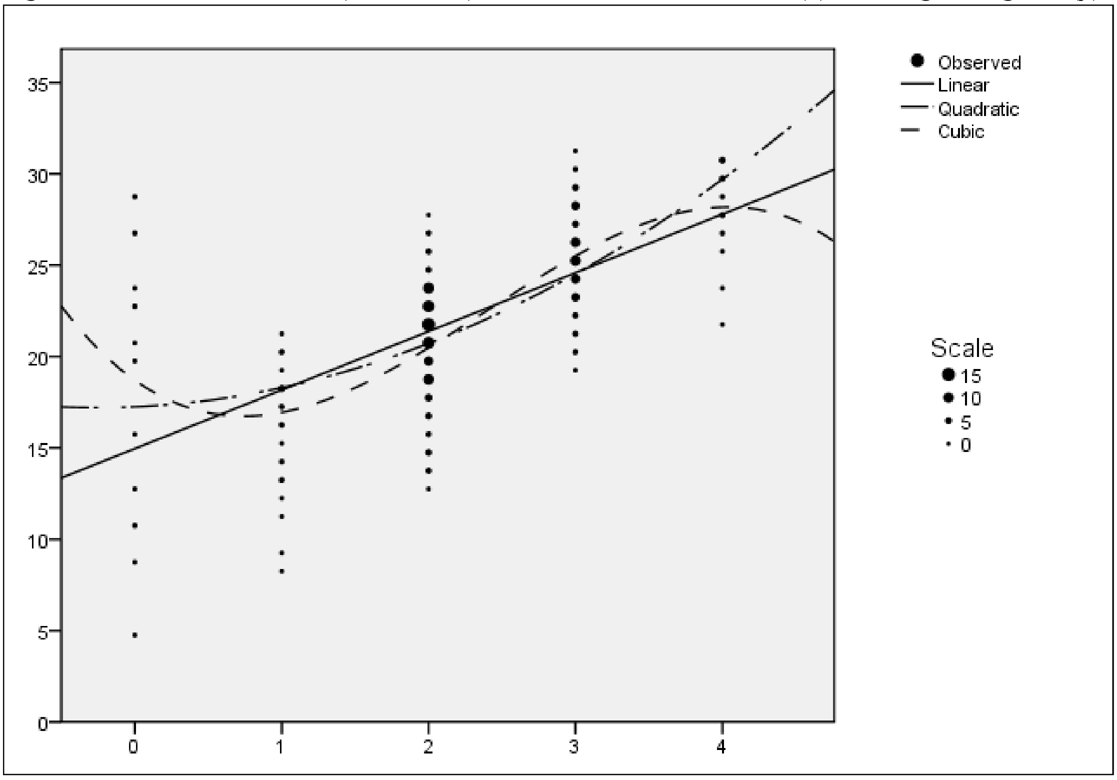
[6] R^2 is the proportional change in the dependent variable (checklist score) due to change in the independent variable (global grade)

[7] It is not necessarily straightforward to calculate adjusted R^2 as there is a choice of methods for doing so. For example, Field (2000) criticises SPSS for using Wherry's equation and agrees with Stevens (1992) that Stein's formula is more appropriate.

[8] In fact, we would recommend as good practice always plotting a scatter graph of checklist marks against global rating, regardless of station metrics.

regression intercept with the y-axis, and increasing the slope of the regression line. For the data indicated in Table 2, the authors removed outliers and recomputed the passing score and individual station pass marks. This made very little difference, increasing the passing score by less than 0.2%.

Figure 3 – Curve estimation (Station 19) – Assessor checklist score (x) versus global grade(y)



This unsatisfactory relationship between checklist marks and global ratings causes some degree of non-linearity, as demonstrated in the accompanying Table 3 (produced by SPSS), where it is clear graphically that the best fit is clearly cubic.[9]

Table 3 Curve estimation table (Station19)

Polynomial fitted	R Square	F	df1	df2	Sig.
Linear	.401	159.889	1	239	.000
Quadratic	.435	91.779	2	238	.000
Cubic	.470	70.083	3	237	.000

The existence of low R^2 values at certain stations and/or a wide spreads of marks for a given grade should prompt a review of the item checklist and station design. In this particular case, although there was intended to be a key emphasis on safe, effective management in the station, re-assessment of the checklist in light of these metrics showed this emphasis was not

[9] Note that mathematically speaking, a cubic will always producer a better fit, but parsimony dictates that the difference between the two fits has to be statistically significant for a higher order model to be preferred. In this example the fit of the cubic polynomial is significantly better than that of the linear.

well represented. It is clear that weaker candidates were able to acquire many marks for 'process' but did not fulfil the higher level expectations of the station (the focus on decision making). This has been resolved through a re-write of the station and the checklist, with plans for re-use of this station and subsequent analysis of performance within a future OSCE.

