Hawks, Doves and Rasch decisions

Understanding the influence of different cycles of an OSCE exam on students' scores using Many Facet Rasch Modelling.

Authors:

Peter Yeates
Stefanie S. Sebok-Syer

Corresponding author:
Peter Yeates
Keele University School of Medicine, David Weatheral Building, Keele University, Stoke-on-Trent, Staffordshire ST5 5BG
Tel: +44 1782 733930
Fax: n/a
Email: p.yeates@keele.ac.uk

Notes on contributors:

Peter Yeates is a lecturer in medical education and consultant in acute and respiratory medicine. His research focuses on assessor variability and assessor cognition within health professionals' education.

Stefanie S. Sebok-Syer is a postdoctoral Fellow at the Centre for Education Research and Innovation specializing in measurement, assessment, and evaluation. Her main interests include exploring the rating behaviour of assessors, particularly in high-stakes assessment contexts.

Cite as: Peter Yeates, Stefanie S Sebok-Syer. Hawks, Doves and Rasch decisions: Understanding the influence of different cycles of an OSCE on students' scores using Many Facet Rasch Modeling. *Medical Teacher* 2016. *On-line*
http://www.tandfonline.com/doi/abs/10.1080/0142159X.2017.1248916

DOI: 10.1080/0142159X.2017.1248916

Abstract (max 200 words)

Introduction: OSCEs are commonly conducted in multiple cycles (different circuits, times, and locations) yet the potential for students' allocation to different OSCE cycles is rarely considered as a source of variance - perhaps in part because conventional psychometrics provide limited insight.

Methods: We used Many Facet Rasch Modelling (MFRM) to estimate the influence of "examiner cohorts" (the combined influence of the examiners in the cycle to which each student was allocated) on students' scores within a fully nested multi-cycle OSCE.

Results: Observed average scores for examiners cycles varied by 8.6% but model adjusted estimates showed a smaller range of 4.4%. Most students' scores were only slightly altered by the model; the greatest score increase was 5.3%, and greatest score decrease was -3.6%, with 2 students passing who would have failed.

Discussion: Despite using 16 examiners per cycle, examiner variability did not completely counter-balance, resulting in an influence of OSCE cycles on students' scores. Assumptions were required for the MFRM analysis; innovative procedures to overcome these limitations and strengthen OSCEs are discussed.

Conclusions: OSCE cycle allocation has the potential to exert a small but unfair influence on students' OSCE scores; these little-considered influences should challenge our assumptions and design of OSCEs.

Practice Points:

OSCEs are often conducted in multiple-cycles, at different times or locations, yet the influence of different cycles on students' scores is rarely considered.

Small differences occurred due to the effect of different OSCE cycles that had an important influence on some students' scores.

Modifications to OSCE procedures are needed to robustly estimate these effects.

Glossary terms:

Many Facet Rasch Modelling (MFRM):

A form of psychometric analysis, derived from item response theory, which models the comparative influence of a number of "facets" (variables that can have an influence) on examinees' scores. The modelling is then able to estimate the score each examinee would have received if they had encountered a completely neutral example of each facet (for example a completely neutral examiner), and provide an estimated "fair score".

Reference: Bond, T., & Fox, C. (2012). Applying the Rasch Model Fundamental Measurement in the Human Sciences (2nd Editio). New York & London: Routledge.

Key words:

Medical Education research < Management

Assessment < OSCE

Psychometrics < Assessment

Introduction

Ensuring that assessments for healthcare professions students are conducted fairly is vital to the validity of assessment results (Kane, 2006) Additionally fairness in assessments is crucial to maintaining trust between an institution and its students (Watling, 2014). Given an institution's duty and responsibility to the public (Epstein & Hundert, 2002), the notion of fairness in assessment practices is imperative. Although "fairness" can be interpreted in numerous ways, in this article we consider one aspect of fairness: how the common practice of conducting Objective Structured Clinical Exams (OSCEs) in multiple circuits, at different times and locations (termed hereafter as multiple "cycles" of an OSCE), may influence the accuracy, or overall standard of judgement, by which students in different cycles are examined. Put differently, we aimed to consider whether the cycle of an OSCE that a student is allocated to substantially impacts their overall score. Describing a comparatively novel approach to understanding these influences, we highlight this important consideration and discuss modifications to the assessment process that could aid fairness in clinical assessments.

Objective structured clinical exams (OSCEs) were developed in the 1970s in response to overt problems with examiner subjectivity and variable case mix in exams (Harden, Stevenson, Downie, & Wilson, 1975). Whilst OSCEs arguably comprise one of the more important innovations in medical education during the 20th century, it is clear that they only partly overcome the problem of examiner subjectivity. In a meta-analysis of studies that report on the reliability of OSCEs, Brannick et al (2011) showed an average Cronbach's alpha value of just 0.66, with lower results for communication-focused stations and less experienced examiners. Haraysm et al (2008) found that variance due to OSCE examiners (44.2%) was more than 4 times greater than variance due to students (10.3%). Moreover, adjusting for examiner variance altered the pass/fail decisions of 11% of students in their study. Other studies have suggested that examiners' scores may be biased depending on the timing of OSCEs (Hope & Cameron, 2015; McLaughlin, Ainslie, Coderre, Wright, & Violato, 2009), the performance of other candidates (Yeates, Moreau, & Eva, 2015) or by different geographical locations (Sebok, Roy, Klinger, & De Champlain, 2015). Consequently, it is not sufficient to simply conduct an OSCE, and believe that the resulting scores are a fair representation of students' performance given the known influences of construct irrelevant variance.

