
Article:
Fuller, R, Homer, MS orcid.org/0000-0002-1161-5938, Pell, G et al. (1 more author) (2017) Managing extremes of assessor judgement within the OSCE. Medical Teacher, 39 (1). pp. 58-66. ISSN 0142-159X

https://doi.org/10.1080/0142159X.2016.1230189

Rogue or rational? Investigating the impact of the extremes of assessor judgement within the OSCE.

Abstract

Context

There is a growing body of research investigating assessor judgements in complex performance environments such as OSCE examinations. Post-hoc analysis can be employed to identify some elements of 'unwanted' assessor variance. However, the impact of individual, apparently 'extreme' assessors on OSCE quality, assessment outcomes and pass/fail decisions has not been previously explored, This paper uses a range of "case studies" as examples to illustrate the impact that 'extreme' examiners can have in OSCEs, and gives pragmatic suggestions to successfully alleviating problems.

Method & Results

We used real OSCE assessment data from a number of examinations where at station level, a single examiner assesses student performance using a global grade and a key features checklist.

Three exemplar case studies where initial post hoc analysis has indicated problematic individual assessor behaviour are considered and discussed in detail, highlighting both the impact of individual examiner behaviour and station design on subsequent judgments.

Conclusions

In complex assessment environments, institutions have a duty to maximise the defensibility, quality and validity of the assessment process. A key element of this involves critical analysis, through a range of approaches, of assessor judgments. However, care must be taken when assuming that apparent aberrant examiner behaviour is automatically just that.

Introduction

High stakes testing of clinical performance is a cornerstone of programmes of healthcare assessment, with the Objective Structured Clinical Examination (OSCE) representing the principal test format across a spectrum of learners from novices to senior postgraduates. The last 40 years have seen considerable development of the OSCE format from its first description (Harden et al 1975), but its original strength, namely the sampling of multiple examiner-candidate interactions across a range of authentic clinical tasks, remains a critical part of its ongoing success (Patricio et al 2009; Harden et al 2015).

The design of the OSCE has traditionally been viewed as advantageous in nullifying individual behavioural judgement effects seen in other test formats through its use of multiple assessor-candidate interactions across stations. However, OSCEs can be prone to high levels of error variance due to inadequate sampling, poor station design and individual differences in assessor decision making, often perhaps simplistically characterised as 'hawks and doves' (Crossley et al 2002). This is usually reconciled, to an extent, by randomisation of assessors and students, balancing out variations in individual assessor judgement across an overall OSCE circuit or test cycle (McManus et al 2006; Harasym et al 2008).

Approaches to addressing assessor variation have been threefold – namely faculty development with specific OSCE training delivered face-face or online (Holmboe et al 2004; Pell et al 2008; Gormley et al 2012), increasing sophistication of post-test analyses (Pell et al 2010; Tavakol & Dennick 2012) and station level enhancement with measurable psychometric improvements (Fuller et al 2013). Sophisticated post-hoc analyses of the OSCE now allow us to identify variance due to non-student factors including those related to the delivery of the OSCE (e.g. OSCEs delivered across multiple centres, timing of parallel examination cycles), scoring rubrics and assessor judgements at overall test level, including the identification of 'extreme' hawks and doves assessor behaviours (Pell et al 2010; Bartman et al 2013).

However, an emergent body of work situated in workplace performance assessment formats has placed the effect of *individual* assessor-candidate interactions and judgments as central in understanding apparent variation in candidate outcomes (Govaerts 2007; Kogan et al 2011). This work has explored the complex, dynamic interaction of performance testing, and challenged some of our assumptions that assessors are 'trainable', either because of inherent, often subconscious traits and biases that mean assessors are 'fallible', or through use of constructivist paradigms illustrating that differing assessor judgements reflect *meaningful* idiosyncrasy (Gingerich et al 2014). The constructivist approach to learning suggests that knowledge evolves continuously after being 'constructed' (Brooks & Brooks 1993), and the role of the assessor is richer and multifaceted (Harden & Crosby 2000). A consensus view would then be that observations of performance assessments are undertaken through multiple, individual lenses, with assessor judgements reflecting a rich tapestry of what is 'experienced' during the assessment, interwoven with the context in which assessors and candidates undertake practice (Govaerts 2016). What assessors can 'measure' with scoring algorithms (whether checklist or global grade) typically fails to capture the rich synopsis of the assessor's experience of the encounter (Kogan et al 2011). This research has led to a persuasive argument that variance in scores amongst assessors

usually reflects legitimate, experience based and contextualised judgements, rather than what might previously been described as ‘error’ or ‘extreme judgments’ (Govaerts et al 2013).