Metric 3 - Inter-grade discrimination

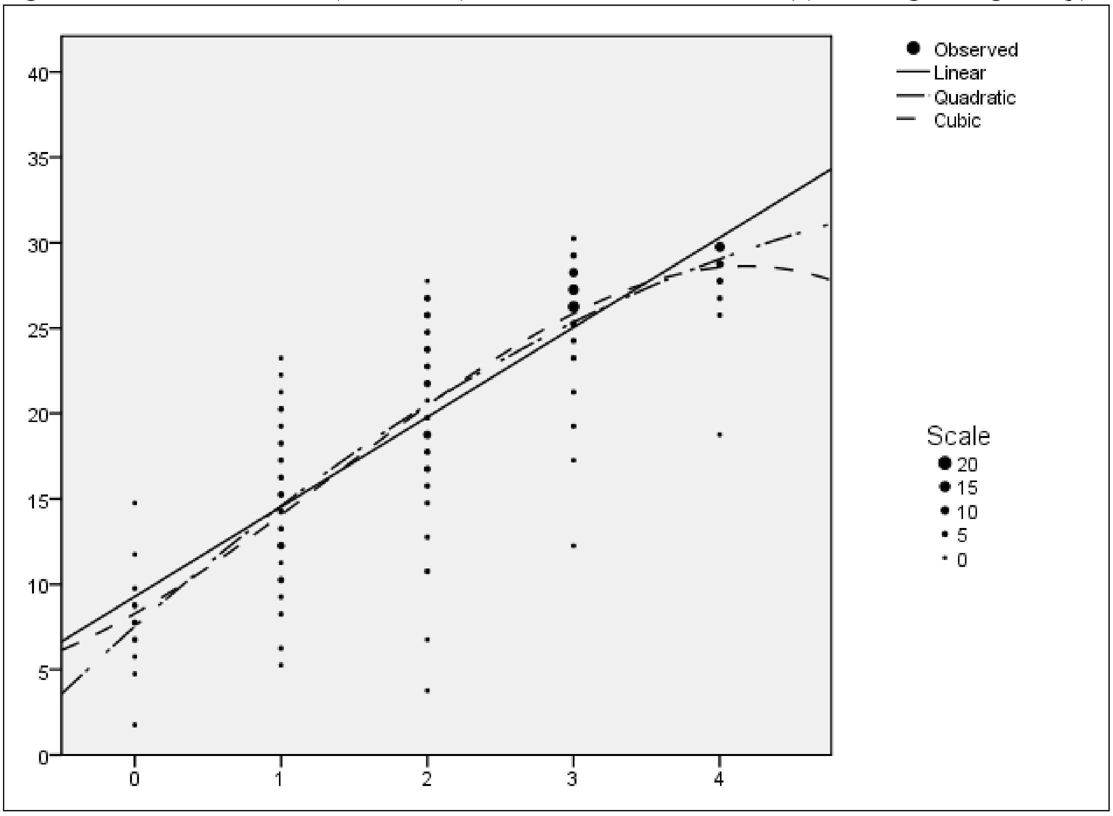
This statistic gives the slope of the regression line and indicates the average increase in checklist mark corresponding to an increase of one grade on the global rating scale. Although there is no clear guidance on 'ideal' values, we would recommend that this discrimination index should be of the order of a tenth of the maximum available checklist mark (which is typically 30-35 in our data).

A low value of inter-grade discrimination is often accompanied by other poor metrics for the station such as low values of R^2 (indicating a poor overall relationship between grade and checklist score), or high levels of assessor error variance (see metric 5 below) where assessors have failed to use a common standard. Too high levels of inter-grade discrimination may indicate either a very low pass mark, or a lack of linearity caused by a small number of badly failing students who tend to steepen the regression line. Where very poor student performance in terms of the checklist score occurs, consideration needs to be given to whether these very low scores should be excluded from standard setting to avoid excessive impact on overall passing scores in a downward direction.