The most traditionally used psychometric analysis for OSCEs is generalisability theory (Crossley, Davies, Humphris, & Jolly, 2002). Generalisability analysis works best when examiners are fully "crossed" with students and stations (i.e., all examiners observe all students at all stations); however, the model can produce useful data when some partial crossing is available (Shavelson, Webb, & Rowley, 1989). With a fully nested design (as typically observed in most medical schools), generalizability is more limited, as it is not able to disentangle variance due to examiner variability from that due to students' abilities (Raymond, Swygert, & Kahraman, 2012). Nesting occurs when, for example, students are examined at each station by just one examiner. Further nesting occurs when examiners are allocated to a particular location, at a specific time. This is problematic because nesting impacts our ability to directly compare the effect of different examiners on students' scores given the lack of data on how other examiners would have scored that specific student. In this paper, we propose a design for investigating examiner variability and student ability by viewing individual examiners as part of a group of examiners (hereafter referred to as "examiner cohorts") using the Many Facet Rasch Model (MFRM).

Reliability is generally understood to increase more with greater numbers of stations than with more examiners at each station (Newblet & Swansons, 1988) and acceptable reliability is likely obtained with 2 to 4 hours of testing time (van der Vleuten & Schuwirth, 2005). Whilst these processes undoubtedly enhance the quality of OSCEs, we contend that the influence of examiner cohorts (i.e., the collective scoring of all examiners that assessed a particular student) is rarely considered. Some larger medical schools run a total of 16 OCSE cycles at 4 locations over 2 consecutive days. We argue that an important element of fairness is ensuring that whether a student passes or fails is not dependent on the cycle of the OSCE to which they are allocated. Perhaps more simplistically, institutions often assume, based on theory (e.g., central limit theorem (Field, 2009, p42), that because examiners have been trained, and because there are 16 examiners in each OSCE cycle, examiners' differences will simply even out, and the *average* influence of examiners in each cycle will be the same.

We contend that there are a number of reasons why the aforementioned assumptions ought to be challenged: 1) examiners are typically not randomised into cycles, but are allocated chiefly based on geography and convenience; 2) examiners in different localities may develop different tacit performance norms, different shared expectations or hidden curricula (Hafferty & Franks, 1994); or different standards of practice (Weiner et al., 1995), which are known to inversely relate to assessors' severity of judgements (Kogan, Hess, Conforti, & Holmboe, 2010); 3) timing of cycles may also influence the standard of judgement (Hope & Cameron, 2015; McLaughlin et al., 2009). Consequently we suggest that in order to understand fairness in OSCEs we should seek to determine whether the combined influence of examiners is trivial or cause for concern. In the remainder of this paper we present our research questions, describe the design for investigating examiner cohort effects, and suggest alterative solutions to the OSCE procedure that could improve our ability to assess students' ability in the OSCE environment.

Research Questions:

1) What is the impact of examiner cohorts on students' scores in a 16 station multi-cycle OSCE? "Examiner cohort" refers to the combined influence of the 16 individual examiners that students' encountered in a given OSCE cycle.
2) What insights can be gleaned about fairness in OSCE assessments when individual examiners are viewed as a collective cohort?

Methods:

Data from a summative undergraduate OSCE in a UK medical school was examined to investigate the impact of examiner cohorts on students' scores. This particular OSCE comprised 16 stations and was conducted at 4 different site locations, on 2 consecutive days, with morning and afternoon cycles on both days. The term "cycle" refers to an administration of the OSCE examination. See Figure 1 for a visual representation of the multi-cycle OSCE administration. Cycles varied in the number of students examined. Most were long, and examined approximately 16 students. Two sites ran short cycles in the afternoons of approximately 8 students. The precise numbers of students in each cycle were varied by local administrators by either adding additional rest stations or allowing gaps in the carousel. Students were quarantined between morning and afternoon cycles to enable the same stations to be re-used within the same day; analogous stations were used on successive days. Stations comprising a range of tasks were blueprinted against learning objectives for the students' stage, and tested communication skills, procedural skills, physical examination skills and history taking skills. Most stations included supplementary tasks such as data interpretation or patient management questions. At each of the 16 stations, examiners scored the performances using domain rating scales. Desirable behaviours were listed for each domain, and examiners rated the extent to which these behaviours were successfully completed. Additionally, each examiner completed an overall judgement of the standard of the performance on a 7 point Likert scale that ranged from *Clear Fail* (points 1 & 2); *Borderline* (fail)(3); *Satisfactory* (4); *Good* (5); *Excellent* (6 & 7). Satisfactory was defined as an acceptable standard of performance in order to progress to the next stage of training. Summative judgements were made based on the overall judgement score that consisted of judgments provided by all 16 examiners that observed students' during the OSCE; domain scores were supplied to students for feedback purposes. All examiners had undergone OSCE examiner training that involved explanation of the scoring format, and video review of cases with discussion. They also underwent a repeat briefing on their role and the scoring format immediately prior to commencing the OSCE.

We analysed the data using the MFRM. Rasch modelling is a variant of item response theory; its main benefit is that the model is able to examine the influence of multiple variables simultaneously. For example, it can simultaneously consider students' ability, examiner severity/leniency, exam location, and station difficulty. As intended audience for this paper are those who are assessment minded (but not necessarily technical experts), this paper will not describe Rasch modelling in great detail. Interested readers may find either the text by Bond and Fox (2012), or (for a briefer explanation) the paper by Sebok, Luu, and Klinger (2014) approachable introductions.