How then should we reconcile these constructivist and psychometric approaches within practical OSCE settings, and to understand the effect of apparently extreme individual assessor judgments on overall candidate outcomes? In other words, to what extent are these examiners truly extreme and contributing to ‘error’ in assessment and/or in fact acting systematically within themselves but differently from their peers? Both of these types of assessor behaviour will affect the defensibility of overall assessment decisions and being able to better differentiate such assessor judgements should help decisions in respect of further interventions (e.g. statistical adjustment of data, targeted examiner training, appropriate examiner feedback relative to their peers). This provides an exploration of the extent to which this variance meaningfully impacts on individual candidates, overall assessment quality and also the setting of appropriate standards.

We hypothesize that this categorisation of extreme assessor behaviour will help the move from ‘mechanistic’ applications of psychometrics [e.g. adjusting all scores systematically to ‘neutralise’ the assessor ‘variance’] (McManus et al 2006), to a more meaningful use of post hoc analysis of OSCE outcomes to identify genuinely aberrant assessor behaviour. As well as providing pragmatic solutions in cases where assessor behaviour is clearly unsatisfactory, we anticipate this work will assist OSCE developers in bridging conceptual models of assessor cognition and judgement and practical station level design and assessor support.

Exemplar case studies and context

The work presents a series of case studies as examples of examining the identification and impact of extreme assessor judgments within an OSCE setting. Such contextualised case studies can help apply evidence in practice-based settings and provide practical advice to test developers (Harden et al 2015).

These studies are drawn from assessment data within a 5-year undergraduate Medicine programme with high stakes OSCEs undertaken at the end of years 3-5 as part of a sophisticated programme of assessment (Schuwirth & van der Vleuten 2011). The year 3 (intermediate level) OSCE is of 'traditional' test format with single examination of testing length ~2.5 hours, coupled with remediation and retesting for the weakest students. Senior OSCEs (years 4 and 5) are delivered via Sequential Test formats, where all candidates sit an initial screening test, and only those in the critical pass/fail region are required to sit a supplementary assessment. Such models of assessment have been described in detail elsewhere (Pell et al 2013). In each cohort in each year there are of the order of 250-280 students. With the exception of candidates who receive extra time as part of reasonable adjustments to assessment (as a result of physical or learning difficulties), all candidates are randomised into groups (i.e. circuits).

All examinations use local clinician (or clinical skills educators) examiners who go through a standard OSCE training programme, which includes familiarisation and practice with scoring rubrics and practical demonstrations on the impact of a range of ‘undesirable’ assessor

behaviours on the OSCE (such as excessive prompting, impact of failure to adhere to examiner instructions). This is further supported by on-the-day briefings and familiarisation with station-specific material in advance of the examination, and post-OSCE examiner feedback showing individual assessors how their typical scores compare with their peers.

The configuration of OSCEs in this programme requires the deployment of multiple parallel circuits across up to 4 sites, using up approximately 300 examiner 'slots'. This spread of the OSCE means there may be up to 12 parallel circuits of the same station underway across 3 sessions across the day, with up to 36 individual examiners assessing a station across multiple sites. Whilst examiners are able to choose their OSCE site (to maintain clinical cover arrangements in individual examination centres and to improve the cost-effectiveness of assessment), they are typically randomised to circuits and/or stations within site.

Each encounter is scored by a single assessor with a key features checklist (Farmer & Page 2005) and a global grade with five categories (Fail, borderline, clear pass, good pass, excellent). Standard setting is undertaken using the Borderline Regression method, with a comprehensive range of whole test and station level quality analyses undertaken post-hoc (Pell and Roberts 2006; Pell et al 2010). These analyses encompass an examination of a range of fixed effects pertinent to the test format described, including time, site and parallel circuits to detect any significant (non student) variance as a result of OSCE delivery and design.