Returning to Table 2, it is clear that the inter-grade discrimination values are generally acceptable across the stations (station maxima being in the region of 30-35 marks), although there are three stations with discrimination values in excess of 5 (e.g. station 14 – a skills station involving completion of a cremation form).

Where there is doubt about a station in terms of its performance based on the discrimination metric, returning to the R^2 measure of variance and curve estimation is often instructive. In Table 2, station 14 has the highest inter-grade discrimination, and it can be seen in Figure 3 that most global grades again encompass a wide range of marks, especially the '*clear pass'* grade - value 2 on the *x*-axis, ranging from 4 to 27, but that the lower of these values are clearly outliers. As the rest of the station metrics are acceptable, this station can remain unchanged but should be monitored carefully when used in subsequent assessments.

Figure 3 –Curve estimation (Station 14)- Assessor checklist score (x) versus global grade(y)



Metric 4 - Number of failures

It would be a mistake to automatically assume that an unusually high number of failures indicates a station that is somehow too difficult. The 'reality check'[10], which is an essential part of borderline methods, will to a large extent compensate for station difficulty.

As previously described, other psychometric data can be used to investigate station design and performance in order to identify problems. Failure rates may be used to review the impact of a change in teaching on a particular topic - with higher such rates indicating where a review of content and methods of teaching can help course design. There are no major outliers for this metric in Table 1, but the difficulties with station 19 have allowed us to identify and deliver additional teaching around elements of patient safety within the final year curriculum, and introduce this specific safety focus into checklists.

[10] i.e. in determining the global rating, assessors should bear in mind the expected performance level of the minimally competent student.

Metric 5 – Between-group variation (including assessor effects)

When performing analysis on data resulting from complex assessment arrangements such as OSCEs, where, by necessity, the students are subdivided into groups for practical purposes, it is vital that the design is fully randomised. Sometimes, however, this is not always possible, with logistical issues including dealing with special needs students who may require more time and have to be managed exclusively within a separate cycle. Any non-random subgroups must be excluded from statistically-based types of analysis that rely on randomness in the data as a key assumption.

In the ideal assessment process, *all* the variation in marks will be due to differences in student performance, and not due to differences in environment (e.g. local variations in layout or equipment), location (e.g. hospital based sites having different local policies for management of clinical conditions), or differences of assessor attitude (i.e. hawks and doves). There are two ways of measuring such effects, either by performing a one-way ANOVA on the station (e.g. with the assessor as a fixed effect), or by computing the proportion of total variance which is group specific. The latter allows an estimation of the proportion of variation in checklist scores that is due to student performance as distinct from other possible factors mentioned above, although this is usually given as the proportion of variance which is circuit specific.

If the variance components are computed, using group (i.e. circuit) as a random effect, then the percentage of variance specific to group can be computed. This is a very powerful metric as it gives a very good indication of the uniformity of the assessment process between groups. It is also relatively straightforward to calculate. Ideally between-group variance should be under 30%, and values over 40% should give cause for concern, indicating potential problems at the station level due to inconsistent assessor behaviour and / or other circuit specific characteristics, rather than student performance.

From Table 2, stations 6, 17 and 19 give cause for concern with regard to this metric, with the highest levels of between-group variance. In addition, station 6 has a poor R^2, and the overall combination of poor metrics at this station tells us that the poor R^2 was probably due to poor checklist design. These observations prompted a review of the design of station 6, and the checklist was found to consist of a large number of low level criteria where weaker candidates could attain high scores through ‘process’ only. In other words, there was a likely mismatch between the nature of the checklist, and the aims and objectives of the station as understood by the assessors. Hence, in redesigning the station, a number of the low-level criteria were chunked (that is, grouped together to form a higher level criterion) in order to facilitate the assessment of higher level processes as originally intended.

Station 17 tells a different story, as the good R^2 coupled with the high between-group variation indicates that assessors are marking consistently within groups, but that there is a distinct hawks and doves effect between groups. In such a case, this ought to be further investigated by undertaking a one-way ANOVA analysis to determine whether this is an individual assessor or a site phenomenon. The amount of variance attributable to different sites is subsumed in the simple computation of within-station between-group variance as describe above. However, its significance may be determined using a one-way ANOVA analysis with sites as fixed effects.