As the 7-point overall performance score on each of the 16 stations was the variable on which assessment decision are based, we used it as our dependent variable. The model comprised 4 facets: students, stations and OSCE site locations as well as a novel facet that we termed "examiner cohorts." Examiner cohorts described the combined influence of the 16 examiners that a particular student encountered during their assigned OSCE cycle, or put differently, the combined influence of a group of examiners operating within a specific time and location. Each examiner cohort represented a distinct group of examiners, without significant overlap with other examiner cohorts. The examiner cohort facet was created because we wanted to explore the combined impact of all the examiners each student encountered during the OSCE. Given that pass or fail decisions are made based upon the judgments provided by 16 different examiners, it seemed logical to investigate the examiners' severity/leniency collectively. An examiner cohort was created based on 16 unique examiner identification numbers (i.e., 16 different examiners). The number of students that an examiner cohort observed varied depending upon whether they were examining a long or short cycle and fluctuations in student numbers as described above. All examiner cohorts were nested within a single cycle. As a result, the number of students any given examiner cohort observed varied from 8 to 16. The examiner cohort facet also provided a link, which allowed us to compare examiners within a single frame of reference. This is typically not feasible given the nested structure of examiners within OSCEs. In most OSCE situations, comparisons among individual examiners are not possible because one cannot compare different examiners observing different students at different stations. However, the creation of the examiner cohort facet allows for comparisons given that each student was assessed on the same 16 stations by a cohort of 16 examiners. The nesting problem that exists with most OSCE assessments is prominent in most medical schools; thus, the examiner cohort facet provides a solution to allow for the use of traditional psychometrics to help disentangle variance between examiners, site locations, and students' ability.

We used a procedure within MFRM known as "anchoring" to anchor the stations according to the level of difficulty initially specified by the model. Anchoring is when a particular measure is fixed in order to equate (or compare) across groups using the same scale. In this instance, anchoring allowed the model to use the stations as common items across the data set, because all students were examined within the same stations. We allowed the model to presume that *average* student ability was likely to be the same across all OSCE cycles. This assumption was made because students were essentially randomly assigned to a site location (unlike examiners who were allocated based on geography and convenience). Whilst average student ability was assumed to be even, the model is able to examine individual students' abilities given that the stations were anchored, and thus, each student has a corresponding logit value (i.e., a measure of ability on the standardized Rasch scale) produced based on their scores on the 16 station OSCE. Given this design, we were also able to estimate the relative influence of examiner cohorts and site locations.

For this study, our principle measure of interest was observed averages (the raw mean of items within each facet) and "fair averages" (the model-derived estimate of the items within each facet once every other influence has been accounted for). We compared the maximal increase and decrease in student scores between student observed average scores and fair average scores, as well as the number of students for whom pass/fail decisions would have been altered. We also examined the reliability of separation for each facet. Notably, reliability of separation indices operate differently to conventional reliability parameters; in particular the reliability of separation for examiner cohorts describes the degree to which examiner cohorts differ from one another

(Sebok, et al., 2014). Consequently it is more akin to a “Hawk-Dove reliability” (Streiner & Norman 2008, p182) than a conventional inter-rater reliability. As such it is desirable that this parameter should be low (rather than high) as low values indicate little difference among the items within the facet.

Data analyses were conducted using Facets software version 3.68.1, available from Winsteps (Linacre 2011). All student data were anonymised before being passed to researchers. The university ethics committee confirmed that as the study performed secondary analysis of anonymous data and ethical review was not required.

Results:

Data from all 16 cycles of the exam, comprising a total of 235 students, and a further 21 “gaps” (available places in an OSCE cycle that were not filled) were analyzed. At any given site location, 57 to 60 students were examined. Findings from the MFRM are summarised in Figure 2. This graphicalisation of the data is known as a Wright Map and is routinely provided by most MFRM software. The Wright Map is useful as it displays all items from each facet on a common scale (known as a logit scale, shown on the left of the image). This common scale allows for direct comparisons among every facet. Consequently more able students; more difficult stations; more stringent examiner cohorts; and sites with a more stringent influence on scores are all positioned towards the top of the chart. Whilst this logit scale is a useful means to compare relative influences of items within and between facets, it is less obvious what these influences would mean to student scores. Consequently the model also provides these data displayed in the assessment scale units (the 7 point overall global rating). These are shown on the right hand side of the graph, although (due to range restriction in the raw score data) they range from approximately 3 to 7.

Fit of data to the model:

Mean square infit and outfit indices were between 0.5 and 1.5 for all stations, all sites and all examiner cohorts. This indicated that the extent to which the data fitted the model was good for all of these facets, and that the model was useful for both describing the data and for making adjustments to the data (Linacre, 1996). Importantly neither of the potential problems with fit (underfit, suggesting that a particular item is poorly described by the model; or overfit suggesting that a particular item was highly predictable or redundant) was indicated by our infit and outfit parameters for the facets of station, site or examiner cohort.