In order to pass each OSCE, students must achieve the overall pass mark (the sum of the individual station pass marks + addition of a Standard Error of Measurement, an estimate of the error in the total scores (Streiner and Norman, 2008, p190-193)), and conjunctive criteria including a minimum number/types of stations. For stations identified as problematic in the post hoc analysis, the physical checklists as manually completed by the examiners are retrieved and inspected for unusual patterns of marking. Before decisions are arrived at in terms of how precisely to ameliorate any problems, the impact of different approaches are modelled in terms of changes to individual student decisions and assessment quality metrics.

Exemplar case studies

We report three cases from recent post-hoc OSCE analyses where individual assessor judgements were identified as sufficiently 'extreme' during initial post-hoc analysis to warrant further investigation. These exemplars are chosen in order to illustrate the range of approaches that can be taken to identify, categorise and assess the impact of extreme examiner scores.

Case 1 – A 'Rogue Examiner'?

In a third year OSCE (traditional format with 21 stations), four out of the 21 stations showed very poor station level metrics (low R^2 and unsatisfactory 'alpha with item deleted'). Further analysis pointed to a single common factor – an individual assessor within a parallel circuit in a single exam centre – as shown clearly in Figure 1 (assessor 1224) for one of these four

stations, station 1 (physical examination). This assessor was an experienced examiner (Consultant Physician) who had undertaken OSCE training, and had examined for a number of years with no prior problems. On investigation, it appeared that he had not filled in the key features checklist correctly, and had done this serially across circuits, but on discussion he could not explain why.

Whilst the global grades for the same station and examiner were within acceptable norms (Figure 2), we inferred that the checklist marks were not correct, potentially affecting a group of students who risked failing the station (and possibly the assessment as a whole) because of extreme low checklist scores.

After consultation between the assessment team, external examiner and student representatives, we decided to replace the low checklist marks for the affected students based on their global grades by using the average checklist marks (within grade) for the rest of the cohort for this station. This was felt to be the most parsimonious solution to the problem, maintaining a unified assessment for all students, and was acceptable to all stakeholders, and resulted in significant impact – two students who would have otherwise failed the OSCE overall passed as result of the adjustment made. This approach ensured that student opinion was considered, based on the constructivist perspective that students should be active participants in all aspects of learning and assessment and contribute to assessment practices (Rushton 2009).

Concerns about examiner judgements were initially raised through poor station level metrics, and further scrutiny allowed differentiation of the 'rogue examiner' (1224) from the rest of the assessor group. We then modelled the impact of these changes to marks on individual pass/fail decisions and to station level metrics, illustrated in Table 1. This highlights not only the impact of individual assessor behaviour at candidate level, but also the impact on overall station level metrics – we see that R^2 improves, as does 'alpha deleted', and that there is a large reduction in between-group variance.

For the other three stations involving this examiner, we carried out a similar adjustment to station checklist marks, although the patterns were less extreme whilst still significantly different from the station checklist mean. These adjustments resulted in one new failing

student and two new passes for the overall assessment and improvement in reliability (from alpha=0.74 to 0.75).

After the adjustment to checklist marks on station 1, the pass mark increased (from 23 to 25), leading to the number of station failures increasing by two (from 25 to 27). However, at the OSCE level, many students with low pre-adjusted checklist marks across affected circuits increased their aggregate score and so generally passed the assessment leading to an overall reduction by 1 in the number of failing students across the assessment as a whole.

Returning to Figure 1, we see that a group of assessors (1222, 1223, 2101, 2103, 2104) award very high checklist marks in this station with an apparent lack of discrimination between students in this group, all or almost all of whom were scoring the maximum available mark. One might argue that these scores too should be adjusted since the global grades demonstrate more discrimination (Figure 2). However, looking at the checklist scores across many assessors in this station, these are relatively high in this station. This indicates a ceiling effect where the majority of the candidates accumulated marks resulting in high scores because of a poorly designed, overly 'process focused' key features checklist. The ceiling effect is a common problem in cases where the checklist does not reflect the appropriate level being examined, and/or fails to capture key elements of the performance which assessors consider relevant and that are reflected in the global grade.