However, care needs to be exercised in making judgements based on a single metric, since, with quite large populations, applying ANOVA to individual stations is likely to reveal at least one significant result (as a result of a type I error due multiple significance tests across a large number of groups[11]). Careful post-hoc analysis will indicate any significant hawks and doves effects, and specific groups should be tracked across other stations to determine general levels of performance. If a completely random assessment model of both students and assessors has been used (mindful of the caveats about local variations in equipment and exam set up), then many of these effects should be largely self-cancelling; it is in the aggregate totals that group-specific fixed effects are important and may require remedial action.

Metric 6 – Between group variance (other effects)

ANOVA analysis can also be of use when there are non-random allocations of either assessors or students, as is the case in some medical schools with large cohorts and associated teaching hospitals where multi-site assessment may occur. Such complex arrangements can result in the non-random assignment of assessors to circuits since it is often difficult for clinical staff to leave their place of work. This may then lead to significant differences due to 'site effects' which can be identified with appropriate action taken in the analysis of results.

Other important fixed effects can also be identified through use of ANOVA. For example, assessor training effects, staff / student gender effects, and associated interactions, which have all been previously described (Pell 2008), and which underline the need for complete and enhanced assessor training as previously highlighted (Holmboe 2004).

[11] In this case cohorts of approximately 250 students per annum, and of the order of 15 parallel circuits.

Metric 7 – Standardised Patients

Most centres that use SPs require them to rate candidates. In keeping with other metrics, a higher than normal proportion of candidates (say over 10%) receiving adverse SP ratings may indicate problems. If this is coupled with a higher than normal failure rate it could be the result of inadequate teaching of the topic. Adverse values of this metric are often accompanied by high rates of between group variance; assessors viewing candidates exhibiting a lower than expected level of competence often have difficulty in achieving consistency.[12]

The overall reliability of the assessment may be increased by adding the SP rating to the checklist score; typically the SP rating should contribute 10-20% of the total station score. (Homer & Pell 2009). An alternative approach, taken within our own institution at graduating level OSCEs, is to set a 'minimum' requirement for SP comments as a proxy for patient satisfaction (using rigorously trained SPs).

Quality control by observation: detecting problems in the run up to OSCEs and on the day

It is informative for those concerned with minimising error variance between groups, to observe the OSCE assessment systematically. As a result, the following causes of between-group error can often be anticipated, observed and corrected:-

- Different layouts at different circuits - for example, the precise placing of disinfectant gel outside a station so that the assessor may not be able to score hand hygiene approaches.
- Assessors (or assistants) not returning equipment to the start or neutral position as candidates change over.
- Unauthorised prompting by assessors (despite training and pre-exam briefings).
- Inappropriate behaviour by assessors (e.g. changing the 'tone' of a station through excessive interaction).

[12] Typically SPs are asked a question such as "Would you like to consult again with this doctor?" with responses:- strongly agree, agree, neither agree nor disagree, disagree, strongly disagree; the two latter responses being regarded as adverse

- Assessors who arrive late and miss the pre-assessment briefing and who therefore fail to adhere adequately to the prescribed methodology.
- Excessively proactive SPs whose questions act as prompts to the students.
- Biased real patients (e.g. gender or race bias). Simulated patients receive training on how to interact with the candidates, but this is not possible with the majority of real patients.

Post hoc remedial action

Even with bad metrics, it is very unlikely that any medical school would call back students for re-assessment. Therefore in such cases action needs to be taken to ensure that the assessment decisions are defensible, and equitable towards students. Typically, fairly simple methods of remediation would include:

1. Adjustment of total marks for site effects

   The easiest method is to adjust to a common mean across all sites. After any such adjustment, the site profile of failing students should be checked to ensure that, for example, all failures are not confined to a single site. The effect of any special needs group located within a single specific site needs to be discounted when computing the adjustment level.
2. Adjustment at the station level

   This is seldom necessary because any adverse effects will tend to cancel each other out. In the rare cases where this does not happen, a station level procedure as in 1. above should be carried out.
3. Removal of a station

   Again, this is a rare event and the criteria for this are usually multiple adverse metrics, the result of which would disadvantage students to such an extent that the assessment decisions are indefensible against appeal.

Conclusion

Using a series of worked examples and 'live data', this guide has focussed on commonly used OSCE metrics and how they can be used to identify and manage problems, and how such an approach helps to anticipate future issues. This methodology therefore naturally feeds into the wider assessment processes – for example, station design, assessor training and so on.