Conversely, the mean square Infit and outfit parameters for students showed that 37 out of 235 (15.7%) students showed either one or both mean square values outside the 0.5-1.5 mean square range. Of these, the majority 25 students or 10.6% of the sample showed underfit, suggesting that their performance was erratic in some way. An example of erratic behaviour would be when a student of high ability scores well on a difficult task but poorly on an easy task, as one would expect a student of high ability level to score well on an easy task. In all but 1 case, this underfit was minor (MS 1.5-2.0). 12 students (5.1%) showed overfit, suggesting that their scores showed less random variation than might be expected. Whilst we might wonder if these students are unusually "test-wise"(Case & Swanson, 1993), the model will still predict their performance well. Nonetheless these findings merely suggest that we would need to be cautious when adjusting the scores of students that did not fit the model.

Examiner cohorts:

Observed scores for different examiner cohorts ranged from 4.6 to 5.2 on the OSCEs 7 point scale, a difference of 8.6% in scores. These observed scores represent the simple unadjusted scores for the examiner cohorts. Model-derived "fair average" scores for these examiner cohorts show less variation, ranging from 4.80 to 5.11, or a difference of 4.4%. Fair average scores are the average score for an item once all other facets have been adjusted to a value of zero logits. Essentially, this is the score that an average student would expect in a particular examiner cohort once all other factors have been accounted for. The concept of the “fair averages” for examiner cohorts may be easier to conceptualise using the logit scale. This indicates that once the influences of other facets has been accounted for, the relative average influence of different examiner cohorts on students’ scores ranged from a reduction of -0.20 (SE ±0.06) logits to an increase of 0.10 (SE ±0.06). The fair average scores illustrate the degree of this influence using the OSCEs scale. The reliability of separation for examiner cohorts was 0.31, indicating that differences in examiner groups were observed, but that the effect was minimal. To illustrate this, inter-rater reliability is akin to 1-reliability of separation, so in this instance would be 1 - 0.31 = 0.69). Values below 0.80 are commonly viewed as desirable for judged events. These data are illustrated in Table 1.

Stations:

Station observed averages ranged from 4.0 to 5.5, with corresponding fair average values ranging from 4.03 to 5.48. Again, the fair average value is the score that an average student might expect at this station once other factors have been accounted for. Notably the differences between the fair averages and the observed averages for station are very small, because the stations were fully crossed with students and examiner cohorts. The same results examined via the logit scale show that stations ranged in difficulty from -0.88 (SE ±0.06) logits to 0.48 (SE±0.07), with stations with a lower logit score tending to reduce students overall scores and stations with a higher logit score tending to increase them. The common logit scale makes it easier to compare the influence of different facets when compared against each other. Therefore, it was observed that the influence of different stations (1.36 logits) was much greater than the influence of different examiner cohorts (0.30 logits). This finding is not surprising as this finding highlights “content specificity,” a well researched area in medical education that illustrates that the content from some OSCE stations is more difficult than other stations and that station difficulty is context specific.

Sites:

Small differences were observed between OSCE sites in the observed scores (range 4.8 to 5.1). Similar to examiner cohorts, the fair average scores for sites showed less of a range (4.91-5.08), and therefore once again the model derived estimates show less influence due to different sites than might be inferred from the observed averages (4.3% difference for observed scores vs 2.4% difference for fair average scores. On the logit scale the influence of site ranged from -0.09 (SE ±0.03) to 0.07 (SE ±0.03). For those who are unfamiliar with logit measures, this represents a very small degree of influence due to site location.

Students:

Students' observed-scores ranged from 3.7 to 6.3. The pass mark was 4.0 or above and the maximum possible score was 7.0. Similar to the other facets, students' fair average scores showed a slightly smaller range, from 3.98 to 6.17. The majority of students' scores received only small adjustments between observed and fair average scores; 64% of students' scores were adjusted by ≤±1%. This suggests that although the model detected small statistical differences, the practical implications of these adjustments were trivial. A small subset of student's observed average scores were adjusted by larger amounts, indicating instances where fairness may have been compromised. The largest upward adjustment was 5.3% from an observed average score of 4.1 to a fair average of 4.47. The largest downward adjustment was -3.6% from an observed average of 4.8 to a fair average of 4.58. Notably neither of these students had infit or outfit parameters that were adverse, and neither were at the extremes of student ability. Overall, two students who would have minimally failed based on their observed scores would have passed based on their fair average scores, although both remained very close to the pass/fail threshold (3.70 increased to 4.06, and 3.90 to 4.04 respectively). No students that passed based on observed scores would have failed based on fair average scores, possibly indicating a reluctance to fail students by some examiners. The reliability of separation for students was 0.64, indicating that the sample for analysis showed a fair degree of heterogeneity in terms of student ability. Given the high baseline educational attainment of medical students, a reliability of separation of 0.64 is adequate.

Discussion:

Summary of findings:

Analysis of the relative influence of different examiner cohorts on students' scores in a multi-site, fully nested OSCE demonstrated that viewing examiners as a cohort can address some of the challenges associated with nested OSCE designs. Model-derived "fair average scores" showed apparent, albeit small, differences in the standard of judgment employed in different examiner cohorts, suggesting that examiner bias can still exist even when 16 examiners are utilized. These differences were smaller than those implied by students' observed average scores, but they were observed nonetheless. This has important implications for medical schools that use 6, 8, 10, or 12 station OSCEs as the variance associated with individual examiners could be more prevalent and could have greater consequences. Different exam sites had a very small influence on scores. For most students, use of the model-derived fair average scores would have produced only a very small correction in their scores in this instance, but for a minority larger adjustments would have resulted. Consequently whilst this OSCE appears to have achieved good levels of fairness overall, it is notable that small effects did occur, that would have altered the pass/fail decisions for a very small number of students around the cut score.