However, the corresponding global grades for these students do show discrimination (Figure 2), highlighting that whilst these checklist scores may initially appear 'extreme', the corresponding grades indicate that these examiners are apparently able to discriminate well between the students they examine. These contrasting assessor behaviours perhaps reflect the tension between what is 'experienced' by the assessor in the station, as recorded in their global grade, and the constraints of the design and scoring in the checklist (which was revised with consequently improved performance in a further OSCE).

Case 2 – 'Mis-scoring Assessors'?

In a different intermediate level OSCE, a similar pattern of assessors generating low checklist scores was identified (Figure 3, assessors 32 & 38). On investigation, it was found that this problem was due to a single, experienced assessor who had assessed two separate groups for the same station during a morning session. On inspecting the scoring sheets it appeared that the assessor had mistakenly reversed the criteria anchors (i.e. had effectively reversed the performance scale, scoring good performance lowly and poor performance highly). The examiner confirmed that this was indeed the case on discussion, and the external examiner was notified of the issue. With the agreement of all parties, the problem was relatively easily fixed by re-scoring these groups for this station in the correct way, reducing station level between group variance from 76.4% to 39.3%.

Of arguably greater interest, Figure 3 also shows that there are a number of assessors giving high marks with little discrimination between students (e.g. assessors 2, 27, 28, 34). This is similar to problems witnessed in the first case study (see Figure 1), and again the grades pertaining to these assessors seemed satisfactory. In this current case, it would suggest that the scoring format in the station means assessors are struggling to discriminate between students, possibly as a result of poor station construct and/or poor assessor guidance as to the 'expected' level. This is a common feature of intermediate/junior undergraduate OSCEs where the candidates' stage of training and assessor expectations of practice at this midway stage can appear to generate excessive variation in group mean scores (Chesser at al 2009, Pell et al 2009). Whilst these assessors might be regarded as giving 'extreme' scores in this case study, it is clear that problem lies at the station, rather than assessor, level.

Case 3 – Assessor judgements in resits and sequential tests

Traditional resits and the second part of a sequential test, where assessors are faced with a non-randomised, extreme subgroup of generally weak candidates, can pose challenges for assessors. Despite training, calibration and good station design, their expectations of acceptable performance may be different to that present in the main test as part of a wider 'contrast effect' (Yeates et al 2015). Our final case is taken from a sequence 2 OSCE consisting of 12 additional stations (the preceding main 'screening' test comprising 13 stations, making a total sequence of 25 stations for weaker candidates). In this second part of the sequential test, we use only stations that have been standard set in the past. Thirty-three students were recalled to undertake the Sequence 2 OSCE, and were split between three parallel circuits run over two sessions. Due to the relatively small size of the cohort, no rotation of assessors between sessions was necessary. This was intended to ensure assessor consistency within stations, which is particularly important given the nature of the candidates examined. All assessors were experienced OSCE examiners, regularly examining main and sequential test candidates and had been carefully briefed at the start of this Sequence 2 OSCE.

Figure 4 shows the extent of assessor problems – the assessor examining the 'blue' parallel circuit had very little discrimination between students, and the red assessor was scoring systematically lower than the other two assessors, with high levels of between group variance (81%). The station in question focused on a complex consultation and patient management encounter that had performed very well when used previously in main screening tests.

To further understand assessor behaviour in this station, Figure 5 shows that there was no overall linear relationship between checklist scores and global grades (r=-0.036, not significantly different from zero), highlighted by the flat line of best fit. The other two dotted curves are quadratic and cubic best fits and confirm that any apparent 'relationship' between checklist scores and global grades is far from linear for this station. This case provides a useful alert for assessments with small cohorts particularly those comprised of weaker/resitting candidates. Given the systematic difference between circuits (Figure 4) and the lack of linear relationship between checklist and global scores (collectively and within groups), there is insufficient information to make any systematic adjustment between circuits, and as result of these findings the station had to be withdrawn from the OSCE.