In the authors' institution there is a close relationship between those who analyse the data, and those who design and administer the clinical assessments and develop/deliver teaching. Revealing mismatches between checklists and global ratings has lead to the redesign of certain OSCE stations with a subsequent improvement of metrics. Some of these redesigns include:

- Chunking of a number of simple criteria into fewer criteria of higher level.
- Chunking to allow for more higher level criteria commensurate with the stage of student progression.
- The inclusion of intermediate grade descriptors on the assessor checklists.
- Ensuring that the majority of checklist criteria have three instead of two anchors thereby allowing greater discrimination by assessors.
- A greater degree of uniformity between the physical arrangements of the different circuits.

The presence of high failure rates at particular stations has lead to a revisiting of the teaching of specific parts of the curriculum, and was followed by changes in the way things were taught, resulting in improved student performance as measured in subsequent OSCEs.

Indications of poor agreement between assessors has, on occasion, lead to a number of changes all of which have been beneficial to the quality of assessment:

- The provision of more detailed support material for assessors.
- Improved assessor briefings prior to the assessment.
- Improved SP briefings prior to the assessment.
- Dummy runs before the formal assessment for both assessors and SPs (this is only really practicable where students numbers are relatively small e.g. resits, and in dental OSCEs with smaller cohorts of students).
- Upgrading of assessor training methods.
- Updating ('refreshing') assessors who were trained some time ago.

The need for all the above improvements would be unlikely to have been apparent from using a single reliability metric, such as Cronbach's alpha or the G Coefficient[13]. It is only when a

[13] For a simple model where there is no double marking, the G Coefficient and the un-standardised alpha have the same value. In the majority of UK medical schools good assessors are in short supply, it is not possible to provide the level of double marking of students which would favour the use of G theory.

family of metrics is used that a true picture of quality can be obtained and the deficient areas identified. Adopting this approach will be rewarded with a steady improvement in the standard of clinical assessment as we have seen locally.

References

NORCINI, J., BURCH, V. (2007) Workplace-based assessment as an educational tool: AMEE guide No. 31. *Med Teacher* 29(9), pp. 855-871

POSTGRADUATE MEDICAL EDUCATION TRAINING AND EDUCATION BOARD (2009) Workplace based assessment. A guide for Implementation (London). www.pmetb.org.uk/fileadmin/user/QA/assessment/PMETB_WPBA_Guide_20090501.pdf (accessed May 11th 2009)

EVA, K.W., ROSENFELD, J.., REITER H., NORMAN G.R. (2004). An Admissions OSCE: the multiple mini-interview. *Medical Education* 38, pp. 314-326

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

COHEN, D.S., COLLIVER, J.A., ROBBS, R.S., SWARTZ M.H. (1997). 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, pp. 209-213

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* 73(9), pp. 993-997

ROBERTS, C., NEWBLE, D., JOLLY, B., REED, M., HAMPTON, K. (2006). Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher* 28(6), pp. 535-543

WASS, V., MCGIBBON, D., VAN DER VLEUTEN, C. (2001). Composite undergraduate clinical examinations: how should the components be combined to maximise reliability? *Medical Education* 35, pp. 326-330

CUSIMANO, M. (1996). Standard setting in Medical Education. *Academic Medicine*, Vol. 71, No. 10, S112-S120.

PELL, G., ROBERTS, T.E. (2006). Setting standards for student assessment. *International Journal of Research & Method in Education* 29(1), pp. 91-103

HOMER, M. AND PELL, G. (2009). The impact of the inclusion of simulated patient ratings on the reliability of OSCE assessments under the borderline regression method. *Medical Teacher*, 31 (5), 420 - 425.

PELL, G., HOMER, M., ROBERTS, T.E. (2008). Assessor Training: Its Effects on Criterion Based Assessment in a Medical Context. *International Journal of Research & Method in Education* 31(2), pp.143-154

HOLMBOE E. (2004). Faculty and the observation of trainees' clinical skills: Problems and opportunities. *Academic Medicine* 79(1), pp. 16-22

FIELD, A. (2000). *Discovering Statistics (using SPSS for windows*), p. 130 (Sage Publications, London)

STEVENS, J. (1992) *Applied multivariate statistics for the social* sciences (2nd ed), Chapter 4 pp.151-182 (Erlbaum, Hillside NJ)