Implications for practice:

These findings have a number of practical implications. Conventional approaches to OSCE assessment recommend that a number of indices are calculated for quality assurance. Examiner variability is usually considered by means of reliability, which, in turn is usually calculated using either Cronbach's alpha or generalizability theory (Schuwirth & van der Vleuten, 2011). Whilst these are useful parameters in developing an argument for the overall validity of the assessment (Downing, 2003), they are unlikely to provide an appreciation of the issues we have illustrated. Generalizability analysis relies on the assumption that error variance is randomly distributed (Streiner & Norman, 2008, p 301). Importantly, the effect we have described would produce a systemic (construct irrelevant) source of variance that is *non*-random, and would therefore not be seen as error within a generalizability analysis. We might expect that this would be (inappropriately) interpreted by the analysis as variation in students' performance and might therefore be expected to *increase* the calculated reliability of the OSCE (See Bond & Fox 2012, p155-157 for a worked example of this phenomenon). Consequently, even when the overall reliability of an OSCE appears good, students could still potentially be unfairly exposed to differing standards of judgement in different cycles of the OSCE.

Whilst the influence of examiner cohorts was small in this instance, there is no reason to presume that it could not be larger in other instances. Consequently, we believe institutions should consider monitoring the influence of these effects. This recommendation is consistent with the intent of prior authors; Pell et al (2010) recommended that the observed averages of students' scores in each exam cycle should be compared using ANOVA. Whilst we agree with their intent, our data suggest that the simple observed averages of each cycle may over-estimate the differences between examiner cohorts. Further theoretical and empirical work is required to more fully understand the relative merits of these different approaches. It is important to create assessment designs that are consistent and reflective of the decision making that result from examiner judgments. At some medical schools, the judgment of one examiner from a single station is enough to justify the overall pass/fail decision. These institutions would benefit from a design where both examiner cycles and stations are anchored to account the possible variations in examiner behaviour.

Strengths and Limitations:

Our study highlighted a previously unidentified area of construct irrelevant variance in OSCE assessments that warranted further examination. Our results indicate that even with an (apparently) optimal number of examiner judgments, variance can still exist within and between examiner groups. Because this effect may arise from a systematic rather than random effect, this suggests that increasing the number of OSCE stations or length of an OSCE examination may not necessarily equate to more accurate and reliable judgments regarding students' ability. Furthermore, this study highlighted an important aspect that requires consideration in order to uphold the principles of fairness in OSCE assessments.

Given that students were allocated to cycles through random assignment, we made the assumption that average student ability within each cycle was even. As a result, we suggest that the procedure in its current form should not be used to make summative adjustment to students' scores. Ideally, baseline data on students' prior performances would have been available and should be used in order to cross-validate this assumption. Without established baseline measures, one cannot state with 100% certainty if differences are due to examiner influence or student ability. Nonetheless we believe that this analysis is useful to inform discussions within institutions and highlight the potential for exploring examiner variance by investigating examiner cohorts. As described in the results, the fit between the data and the model was not perfect in every instance. As previously stated, 12% of students showed poor fit to the model, which unfortunately suggests limited utility with respect to assessing those students' ability.

Recommendations for development:

As stated in the methods section, using MFRM in this manner required an assumption that average student ability was likely to be even across OSCE circuits. A number of procedures are conceivable that would enable us to either test this assumption or to proceed without the need for it. Firstly, we could use formal randomization to allocate students to cycles of the OSCE (as recommended by Pell et al. (2010)). This would help to strengthen our assumption that average student ability was even between cycles, but (as with all sampling distributions), some variation in average ability would still be expected. Secondly, it would be possible to obtain baseline data from other assessments to verify whether students' ability did indeed appear to have been evenly distributed. Such data could be drawn from contemporaneous knowledge-based exams or prior OSCEs. Both of these approaches have limitations, in that students' real abilities may have changed with time or differ between assessment modalities. Moreover, if students' abilities were shown to be unevenly distributed across OSCE cycles, there would be little way to adjust the model accordingly.

A better approach would be to find a means to make comparisons within the data set. As described earlier a fully crossed design is desirable, but almost never observed in practice. MFRM is capable of equating between items that are otherwise nested if there is a sufficient amount of cross-over material, for example by rotating examiners between locations (Bond & Fox, 2012). Therefore, if 4 examiners from the morning cycle in each location moved to the afternoon cycle in a different location, then sufficient crossover would be available to link all the sites. If a different 4 examiners remained in each site for both the morning and the afternoon and a further subset of 4 examiners were persuaded to attend on both days, then all sites, days and parts of the day would be linked. As a consequence, there would be adequate linkage across the dataset to enable the model to directly compare the standard of judgement in each examiner cohort. This would remove the need for the assumption regarding average student ability, and make the model more dependable. An alternative approach to linkage could involve examiners rating a common pool of video performances in addition to the OSCE circuit that they examined. Regardless, some process to link the currently fully nested OSCE circuits would improve the ability of MFRM to equate between examiner cohorts. All of these suggestions require research in order to determine whether they are practical and to determine the extent to which they produce dependable results. Finally, these recommendations for development all stress the importance of having a good design in place for collecting assessment data from OSCEs.