Discussion

Within assessments that use groups of assessors (typically as single station level assessors across multiple parallel circuits) to examine the performance of individual candidates, we have tended to believe that the characteristics of any individual examiner are balanced by those of colleagues, assuming a randomised allocation of students and examiners. Small bodies of work have started to concentrate on identifying examiners that are at the 'extremes' of marking at station level, revealing in one study that <0.3% of examiners can be classed as extreme [defined as an individual rater's mean score being greater or less than three standard deviations beyond the mean of all raters] (Bartman et al 2013). Whilst the overall proportion of such 'extreme' assessors appears small in these case studies, it remains important for us to understand whether these judgements' are truly 'extreme' or are a reflection of a complex environment which exerts powerful effects on assessment through test constructs and scoring formats and judgements about safe clinical care (Kogan et al 2014).

This paper uses three exemplar case studies that illustrate apparently extreme assessor judgments. To what extent were these judgments (and levels of variance) undesirable? We would argue that any numerical classification of extreme behaviour is arbitrary, and that any apparent 'extreme' behaviour should generate a close inspection of station quality and the impact on candidates. In two cases, behaviours at an individual assessor level meant that extreme (low) assessor scores risked candidate failure. Whilst the detection of these cases through application of recognised post-hoc analyses and subsequent exploration of impact use established methods, managing these effects correctly poses philosophical and policy challenges.

Re-examining a whole cohort is clearly impractical and difficult to justify, but the removal of (multiple) stations where assessor effects have impact on a number of candidate's poses risks to validity through effects on blueprinting and sampling. Imputation to replace 'bad' scores (as practiced in case 1) does ensure maximum use of data as part of the overall analysis, but requires careful modelling of the effect on both candidates obviously affected by 'extreme' assessor behaviour as well as the broader cohort who may be affected as a result of changes to station level passing scores. These cases also challenge assumptions that individual assessor judgements have little effect on the overall quality of the test, as illustrated by the significant impact on station level metrics and crucially, on the station passing score.

Of arguably greater interest in this study is the larger group of examiners who generated relatively high scores with little discrimination in comparison with their peers examining the same station (as seen in both cases 1 and 2). Multiple factors in OSCEs, both individual (cognitive overload) and external (poor assessor support, poor quality scoring instruments) can easily lead to examiners being labelled as 'extreme'. A focus on scores alone to identify such variation is overly simplistic and closer analysis of assessors' use of scores and global grades does show evidence of discrimination, aligning with the conceptual frameworks outlined by Govaerts and others of highly contextualised, individual examiner judgments (Govaerts 2016). In both cases, we inferred that the problem lay largely at a station design level, promoting a review of existing scoring instruments and potential ceiling effects. Ongoing work within our own institution seeks to quantify the extent and impact of such ceiling effects in scoring in greater detail (for example on station level metrics and standard setting), and to better understand this complexity via an exploration of the (written) narrative feedback our assessors provide for each candidate within an OSCE station.

The final case illustrated in this study reveals some of the significant challenges associated with the examination of sub-groups of weak candidates, for example in resits or the second sequence in a sequential test. Candidates tend to perform better in repeat or supplementary assessments, and recent literature highlights that some of the difficulties in setting the standard for such an extreme subgroup, where students numbers may be smaller and individual assessor effects more pronounced [in this example, typically 30 students and 15-20 assessors] (McManus & Ludka 2012; Pell et al 2012; Homer et al 2015). There are likely to be multiple factors contributing to the variance seen in Case 3, including contrast effects between weaker candidates and the 'duty' examiners face as educators, assessors and practitioners are likely to exert powerful effects (Kogan et al 2014; Yeates 2015).

One of the potential limitations of this work is that it is situated within a single institution, although these findings are spread across different examiner and student cohorts and different years of study, suggesting a wider effect that would be seen in other institutions' OSCEs. The nature of these case studies (revealed through post hoc analysis) meant we were unable to perform direct observation of the assessor-candidate encounters to better understand the challenges of a complex performance assessment, nor undertake detailed interview work with assessors in a 'live' OSCE.