Conclusions:

In this study we have demonstrated the potential for multi-cycle OSCEs to be unfairly biased towards students examined in different cycles of the OSCE. Whilst this effect displayed minimal practical significance in the particular dataset, we have shown how common assumptions could lead to bias in OSCE settings. We have illustrated a procedure that enables the MFRM to provide insight into this potential, while also noting modifications to the exam process would be required for this form of modelling to be adequately robust to enable adjustment of students' scores. Above all, we wish to draw attention to the potential for different standards of judgement to be employed in different cycles of OSCE exams and to stimulate debate around the assumptions and design of multi-cycle OSCE exams in order to enhance the fairness and defensibility of assessments for our students.

Acknowledgements:

Dr. Emyr Benbow, Manchester Medical School for invaluable expertise in our discussions of OSCE design and conduct.

Dr. Peter D. MacMillan for his thoughtful review and suggestions for improvement of this manuscript.

Declaration of interest

The authors report no declarations of interest.

Funding

None

Contributorship

None

References

Bond, T., & Fox, C. (2012). Applying the Rasch Model Fundamental Measurement in the Human Sciences (2nd Editio). New York & London: Routledge.

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. Medical Education, 45, 1181–1189.

Case, S. M., & Swanson, D. B. (1993). Extended-matching items: A practical alternative to free-response questions. Teaching and Learning in Medicine, 5(2), 107–115.

Crossley, J., Davies, H., Humphris, G., & Jolly, B. (2002). Generalisability: a key to unlock professional assessment. Medical Education, 36(10), 972–8.

Downing, S. M. (2003). Validity: on meaningful interpretation of assessment data. Medical Education, 37(9), 830–7.

Epstein, R. M., & Hundert, E. M. (2002). Defining and Assessing Professional Competence. JAMA, 287(2), 226–235.

Field, A. (2009). Discovering statistics using SPSS (3rd ed.). Los Angeles: Sage.

Hafferty, F. W., & Franks, R. (1994). The Hidden Curriculum, Ethics Teaching, and the Structure of Medical Education. Academic Medicine, 69(11), 861–871.

Harasym, P. H., Woloschuk, W., & Cunning, L. (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. Advances in Health Sciences Education, 13(5), 617–32.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Medical Education Assessment of Clinical Competence using Objective Structured Examination. British Medical Journal, 1, 447–451.

Hope, D., & Cameron, H. (2015). Examiners are most lenient at the start of a two-day OSCE. Medical Teacher, 37, 81–85.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), Educational measurement (4th ed., pp. 17–64). Westport, Conneticut: American Council on Education/Praeger.

Kogan, J. R., Hess, B. J., Conforti, L. N., & Holmboe, E. S. (2010). What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. Academic Medicine, 85(10 Suppl), S25–8.

Linacre, J. M. (1996). FACETS: A computer program for analysis of examinations with multiple facets. Chicago: MESA.

Linacre, J. M. (2011) Facets computer program for many-facet Rasch measurement, version 3.68.1 Beaverton, Oregon: Winsteps.com

McLaughlin, K., Ainslie, M., Coderre, S., Wright, B., & Violato, C. (2009). The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Medical Education, 43(10), 989–92.

Newblet, D. I., & Swansons, D. B. (1988). Psychometric characteristics of the objective structured clinical examination, 22, 325–334.

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), 802–11.

Raymond, M. R., Swygert, K. a, & Kahraman, N. (2012). Measurement precision for repeat examinees on a standardized patient examination. Advances in Health Sciences Education, 17(3), 325–37.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). General overview of the theories used in assessment: AMEE Guide No. 57. Medical Teacher, 33(10), 783–97.

Sebok, S. S., Luu, K., & Klinger, D. a. (2014). Psychometric properties of the multiple mini-interview used for medical admissions: findings from generalizability and Rasch analyses. Advances in Health Sciences Education, 19(1):71-84.

Sebok, S. S., Roy, M., Klinger, D. a, & De Champlain, A. F. (2015). Examiners and content and site: Oh My! A national organization's investigation of score variation in large-scale performance assessments. Advances in Health Sciences Education, 20(3):581-94.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44(6), 922–932.

Streiner, D., & Norman, G. (2008). Health Measurement Scales (4th ed.). Oxford: Oxford University Press.

van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: from methods to programmes. Medical Education, 39(3), 309–17.

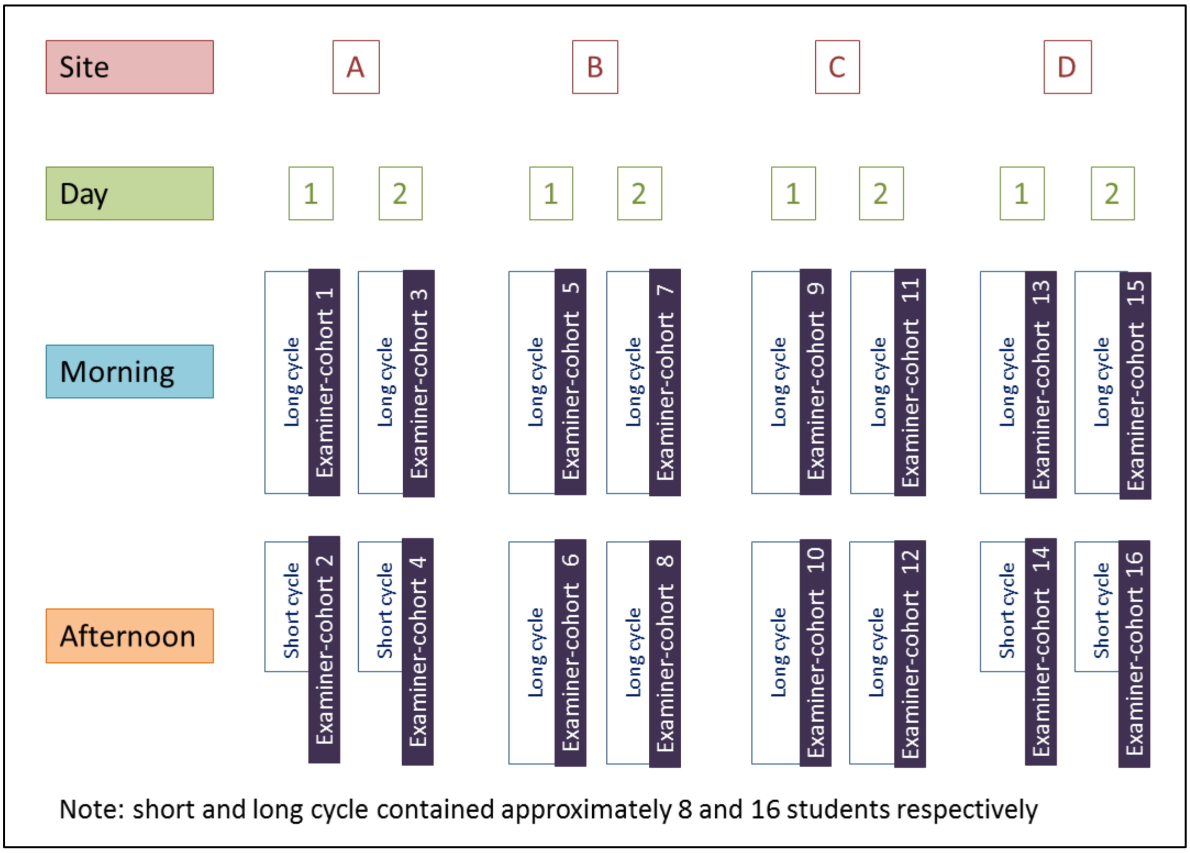
Watling, C. J. (2014). Unfulfilled promise, untapped potential: Feedback at the crossroads. Medical Teacher, 36, 692–697.

Weiner, J. P., Parente, S., Garnick, D., J, P., Lawthers, A., & Palmer, R. (1995). Variation in office based quality: a claims based profile of care provided to Medicare patients with diabetes. JAMA, 273, 1503–1508.

Yeates, P., Moreau, M., & Eva, K. (2015). Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects? Academic Medicine, 90(7), 975–980.

Tables and figures:

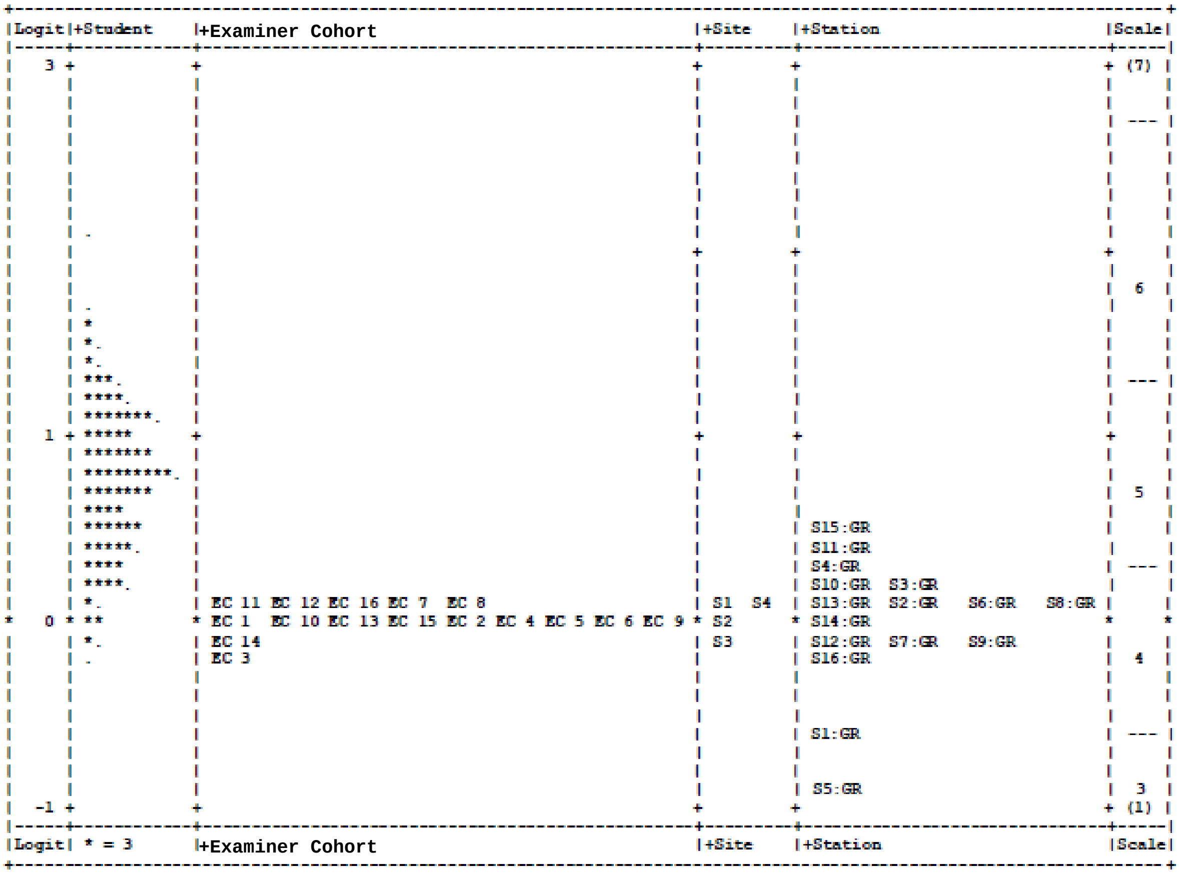
Figure 1: Schematic of OSCE cycle design