Whilst the number of cases we are able to present in this work is small, the impact of these 'extremes' of marking is important, particularly for pass/fail decisions in the critical region, reflecting the findings of others (Bartman et al 2013). It is also tempting to postulate that such problems would not arise through the use of global grades only (as employed by many institutions), but we would argue that the use of key feature scoring/checklist formats and global grades allowed the generation of sophisticated station level metrics which consequently aided in detection and categorisation of assessor judgements judged to be at 'the extreme'. We also find examples where the global grades are themselves problematic, for example with 'hawks and doves' and/or lack of discrimination within circuits. It is also noteworthy that all the cases detected in this study related to experienced examiners who had participated in training and had assessed on multiple previous OSCEs.

In order to best understand the implication and formulation of extreme assessor behaviour in its entirety, our research direction will develop based on the exploration of the constructivist approach to learning and assessment. Constructivism has a long standing history in educational research and underpins the rationale for explorative qualitative research (Driver & Oldham 1986). How assessors construct their knowledge, and how this interacts with self-regulation of knowledge and impacts their behaviour is clearly key. Moving forward, we aim to further unpick the complexity of assessor variance, and the formulation of examiner judgements using these tools. Focus groups and interviews with students and assessors will help to progress the research in ways that best complement psychometric analysis.

In this 'post-psychometric' era, our focus should move from the mechanistic application of psychometrics and the 'correction' of 'extreme' scores to a model including measurement *and* exploration of judgements (Eva & Hodges, 2012). In doing so, we begin to identify and unpick the multiple factors which contribute to score variance in OSCEs, bridging meaningful psychometric analyses and constructivism to help better understand examiner judgements within the OSCE.

References

Bartman I, Smee S, Roy M (2013). A method for identifying extreme OSCE examiners. *Clinical Teacher* 10:27-3

Brooks, JG & Brooks, MG. (1993) In Search of Understanding: The Case for Constructivist Classrooms (Alexandra VA, Association for Supervision and Curriculum Development).

Chesser A, Cameron H, Evans P, Cleland J, Boursicot K, Mires G (2009). Sources of variation in performance on a shared OSCE station across four UK medical schools. *Med Educ*. 43(6):526–32.

Crossley, J., Davies, H., Humphris, G. and Jolly, B. 2002. Generalisability: a key to unlock professional assessment. *Med Educ*. 36(10):972–978.

Driver, R, Oldham V (1986) A Constructivist Approach to Curriculum Development in Science, *Studies in Science Education*, 13 (1), 105-122.

Eva K, Hodges B (2012). Scylla or Charybdis? Can we navigate between objectification and judgement in assessment? *Med Educ* 46(9):914-919

Farmer E, Page G (2005). A practical guide to assessing clinical decision-making skills using the key features approach*. Med Educ*. 39:1188–94.

Fuller R, Homer M, Pell G (2013). Longitudinal interrelationships of OSCE station level analyses, quality improvement and overall reliability. *Med Teach*. 35(6):515–7.

Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E (2014). Seeing the 'black box' differently: assessor cognition from three research perspectives. *Med Educ* 48(11):1055-68

Gormley G, Johnston J, Thomson C, McGlade K (2012). Awarding global grades in OSCEs: Evaluation of a novel eLearning resource for OSCE examiners. *Med Teach* 34(7):587-589.

Govaerts M, van der Vleuten C (2007). Broadening perspectives on clinical performance assessment: rethinking the nature of in training assessment. *Adv Health Sci Educ Theory* 12:239-260

Govaerts M, van der Viel M, Schuwirth L, van der Vleuten C, Muijtjens A (2013). Workplace-based assessment: raters' performance theories and constructs*. Adv Health Sci Educ Theory Pract* 18(3)*:*375–96

Govaerts M (2016). Competence in Assessment: Beyond Cognition. *Med Educ*. 50(5):502-504

Harasym P, Woloschuk W, Cunning L (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv in Health Sci Educ* 13:617-632

Harden RM, Crosby J (2000) AMEE Guide No 20: The good teacher is more than a lecturer - the twelve roles of the teacher, *Medical Teacher*, 22 (4), 334-347

Harden RM, Stevenson M, Downie WW, Wilson GM (1975). Assessment of clinical competence using objective structured examination. *British Medical Journal* 1:447–51.