Site	A		B		C		D	
Day	1	2	1	2	1	2	1	2
Morning	Long cycle Examiner-cohort 1	Long cycle Examiner-cohort 3	Long cycle Examiner-cohort 5	Long cycle Examiner-cohort 7	Long cycle Examiner-cohort 9	Long cycle Examiner-cohort 11	Long cycle Examiner-cohort 13	Long cycle Examiner-cohort 15
Afternoon	Short cycle Examiner-cohort 2	Short cycle Examiner-cohort 4	Long cycle Examiner-cohort 6	Long cycle Examiner-cohort 8	Long cycle Examiner-cohort 10	Long cycle Examiner-cohort 12	Short cycle Examiner-cohort 14	Short cycle Examiner-cohort 16

Note: short and long cycle contained approximately 8 and 16 students respectively

Figure 2: Wright Map of the OSCE data showing relative influence of Students, Examiner cohorts, Sites and Stations



```
+-------------------------------------------------------------------------------------------------------------------------+
|Logit|+Student     |+Examiner Cohort                                   |+Site     |+Station                        |Scale|
|-----+-------------+---------------------------------------------------+----------+--------------------------------+-----|
|   3 +             +                                                   +          +                                + (7) |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                | --- |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     | .           |                                                   |          |                                |     |
|     |             |                                                   +          +                                +     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |  6  |
|     | .           |                                                   |          |                                |     |
|     | *           |                                                   |          |                                |     |
|     | *.          |                                                   |          |                                |     |
|     | *.          |                                                   |          |                                |     |
|     | ***.        |                                                   |          |                                | --- |
|     | ****.       |                                                   |          |                                |     |
|     | *******.    |                                                   |          |                                |     |
|   1 + *****       +                                                   +          +                                +     |
|     | *******     |                                                   |          |                                |     |
|     | *********.  |                                                   |          |                                |     |
|     | *******     |                                                   |          |                                |  5  |
|     | ****        |                                                   |          |                                |     |
|     | ******      |                                                   |          | S15:GR                         |     |
|     | *****.      |                                                   |          | S11:GR                         |     |
|     | ****        |                                                   |          | S4:GR                          | --- |
|     | ****.       |                                                   |          | S10:GR  S3:GR                  |     |
|     | *.          | EC 11 EC 12 EC 16 EC 7  EC 8                      | S1   S4  | S13:GR  S2:GR   S6:GR   S8:GR  |     |
*   0 * **          * EC 1  EC 10 EC 13 EC 15 EC 2 EC 4 EC 5 EC 6 EC 9  * S2       * S14:GR                         *     *
|     | *.          | EC 14                                             | S3       | S12:GR  S7:GR   S9:GR          |     |
|     | .           | EC 3                                              |          | S16:GR                         |  4  |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          | S1:GR                          | --- |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          |                                |     |
|     |             |                                                   |          | S5:GR                          |  3  |
|  -1 +             +                                                   +          +                                + (1) |
|-----+-------------+---------------------------------------------------+----------+--------------------------------+-----|
|Logit| * = 3       |+Examiner Cohort                                   |+Site     |+Station                        |Scale|
+-------------------------------------------------------------------------------------------------------------------------+
```

Table 1: Observed and model adjusted scores for examiner cohorts

	Scores			Logits	
Examiner Cohort	Observed Average	Model-derived Fair Average	Difference (% max score)	Measure	SE
1	5	4.98	0.02 (0.4%)	-0.02	0.06
2	4.9	5	-0.1 (-2.0%)	-0.01	0.06
3	4.6	4.8	-0.2 (-4.3%)	-0.2	0.06
4	4.9	4.98	-0.08 (-1.6%)	-0.02	0.07
5	4.9	5.01	-0.11 (-2.2%)	0	0.07
6	5	4.97	0.03 (0.6%)	-0.04	0.08
7	5.2	5.11	0.09 (1.7%)	0.1	0.08
8	5.2	5.1	0.1 (1.9%)	0.09	0.06
9	5	5.05	-0.05 (-1.0%)	0.04	0.07
10	4.9	5.02	-0.12 (-2.4%)	0.02	0.07
11	5.3	5.1	0.2 (3.8%)	0.09	0.06
12	5	5.09	-0.09 (-1.8%)	0.08	0.06
13	4.9	4.96	-0.06 (-1.2%)	-0.05	0.06
14	4.8	4.86	-0.06 (-1.3%)	-0.14	0.08
15	4.9	5	-0.1 (-2.0%)	-0.01	0.06
16	5.1	5.06	0.04 (0.8%)	0.05	0.08
Mean	5	5.01		0	0.07
S.D. (popn)	0.2	0.08		0.08	0.01
S.D. (sample)	0.2	0.09		0.08	0.01

Abbreviations:

% dif = Percentage difference between observed score and model derived fair score, as a percentage of the observed score.

SE = standard error

MnSq = Mean Square