Harden R, Lilley, P, Patricio, P. *The Definitive Guide to the OSCE: The Objective Structured Clinical Examination as a performance assessment*, 2015. First edition. Edinburgh ; New York: Churchill Livingstone.

Holmboe ES, Hawkins RE, Huot SJ (2004). Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Ann Intern Med*. 140(11):874–81.

Homer M, Pell G, Fuller R, Patterson J (2015). Quantifying error in OSCE standard setting for varying cohort sizes: A resampling approach to measuring assessment quality. *Med Teach*. 24:1–8.

Kogan JR, Conforti L, Bernabeo E, Iobst W, Holmboe E (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Med Educ*. 45(10):1048–60.

Kogan JR, Conforti LN, Iobst WF, Holmboe ES (2014). Reconceptualizing variable rater assessments as both an educational and clinical care problem. *Acad Med*. 89(5):721–7.

McManus IC, Thompson M, Mollon J (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES): using multi-facet Rasch modelling. *BMC Medical Education* 6:42

McManus IC, Ludka K (2012). Resitting a high-stakes postgraduate medical examination on multiple occasions: nonlinear multilevel modelling of performance in the MRCP(UK) examinations. *BMC Medicine*. 10(1):60.

Patricio M, Juliao M, Fareleira F, Young M, Norman G, Vaz Carneiro A (2009). A comprehensive checklist for reporting the use of OSCEs. *Med Teach* 31(2):112-124.

Pell RG; Roberts TE. Setting standards for student assessment (2006). *International Journal of Research and Method in Education*. 29(1):91-103.

Pell G, Homer MS, Roberts TE. Assessor training: its effects on criterion-based assessment in a medical context (2008). *International Journal of Research & Method in Education.* 31(2):143–54.

Pell G, Fuller R, Roberts T, Homer M (2009). Comments on within-station between-sites variation. *Med Educ*. 43(10):1021–2.

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(10):802–11.

Pell G, Fuller R, Homer M, Roberts T (2012). Is short-term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments. *Med Teach*. 34(2):146–50.

Pell G, Fuller R, Homer M, Roberts T (2013). Advancing the objective structured clinical examination: sequential testing in theory and practice. *Med Educ*. 47(6):569–77.

Rushton A (2005) Formative assessment: a key to deep learning?. Medical Teacher, 27 (6), 509-513

Schuwirth, Lambert WT, and Cees PM van der Vleuten (2011). Programmatic assessment: from assessment of learning to assessment for learning. *Med Teach* 33(6):478-485.

Streiner, D. and Norman, G. 2003. *Health Measurement Scales: A practical guide to their development and use.* 4th ed. OUP Oxford.

Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66 (2012). *Med Teach* 34(3):e161–e175.

Yeates P, Cardell J, Byrne G, Eva KW (2015). Relatively speaking: contrast effects influence assessors' scores and narrative feedback. *Med Educ*. 49(9):909–19

Glossary

Sequential testing

In a sequential testing format, all candidates sit an initial 'screening' test and only those candidates who fail to demonstrate sufficient competence on this part are required to sit a supplementary test, usually of a similar size to the first part (Pell et al, 2013). Pass/fail decisions for this weaker group are made using performance across both parts of the assessment.

Pell G, Fuller R, Homer M, Roberts T (2013). Advancing the objective structured clinical examination: sequential testing in theory and practice. *Med Educ*. 47(6):569–77.

Standard error of measurement

All assessments are subject to measurement error (i.e. to error in the test scores). The standard error of measurement is an estimate of the size of this error, and is related to the reliability of the assessment (so small standard errors of measurement correspond to high levels of reliability and vice versa). (Streiner and Norman, 2008, p190-193).

Streiner, D. and Norman, G. 2003. *Health Measurement Scales: A practical guide to their development and use.* 4th ed. OUP Oxford.

Practice points

- Institutions have a duty to maximise the defensibility, quality and validity of the assessment process.
- Individual examiner behaviour and OSCE station design can impact on measures of assessment quality, and assessment outcomes, including on pass/fail decisions.
- Careful analysis of assessor judgments, through a range of post hoc approaches, can highlight possible aberrant behaviours.
- Care must be taken when assuming that apparent aberrant examiner behaviour is automatically just that.